

The 17th **Knowledge and Systems Engineering** International Conference

Editors:

Duong Van Hai
Shinobu Hasegawa
Hieu Dinh Vo
Son Nguyen

November 2025
Da Lat University, Vietnam



**THE 17TH INTERNATIONAL CONFERENCE ON
KNOWLEDGE AND SYSTEMS ENGINEERING**

KSE 2025

November 2025
Da Lat University, Vietnam

COPYRIGHT

The 17th International Conference on
Knowledge and Systems Engineering (KSE 2025)

Part number: CFP2503I-ART
ISBN: 979-8-3315-8900-4
Online ISSN: 2694-4804

Copyright and Reprint Permission: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For reprint or republication permission, email to IEEE Copyrights Manager at pubs-permissions@ieee.org. All rights reserved. Copyright ©2025 by IEEE.

Welcome Message from The KSE 2025 General Committee

The 2025 IEEE International Conference on Knowledge and Systems Engineering (KSE 2025) marks the 17th meeting of the KSE series. This year, the conference is proudly hosted onsite at Da Lat University, Da Lat City, Vietnam, from November 6-8, 2025.

Since its inception in 2009 by the College of Technology (now University of Engineering and Technology), Vietnam National University (VNU), Hanoi, the KSE conference has grown into a reputable annual scientific forum for researchers, academics, and practitioners in the fields of Knowledge Engineering, Artificial Intelligence, and Systems Engineering. Over the years, KSE has been hosted by various leading institutions across Vietnam and internationally, including Hanoi, Ho Chi Minh City, Hue, Da Nang, Can Tho, Thailand, and Malaysia, with KSE 2024 successfully organized in Kuala Lumpur, Malaysia.

This year, KSE 2025 returns to Vietnam and is held in the picturesque city of Da Lat, well known for its natural beauty, academic atmosphere, and vibrant research community. The conference continues its mission to foster collaboration, share innovative research ideas, and promote interdisciplinary discussion among scholars and industry experts.

On behalf of the KSE 2025 Organizing Committee, we would like to express our deep appreciation to all keynote speakers for sharing their invaluable insights, to our Program Chairs, Special Session Chairs, and Reviewers for their dedicated efforts in ensuring the scientific quality of the program, and to all authors and participants for their valuable contributions.

We extend our sincere gratitude to our hosts and co-hosts, Da Lat University (DLU), VNU University of Engineering and Technology (UET), Japan Advanced Institute of Science and Technology (JAIST), and the Vietnam Association of ICT Faculties, Institutes, Schools, and Universities (FISU Vietnam), for their strong support. We also warmly thank our sponsors and partners for their generous contributions to making this event possible.

Special thanks go to the Local Organizing Committee at Da Lat University for their excellent work in preparing and hosting the conference, ensuring a productive and memorable experience for all participants.

We believe that KSE 2025 will continue the strong tradition of the KSE series, fostering international collaboration, inspiring innovation, and advancing research in Knowledge and Systems Engineering. We hope that all participants will find the conference intellectually stimulating and that their stay in Da Lat will be enjoyable and enriching.

KSE 2025 General Co-chairs

Nguyen Le Minh (JAIST, Japan)

Le Sy Vinh (VNU University of Engineering and Technology, Vietnam)

Bui Thu Lam (Academy of Cryptography Techniques, Vietnam)

Author Index

Ahmed Mohammed, Muzakkiruddin	417
Al Mandalawi, Shames	417
Bui Huynh, Trung Nam	85
Bui, Cong Thanh	483
Bui, Quan	222
Bui, Quang-Thinh	406
Bui, Thu Lam	334
Bui, Trong Tung	544
Bui, Van-Tan	328
Cakmak, Mert Can	417
Cao Thanh, Truc	298
Cao Thi, Nhan	358
Cao, Chi-Hai	501
Chau, The Khanh	43
Christoff, Zoé	180
Chu Thi, Linh	310
Claasssens, Leon	19
Cong Thuong, Le	109
D. Nguyen, Hien	454
Dang Quoc, Huu	274
Dang, Chi-Nguyen	435
Dang, Minh	394
Dang, Nhat Duy	544
Dang, Nhat-Minh	304
Dang, Thanh Hai	7
Dang, Thanh Trung	228
Dang, Thanh-Hai	400
Dinh, Viet Sang	114, 132
Do Thi, Thanh-Tam	328
Do, Dinh Truong	477
Do, Nhu-Tai	174
Do, Thanh	370
Do, Thi-Thu-Trang	304
Do, Trong-Hop	435
Do-Chi, Cuong	394
Doan, Dang Phuong Nam	465
Doan, Nhat Quang	120
Doan, Nhat-Quang	144
Doan, Quang-Minh	49
Dong, Peizhe	25
Duc Hanh, Dang	250
Duc Minh, Nguyen	531
Duc, Le Xuan	61
Dung, Le The	340
Dung, Nguyen Trung	61
Duong, Hai	262, 286
Duong, Minh Ngoc	364
Duong, Minh-Quan	150
Duong, Trinh-Hoai-An	501
Duong, Trung	394
Duong-Bao, Ninh	358
Duong-Thi, Mong-Thuy	435
Giang, Anh Tuan	120, 144
Ha Minh, Tuan	103
Ha Thi Xuan, Chi	298

Ha, Thi Lien Doanh	412
Hasegawa, Shinobu	192, 216
Hiep, Nguyen	222
Hoang Ngoc, Quan	126
Hoang Thi, Thanh-Nga	376
Hoang, Cuong	550
Hoang, Diep Thi	388
Hoang, Dung	429
Hoang, Gia Hung	364
Hoang, Quoc-Viet	304
Hoang, Thanh Tung	138
Hoang, Thu-Huong	394
Hoang, Tien	286
Hoang, Van-Dung	423
Hoang-Van, Quyen	210
Hoshikoshi, Daichi	519
Hung, Nguyen The	61
Huynh, Hiep	406
Huynh, Hieu M.	471
Katsurada, Kouichi	513, 519
Khuong, Vu Tram Anh	91
La, Thang	262
Le Dinh, Thanh	234
Le Hong, Hai	495
Le Khanh, Trinh	210
Le Ngoc, Tuyen	316
Le Nguyen Hai, Dang	31
Le, Bac	262, 286, 340
Le, Bang Giang	168
Le, Cong Thuong	138
Le, Duc-Trong	55, 204
Le, Hai	346
Le, Hoai Nam	120
Le, Huu Ton	120, 144
Le, Minh-Khoi	204
Le, Nguyen-Khang	192, 222
Le, Quan Minh	489
Le, Thanh Ha	91, 97, 109, 138
Le, Thi Quynh	483
Le, Thu	340
Le, Tri	454
Le, Trong-Dai Truong	126
Le, Tuan Anh	412
Le, Tung	465, 489
Le, Van Dong	544
Le, Viet Nam	138
Le, Viet Quan	186
Le, Vu Minh	156, 198
Le, Xuan Duc	483
Le, Xuan Hien	228
Lee, Shee Weng	240
Lee, Tze Houng	240
Linh Hoang, Chi	256
Loi, Dinh-Nhat	400
Luong, Duy-Quan	150
Luu, Son	222
Luu, Thanh Son	477
Ly, Dat	454

Ly, Nhat Huy	144
Lê, Ngoc	394
Ma, Kim-Phat	435
Ma, Thanh	43
Mac, Quang Dat	376
Maclean, Hendrika	417
Mai, Dinh Sinh	412
Mai, Khang Huy	471
Mai, Son Thai	370
Man, Thi Bich Phuong	91, 97
Marais, Adriaan	19
Mohammed, Muzakiruddin Ahmed	19
N. Nguyen, Hai	103
Nakamura, Makoto	507
Ngo Manh, Hung	316
Ngo Van, Trung	310
Ngo, Thao	43
Ngo, Thi Duyen	91, 97, 138
Ngoc Hung, Pham	210
Ngoc Le, Tuyen	544
Nguyen Duc, Dinh	268
Nguyen Hoang, Dac	174
Nguyen Huynh, Tra My	186
Nguyen Le, Minh	31
Nguyen Le, Tuong	280
Nguyen Linh, Giang	316
Nguyen Minh, Hien	388
Nguyen Ngoc Kim, Ngan	298
Nguyen Quoc, Huy	174
Nguyen The, Loc	274
Nguyen Thi, Hao	274
Nguyen Thi, Luong	358
Nguyen Van, Hop	280
Nguyen Van, Vinh	245
Nguyen Viet, Quang	109
Nguyen Vo, Ngoc Khue	114, 132
Nguyen, Anh Kiet	465
Nguyen, Anh Tu	483
Nguyen, Ba-Duy	423
Nguyen, Binh	429
Nguyen, Cam-Van Thi	13
Nguyen, Cong Thanh	334
Nguyen, Dieu-Hien	192
Nguyen, Do Phuong	525
Nguyen, Duc Phu	352
Nguyen, Duc-Hieu	447
Nguyen, Duy Hai	228
Nguyen, Giang	322, 544
Nguyen, Ha	394
Nguyen, Hien	429
Nguyen, Hieu Nghia	126
Nguyen, Hoa	346
Nguyen, Hoang Ha	120, 144
Nguyen, Hoang Lam	114, 132
Nguyen, Hung	346, 406, 454
Nguyen, Huy Tien	465, 489
Nguyen, Ke-Luong	79
Nguyen, Kiet Van	471

Nguyen, Lam M. H.	471
Nguyen, Le Minh	477
Nguyen, Le-Minh	192, 222, 550
Nguyen, Long	268
Nguyen, Luu Tu	91, 97
Nguyen, Mau	216
Nguyen, Minh	222, 346
Nguyen, Minh-Huong	328
Nguyen, Minh-Tien	501
Nguyen, Ngoc An	352
Nguyen, Ngoc Hoa	352
Nguyen, Ngoc-Linh	382
Nguyen, Phong Tuan	489
Nguyen, Phuong-Thai	73, 328
Nguyen, Quang Phong	525
Nguyen, Quang Tri	114, 132
Nguyen, Quang-Hoan	49
Nguyen, Quoc Huy	168
Nguyen, Quy-Nhan	501
Nguyen, Son	150, 162, 394
Nguyen, Tan Hai	228
Nguyen, Thanh	292
Nguyen, Thanh Thai	465
Nguyen, The Hung	483
Nguyen, Thi Quynh Hoa	228
Nguyen, Thi-Hau	292
Nguyen, Thu-Trang	150, 162
Nguyen, Thuan Dinh	370
Nguyen, Tien-Khoi	73
Nguyen, Tri-Thanh	376, 382, 400
Nguyen, Trong Tuan	114, 132
Nguyen, Trung Dung	483
Nguyen, Trung-Hieu	150, 162, 204
Nguyen, Truong-Phuc	501
Nguyen, Tuyen	346
Nguyen, Van Anh	483
Nguyen, Van Kiet	126
Nguyen, Van-Hieu	67
Nguyen, Van-Quyet	304
Nguyen, Van-Vinh	73
Nguyen-Huu, Khanh	358
Nguyen-Le, Hai-Chau	109
Nhu, Mac Thi Quynh	61
Ohmura, Hidefumi	513, 519
Pham Ngoc Thuy, Trang	216
Pham Thi, Quynh Trang	7
Pham Thi, Quynh-Trang	376, 382, 400
Pham Xuan, Vinh	245
Pham, An Duc Vinh	7
Pham, Duy Trung	334
Pham, Hau	429
Pham, Linh	394
Pham, Nguyen Khoi	370
Pham, Quang Long	334
Pham, Thang Q.	471
Pham, Truong Son	525
Pham, Tung-Thuy	150, 162
Phan Thi Thanh, Nga	358

Phan, Manh-Cuong	501
Phuc, Vu Hoang	61
Q. Nguyen, Hung	454
Quan Thanh, Tho	31
Quan, Xuan-Son	73
Saito, Daichi	507
Sawada, Shun	513, 519
Shoshin, Nomura	222
Ta, Quang-Chieu	49
Ta, Son	394
Ta, Viet Cuong	156, 168, 198
Ta, Xuan Duy	156, 198
Talburt, John	19, 417
Tang Vinh, Ha	388
Thai, Phu An	43
Thi Bich Pham, Van	310
Thi Hoang, Ngoc-Trang	55
Thi-Hoang Nguyen, Anh	435
To Thi, Thu	268
Tojo, Satoshi	519
Tong, Van	316, 322
Tran Thi Thanh, Dung	103
Tran, Anh	262
Tran, Anh Tu	537
Tran, Duc	316, 322
Tran, Dung	429
Tran, Hai Anh	322
Tran, Hoang Tung	120, 144
Tran, Hong-Viet	49, 73, 328
Tran, Huu-Tri	1
Tran, Huy-Thinh	162
Tran, Khanh	262
Tran, Nha	454
Tran, Tien-Bang	37
Tran, Trong-Binh	423
Tran, Tuan Long	228
Tran, Van-Khanh	1, 37, 67, 79
Tran, Vu	25, 550
Tran-Diem-Hanh, Nguyen	460
Trieu, Huy	465
Trinh Hoang Viet, La	250
Trinh, Hai Tien	91
Trinh, Linh	370
Trinh, Quang-Hung	162
Trinh, Tuan-Dat	441
Truong Dinh, Hung	103
Truong Thi, Kieu-Anh	55
Truong, Dat Tien	489
Truong, Gia-Bach	441
Truong, Minh Thai	43
Truong, Tin	286
Truong, Xuan Hung	537
Tsuchida, Kohei	513
Ung, Thanh Tra	334
Van Kieu, Tuyen	256
Van To, Khanh	256
Van, Pham Thi Bich	61
Vandierendonck, Hans	370

Vi, Thi Thoa	334
Viet Hung, Nguyen	531
Vo Hoang Nhat, Khang	31
Vo, Hieu Dinh	150, 162
Vo, Tan	43
Vo, Thanh Nghia	465
Vo, Trung	222
Vu Van, Long	495
Vu, Hoang-Duy	55
Vu, Hoang-Long	13
Vu, Ly	310
Vu, Van Canh	138
Vu-Thanh, Trung	358
Vuong, Dinh Minh	114, 132
Vuong, Nhan	340
Yoshitaka, Atsuo	85
Zheng, Xianwen	216
Ågotnes, Thomas	180

Program Committee

Fahad Ahmad
Razvan Beuran
Quang-Thinh Bui
Lam Bui Thu
Thanh-Duc Chau
Huy Chu
Nguyen-Hoang Chu
Dai-Tho Dang

Duc-Hanh Dang
Thanh Hai Dang
Minh Son Dao
Duong Dinh Tran
Cong Do
Kha Do
Truong Do
Nguyen Do-Van

Thanh-Nghi Doan
Thu Doan
Wen Gu
Shinobu Hasegawa
Lam Ho
Van Lam Ho
Diep Thi Hoang
Xuan Dau Hoang
Xiem Hoangvan
Tzung-Pei Hong

Tho Huynh Ngoc
Dương Văn Hải

Shunsuke Kidani
Dang Kieu
Nattapol Kritsuthikul

Duc-Trong Le
Khang Le
Tung Le

Vinh Le
Trinh Le Khanh
Nhien-An Le-Khac
Tuan Le-Viet
Jinghong Li
Son Luu
Chuc Man
Linh Manh Pham
Canh Minh Do
Le Nguyen Hoai Nam
Thi Duyen Ngo
Tuan Ngo
Xuan Truong Ngo
Anh Thi Van Nguyen
Binh Nguyen

University of Portsmouth
Japan Advanced Institute of Science and Technology
Tien Giang University
ACTVN
University of Science, HCMUS
Japan Advanced Institute of Science and Technology
Japan Advanced Institute of Science and Technology
Vietnam-Korea University of Information and Communication Technology, The University of Danang
VNU University of Engineering and Technology
VNU University of Engineering and Technology
National Institute of Information and Communications Technology
JAIST
Smart Computing Lab
Latrobe
Japan Advanced Institute of Science and Technology (JAIST)
National Institute of Information and Communications Technology (NICT), Japan
An Giang University
GSSI
Nagoya Institute of Technology
JAIST
Dalhousie University
Quy Nhon University
Vietnam National University, Hanoi
Posts and Telecommunications Institute of Technology (PTIT)
VNU University of Engineering and Technology
Department of Computer Science and Information Engineering, National University of Kaohsiung
IRISA/Université Européenne de Bretagne/TELECOM Bretagne
Department of Mathematics and Computer Science, University of Dalat, Vietnam
Japan Advanced Institute of Science and Technology
VinUniversity
National Electronics and Computer Technology Center (NECTEC), Thailand
VNU University of Engineering and Technology
Japan Advanced Institute of Science and Technology
Faculty of Information Technology, University of Science, VNU-HCM, Vietnam
Ho Chi Minh University of Technology and Education
VNU University of Engineering and Technology
School of Computer Science, University College Dublin
HoChiMinh City Open University
J-MAX Co., Ltd.
Japan Advanced Institute of Science and Technology
VNU University of Engineering and Technology
VNU University of Engineering and Technology
Japan Advanced Institute of Science and Technology
VNUHCM - University of Science
VNU University of Engineering and Technology
USC
VNU University of Engineering and Technology
Yokohama National University
University of Science

Binh Minh Nguyen	Hanoi University of Science and Technology
Dai Tho Nguyen	VNU University of Engineering and Technology
Dieu-Hien Nguyen	Japan Advanced Institute of Science and Technology
Duc Anh Nguyen	University of Engineering and Technology
Ha-Thanh Nguyen	National Institute of Informatics
Hien Nguyen	VNUHCM - University of Information Technology
Hiep Nguyen	Japan Advanced Institute of Science and Technology
Hung Nguyen	HCMUE
Khanh Nguyen	SCIT
Minh-Phuong Nguyen	Japan Advanced Institute of Science and Technology
Minh-Tien Nguyen	UTEHY
Ngoc-Minh Nguyen	Japan Advanced Institute of Science and Technology
Son Nguyen	VNU University of Engineering and Technology
Tan Nguyen	VNU University of Engineering and Technology
Tan-Minh Nguyen	Japan Advanced Institute of Science and Technology
Thanh Tuan Nguyen	Faculty of Information Technology, HCMC University of Technology and Education
The-Hai Nguyen	Japan Advanced Institute of Science and Technology
Thu-Trang Nguyen	University of Engineering and Technology, VNU
Tri-Hai Nguyen	Van Lang University
Trinh-Dong Nguyen	Faculty of Software Engineering, UIT - VNU-HCM
Viet Anh Nguyen	University of Engineering and Technology, Vietnam National University Hanoi
Viet Cuong Nguyen	HPC SYSTEMS Inc.
Viet Hung Nguyen	LQDTU
Minh Nguyen Le	JAIST
Thin Nguyen Si	Vietnam - Korea University of Information and Communication Technology
Duc Nguyen Trung	Yeungnam University
Thien B. Nguyen-Tat	UIT
Cong Thang Pham	The University of Da Nang - University of Science and Technology, 54 Nguyen Luong Bang, Da Nang, Viet Nam
Xuan Hau Pham	Quang Binh University
Thang Pham	VU University Medical Center
Trang Pham	Japan Advanced Institute of Science and Technology
Truong Son Pham	Le Quy Don University
Viet Anh Phan	LQDTU
Tho Quan	Hochiminh City University of Technology
Viet Cuong Ta	VNU University of Engineering and Technology, Hanoi
Nguyen Tien Huy	University of science HCM
Hlaing Htake Khaung Tin	University of Information Technology
Duy-Hoang Tran	University of Science, VNUHCM
Hanh Nhi Tran	St-Cyr Coetquidan Military Academy, University of Toulouse 3
Hoang-Viet Tran	VNU University of Engineering and Technology
Trong Hieu Tran	Vietnam National University, Hanoi
Van-Khanh Tran	Institute of Applied Science and Technology, Thai Nguyen University of Information and Communication Technology
Vu Tran	JAIST
An Trieu	Japan Advanced Institute of Science and Technology
Tin Truong	Dalat University
Van Truong	Le Quy Don Technical University
Vinh Truong Hoang	Ho Chi Minh City Open University
Dung V. Dang	University of Information Technology, Ho Chi Minh city, Vietnam
To Van Khanh	VNU University of Engineering and Technology
Vinh Van Nguyen	College of Technology, VNU
Hieu Vo	VNU University of Engineering and Technology
Trung Vo	Japan Advanced Institute of Science and Technology
Dinh Phai Vu	University of Technology Sydney

Hien Vu	NEU
Ly Vu	Le Quy Don Technical University
Nhan Vu	VNU University of Engineering and Technology
Sang Vu	University of Information Technology, VNU-HCM
Trung Nghia Vu	Karolinska Institutet
Viet Anh Vu	University of Cambridge
Viet Thang Vu	Hanoi University of industry
Thi-Hai-Yen Vuong	VNU University of Engineering and Technology
Haoran Xie	JAIST
Atsuo Yoshitaka	Japan Advanced Institute of Science and Technology
Nurulhuda Zainuddin	Universiti Teknologi MARA

Additional Reviewers

Binh, Vo Phuong; Bui, Quang-Cuong; Dang, Binh; Dang, Do; Do, Dinh Truong; Ha, Minh-Quyet; Hanh, Nguyen Thi; Hao, Do; Htake Khaung Tin, Hlaing; Hung, Bui Thanh; Kieu, Dang; Kieu, Hai-Dang; Le, Khanh-Duy; Le, Nguyen-Khang; Le, Thi Huong; Le, Tung; Luân, Phạm; Ma, Thanh; Nghiem, Nguyen Viet Dung; Nguyen, Hanh; Nguyen, Huy-Son; Nguyen, Lam; Nguyen, Manh Hung; Nguyen, Ngoc Minh; Nguyen, Thanh Binh; Nguyen, Thi Minh Tuyen; Nguyen, Tuan-Phong; Nguyễn, Hoàng Hà; Nguyễn Xuân, Trường; Phạm, Nga; Sơn, Trang Hồng; Tà, Thang Hoang; Tien Huy, Nguyen; Tran, Manh Cuong; Tran, Tu; Tran Thi Thu, Phuong; Trinh, Van-Giang; Truong, Cong Doan; Tung, Hoang Thanh; Vo, Trung; Vu, Dinh Phai; Vu, Thanh; Vu, Tien Sinh

Table of Contents

Large Language Models, NLP, and Conversational AI	
An Investigation of Collocation Segmentation Approaches in Neural Machine Translation	1
<i>Van-Khanh Tran and Huu-Tri Tran</i>	
ANAVIT: Enhancing Document-Level Relation Extraction with Anaphor Nodes and Visual Transformation	7
<i>An Duc Vinh Pham, Quynh Trang Pham Thi and Thanh Hai Dang</i>	
Integrating Cause-awareness for Improving Emotional Support in Conversational AI	13
<i>Hoang-Long Vu and Cam-Van Thi Nguyen</i>	
Retrieval-Augmented Multi-LLM Ensemble for Industrial Part Specification Extraction	19
<i>Muzakkiruddin Ahmed Mohammed, John Talburt, Leon Claasssens and Adriaan Marais</i>	
SLB:A Bridge Module of Local SLM and Online LLM for Smart Home Device Control	25
<i>Peizhe Dong and Vu Tran</i>	
Multimodal Large Language Model for out-of-context Problems in Fake News Detection	31
<i>Dang Le Nguyen Hai, Khang Vo Hoang Nhat, Minh Nguyen Le and Tho Quan Thanh</i>	
Fusion-based Document Retrieval for Low-Resource Vietnamese Legal Texts	37
<i>Van-Khanh Tran and Tien-Bang Tran</i>	
SHREADS: A Shrimp Health Diagnosis and Advisory System using Agentic RAG	43
<i>Thao Ngo, Phu An Thai, Tan Vo, The Khanh Chau, Thanh Ma and Minh Thai Truong</i>	
Multi-Stage User Intent Classification for Educational AI Agents	49
<i>Quang-Minh Doan, Hong-Viet Tran, Quang-Chieu Ta and Quang-Hoan Nguyen</i>	
RAID: Robust Out-of-Distribution Intent Detection with Augmented and Discriminative Training .	55
<i>Hoang-Duy Vu, Kieu-Anh Truong Thi, Ngoc-Trang Thi Hoang and Duc-Trong Le</i>	
EffWhiS: A Proposed Efficient Approach for Speech-to-Text Streaming Whisper	61
<i>Mac Thi Quynh Nhu, Nguyen The Hung, Vu Hoang Phuc, Nguyen Trung Dung, Le Xuan Duc and Pham Thi Bich Van</i>	
LAURA: A Legal Agent Using Reasoning and Acting for Vietnamese Law	67
<i>Van-Khanh Tran and Van-Hieu Nguyen</i>	
Toward Effective Domain Adaptive Continual Pre-training for Vietnamese Large Language Models.	73
<i>Tien-Khoi Nguyen, Xuan-Son Quan, Hong-Viet Tran, Van-Vinh Nguyen and Phuong-Thai Nguyen</i>	
Improving Legal Reasoning in LLMs through Structured Prompting	79
<i>Van-Khanh Tran and Ke-Luong Nguyen</i>	
Computer Vision and Generative Models	
Optimizing Image Dehazing with Balanced Loss and Feature Map Enhancement	85
<i>Trung Nam Bui Huynh and Atsuo Yoshitaka</i>	
Enhancing Micro-Expression Recognition via Multi-Task Learning with Demographic-Aware Auxiliary Supervision	91
<i>Hai Tien Trinh, Thi Bich Phuong Man, Vu Tram Anh Khuong, Luu Tu Nguyen, Thanh Ha Le and Thi Duyen Ngo</i>	
A lightweight visual deception detection method using dynamic images and a customized SmallCNN-BiLSTM architecture	97
<i>Thi Bich Phuong Man, Luu Tu Nguyen, Thanh Ha Le and Thi Duyen Ngo</i>	
Deep learning-based segmentation of Rooftop photovoltaic panels in Ho Chi Minh City, Vietnam . .	103
<i>Hai N. Nguyen, Dung Tran Thi Thanh, Hung Truong Dinh and Tuan Ha Minh</i>	

HySUP-F: Face-Enhanced Full-Body Reconstruction from a Single Image.....	109
<i>Le Cong Thuong, Quang Nguyen Viet, Hai-Chau Nguyen-Le and Thanh Ha Le</i>	
ZenID: FaceFusionNet-Guided Diffusion for Parent-to-Child Facial Prediction.....	114
<i>Hoang Lam Nguyen, Dinh Minh Vuong, Trong Tuan Nguyen, Quang Tri Nguyen, Ngoc Khue Nguyen Vo and Viet Sang Dinh</i>	
Towards Real-ESRGAN Refinement for Realistic Image Restoration via Advanced Degradation and Knowledge Distillation.....	120
<i>Hoai Nam Le, Nhat Quang Doan, Anh Tuan Giang, Hoang Tung Tran, Huu Ton Le and Hoang Ha Nguyen</i>	
Data Generation Based on Multimodal Language Models for Vietnamese Visual Question Answering.....	126
<i>Trong-Dai Truong Le, Quan Hoang Ngoc, Hieu Nghia Nguyen and Van Kiet Nguyen</i>	
Sketch2Comic: A Diffusion-Based Framework for Consistent Multi-Character Comic Synthesis.....	132
<i>Hoang Lam Nguyen, Trong Tuan Nguyen, Dinh Minh Vuong, Ngoc Khue Nguyen Vo, Quang Tri Nguyen and Viet Sang Dinh</i>	
3D Vietnamese Woodblocks Repair with Normal-Enhanced Fused Shape Representation.....	138
<i>Viet Nam Le, Thanh Tung Hoang, Van Canh Vu, Cong Thuong Le, Thi Duyen Ngo and Thanh Ha Le</i>	

Multi-Agent Systems, Reinforcement Learning, and Recommender Systems

Enhancing Urban Traffic Control with Priority-Aware Multi-Objective Reinforcement Learning....	144
<i>Nhat Huy Ly, Nhat-Quang Doan, Anh Tuan Giang, Huu Ton Le, Hoang Ha Nguyen and Hoang Tung Tran</i>	
CABENCH: Benchmarking Composable AI for Solving Complex Tasks through Composing Ready-to-Use Models.....	150
<i>Tung-Thuy Pham, Duy-Quan Luong, Minh-Quan Duong, Trung-Hieu Nguyen, Thu-Trang Nguyen, Son Nguyen and Hieu Dinh Vo</i>	
An Experimental Study of Linear Attention in Multi-Agent Transformer.....	156
<i>Xuan Duy Ta, Vu Minh Le and Viet Cuong Ta</i>	
ABSTRACT-THEN-CONCRETIZE Automated Design of Multi-Agent System.....	162
<i>Trung-Hieu Nguyen, Huy-Thinh Tran, Quang-Hung Trinh, Tung-Thuy Pham, Thu-Trang Nguyen, Son Nguyen and Hieu Dinh Vo</i>	
Enhance Hyper-Actor Critic Framework in Session-based Recommendation System with Nash Social Welfare.....	168
<i>Quoc Huy Nguyen, Bang Giang Le and Viet Cuong Ta</i>	
Explainable Decomposed Reward Q-learning with Roulette Wheel Selection in Maze Navigation ...	174
<i>Dac Nguyen Hoang, Nhu-Tai Do and Huy Nguyen Quoc</i>	
Reasoning about Coalitional Ability in Network Threshold Models.....	180
<i>Thomas Ågotnes and Zoé Christoff</i>	
Manifold-Preserving Social Recommendation in Hyperbolic Space.....	186
<i>Tra My Nguyen Huynh and Viet Quan Le</i>	
Scigent: Multi-agent for End-to-end Question Answering on Scientific Knowledge Graph.....	192
<i>Dieu-Hien Nguyen, Nguyen-Khang Le, Shinobu Hasegawa and Le-Minh Nguyen</i>	
Learning Parameter Efficient Coordination in Autonomous Vehicles with Online Knowledge Distillation.....	198
<i>Vu Minh Le, Xuan Duy Ta and Viet Cuong Ta</i>	
Edge++: Enhanced Direction Attention via Top-k Refinement for Long-Tail Recommendation.....	204
<i>Trung-Hieu Nguyen, Minh-Khoi Le and Duc-Trong Le</i>	

Software Engineering, AI in Education, and AI Systems

EmbedRank-LM: Natural Language-Driven Web Element Identification for Automated UI Testing .	210
<i>Trinh Le Khanh, Quyen Hoang-Van and Pham Ngoc Hung</i>	
Enhancing Personalized Learning With Real-Time Data Warehousing and Linked Open Data Integration For Open Knowledge Domain.....	216
<i>Trang Pham Ngoc Thuy, Xianwen Zheng, Mau Nguyen and Shinobu Hasegawa</i>	
Finetuning LLMs for Automatic Form Interaction on Web-Browser in Selenium Testing Framework	222
<i>Nguyen-Khang Le, Nguyen Hiep, Minh Nguyen, Son Luu, Trung Vo, Quan Bui, Nomura Shoshin and Le-Minh Nguyen</i>	
Improving Adaptive Learning by Incorporating Question Attributes: A System for Special Needs Education in Vietnam.....	228
<i>Thi Quynh Hoa Nguyen, Duy Hai Nguyen, Tan Hai Nguyen, Tuan Long Tran, Thanh Trung Dang and Xuan Hien Le</i>	
iTestLayout: An Automatic Pagination Algorithm for Print-Ready Examination Papers.....	234
<i>Thanh Le Dinh</i>	
Mathematical Kanban Under Uncertainty: A Systems Engineering Approach.....	240
<i>Shee Weng Lee and Tze Houng Lee</i>	
Optimizing LLM with FP8.....	245
<i>Vinh Pham Xuan and Vinh Nguyen Van</i>	
Transforming Operationalised iStar Models into Temporal Semantics and Event-B for Consistency Assurance	250
<i>La Trinh Hoang Viet and Dang Duc Hanh</i>	

Optimization, Data Mining, and Evolutionary Computation

SAT-Based Approaches for Two-Dimensional Bin Packing: A Comprehensive Comparison with MIP and CP Methods.....	256
<i>Tuyen Van Kieu, Chi Linh Hoang and Khanh Van To</i>	
An Efficient Method for Mining Closed Frequent High Average-Utility Occupancy Patterns	262
<i>Thang La, Khanh Tran, Anh Tran, Hai Duong and Bac Le</i>	
An interactive time suggestion for multi-objective evolutionary algorithms	268
<i>Thu To Thi, Long Nguyen and Dinh Nguyen Duc</i>	
A Balanced Genetic Algorithm for Earliness-Tardiness MS-RCPSP	274
<i>Hao Nguyen Thi, Loc Nguyen The and Huu Dang Quoc</i>	
An integrated workforce planning with LLM support for blood bank operations optimization.....	280
<i>Tuong Nguyen Le and Hop Nguyen Van</i>	
Efficient Mining of Frequent High Average-Utility Occupancy Sequences	286
<i>Tien Hoang, Hai Duong, Tin Truong and Bac Le</i>	
EvoPort: An Evolutionary Framework for Portfolio Optimization via Randomized Alpha Discovery and Ensemble-Based Allocation	292
<i>Thanh Nguyen and Thi-Hau Nguyen</i>	
Simulation-Optimization for Healthcare Supplier Selection Problem under Intuitionistic Fuzzy Environment	298
<i>Ngan Nguyen Ngoc Kim, Chi Ha Thi Xuan and Truc Cao Thanh</i>	

Security, Federated Learning, and AI for Finance

Scalable Federated Learning for Video Anomaly Detection in Distributed Environments.....	304
<i>Nhat-Minh Dang, Thi-Thu-Trang Do, Quoc-Viet Hoang and Van-Quyet Nguyen</i>	
FedSAE-HS: Hybrid Anomaly Scoring for Semi-Supervised Federated Intrusion Detection	310
<i>Linh Chu Thi, Van Thi Bich Pham, Trung Ngo Van and Ly Vu</i>	

Building Reliable Covert Timing Channels via Error Correction Codes	316
<i>Hung Ngo Manh, Duc Tran, Van Tong, Giang Nguyen Linh and Tuyen Le Ngoc</i>	
Privacy-Preserving Personalized Federated Learning for Encrypted Traffic Classification	322
<i>Hai Anh Tran, Van Tong, Duc Tran and Giang Nguyen</i>	
Predicting Portfolio Price Crashes in the Vietnamese Stock Market Using Large Language Models and Deep Learning Models with Temporal Relational Reasoning	328
<i>Minh-Huong Nguyen, Thanh-Tam Do Thi, Hong-Viet Tran, Van-Tan Bui and Phuong-Thai Nguyen</i>	
A New Two-Stage Fusion-Based Approach to Cyber Threat Detection Using Graph Neural Networks	334
<i>Thu Lam Bui, Thi Thoa Vi, Cong Thanh Nguyen, Quang Long Pham, Thanh Tra Ung and Duy Trung Pham</i>	
T-GNN++: Advancing Stock Market Forecasting with Knowledge Representation and Explainable AI	340
<i>Thu Le, Nhan Vuong, Le The Dung and Bac Le</i>	
CSFS: Enhancing Industrial Threat Detection Performance with Covariate Shift Adaptation and SHAP-powered Feature Selection	346
<i>Tuyen Nguyen, Hung Nguyen, Minh Nguyen, Hai Le and Hoa Nguyen</i>	
Core Machine Learning Models and Applications	
Novel Composite Features for Detection of Arc Fault Phenomenon in Electrical Devices Using Machine Learning	352
<i>Duc Phu Nguyen, Ngoc An Nguyen and Ngoc Hoa Nguyen</i>	
A Comparative Study of Kolmogorov-Arnold Network Models for Robust WiFi Fingerprinting-based Positioning	358
<i>Luong Nguyen Thi, Nga Phan Thi Thanh, Nhan Cao Thi, Ninh Duong-Bao, Trung Vu-Thanh and Khanh Nguyen-Huu</i>	
A multi-sensor fusion approach to sensor management using void probability	364
<i>Minh Ngoc Duong and Gia Hung Hoang</i>	
Dynamic Knowledge Distillation for Climate-Driven Diarrhoea Incidence Prediction	370
<i>Thanh Do, Thuan Dinh Nguyen, Nguyen Khoi Pham, Son Thai Mai, Linh Trinh and Hans Vandierendonck</i>	
CASE-LoRA: Continual Adaptation with Structurally Evolving LoRA for Class-Incremental Learning	376
<i>Quang Dat Mac, Thanh-Nga Hoang Thi, Tri-Thanh Nguyen and Quynh-Trang Pham Thi</i>	
FDRL: Semantic-Aware Representation Learning with Feature Drift-Resistant Space Regularization	382
<i>Ngoc-Linh Nguyen, Tri-Thanh Nguyen and Quynh-Trang Pham Thi</i>	
A Comparative Analysis of Deep Learning Models for Interpretable Multi-omics Cancer Subtyping using Integrated Gradients	388
<i>Ha Tang Vinh, Hien Nguyen Minh and Diep Thi Hoang</i>	
AI-Based Modeling of Eutrophication Monitoring Using Remote Sensing and Field Data to Support Aquaculture: A Case Study of Hoa Binh Reservoir	394
<i>Ngoc Lê, Trung Duong, Cuong Do-Chi, Linh Pham, Son Nguyen, Minh Dang, Son Ta, Ha Nguyen and Thu-Huong Hoang</i>	
ELLA: Enhancing Long-Tailed Online Continual Learning with Layer Augmentation	400
<i>Dinh-Nhat Loi, Tri-Thanh Nguyen, Thanh-Hai Dang and Quynh-Trang Pham Thi</i>	
A Reliability-Enhanced Approach for Neighborhood-Aware Possibilistic Semi-Supervised Fuzzy Clustering	406
<i>Hiep Huynh, Hung Nguyen and Quang-Thinh Bui</i>	

An Enhanced Semi-Supervised Gaussian Fuzzy C-Means Clustering Algorithm for Remote Sensing Image Data Analysis	412
<i>Tuan Anh Le, Dinh Sinh Mai and Thi Lien Doanh Ha</i>	

Special Session: Intelligent Software and Knowledge Representation

Policy-Aware Generative AI for Safe, Auditable Data Access Governance	417
<i>Shames Al Mandalawi, Muzakkiruddin Ahmed Mohammed, Hendrika Maclean, Mert Can Cakmak and John Talburt</i>	
Enhancing healthcare diagnostics with attention-based multi-layer perceptron.....	423
<i>Ba-Duy Nguyen, Van-Dung Hoang and Trong-Binh Tran</i>	
Emotionally Expressive Vietnamese Text-to-Speech Synthesis Using Variational Inference Models ..	429
<i>Dung Hoang, Binh Nguyen, Hau Pham, Dung Tran and Hien Nguyen</i>	
Vehicle Re-Identification for Comprehensive Multi-Camera Tracking in Vietnamese Urban Infrastructure	435
<i>Kim-Phat Ma, Mong-Thuy Duong-Thi, Chi-Nguyen Dang, Anh Thi-Hoang Nguyen and Trong-Hop Do</i>	
Conversational Product Navigation for E-commerce via Knowledge-Based Multi-Agent Chatbots...	441
<i>Gia-Bach Truong and Tuan-Dat Trinh</i>	
Formalizing and Conformance Testing ABAC Policies in Process Logs via an ABAC Metamodel ...	447
<i>Duc-Hieu Nguyen</i>	
Group learning Integrating Behaviors and Emotions recognition to Assess Student Engagement in Classroom Context	454
<i>Hung Q. Nguyen, Dat Ly, Nha Tran, Tri Le, Hung Nguyen and Hien D. Nguyen</i>	
Vector-based Similarity for Personalized Recommendation Systems	460
<i>Nguyen Tran-Diem-Hanh</i>	

Special Session: International Symposium on Artificial Intelligence and Legal Documents

Flame Reavers@ALQAC 2025: Integrating Learned Rankers and LLM Reasoning in a Dynamic Hybrid Architecture for Legal Retrieval.....	465
<i>Huy Trieu, Dang Phuong Nam Doan, Anh Kiet Nguyen, Thanh Thai Nguyen, Thanh Nghia Vo, Tung Le and Huy Tien Nguyen</i>	
Local Focus, Global Coverage: A Hybrid Retrieval Architecture for Legal Text	471
<i>Hieu M. Huynh, Lam M. H. Nguyen, Thang Q. Pham, Khang Huy Mai and Kiet Van Nguyen</i>	
A Summary of the ALQAC 2025 Competition	477
<i>Dinh Truong Do, Thanh Son Luu and Le Minh Nguyen</i>	
H-OCR: A Hybrid OCR Approach to Vietnamese Scanned Administrative Text Recognition	483
<i>Cong Thanh Bui, The Hung Nguyen, Trung Dung Nguyen, Anh Tu Nguyen, Van Anh Nguyen, Thi Quynh Le and Xuan Duc Le</i>	
Hallucineers@ALQAC 2025: Unified LLM Pipelines for Legal Document Filtering and Question Answering	489
<i>Phong Tuan Nguyen, Quan Minh Le, Dat Tien Truong, Tung Le and Huy Tien Nguyen</i>	
Combining Large Language Model and Supportive Models for Legal Question Answering	495
<i>Long Vu Van and Hai Le Hong</i>	
UTEHY-NLU@ALQAC 2025: Dynamic Weighted Ensemble and Adaptive Reasoning for Vietnamese Legal Text Processing	501
<i>Truong-Phuc Nguyen, Quy-Nhan Nguyen, Manh-Cuong Phan, Chi-Hai Cao, Trinh-Hoai-An Duong and Minh-Tien Nguyen</i>	

Special Session: New Applications of Evolutionary Linguistics

The Evolution of Defined Terms in Law: A Diachronic Study Using Open Legal Data	507
<i>Makoto Nakamura and Daichi Saito</i>	
Analysis of Section-Wise Syllable Usage and Its Relationship to Instrumentation in J-POP A Cappella Arrangements.....	513
<i>Kohei Tsuchida, Shun Sawada, Kouichi Katsurada and Hidefumi Ohmura</i>	
Personalized Tonal Harmony Parsing via Scoring-Based Tree Selection	519
<i>Daichi Hoshikoshi, Shun Sawada, Kouichi Katsurada, Satoshi Tojo and Hidefumi Ohmura</i>	
Special Session: AI for Security, Security for AI	
Enhancing Multivariate Time Series Anomaly Detection with Feature-Scaled Reconstruction in LSTM Autoencoders	525
<i>Truong Son Pham, Quang Phong Nguyen and Do Phuong Nguyen</i>	
Improving Targeted Password Guessing with Pattern-Guided Vector Quantization	531
<i>Nguyen Duc Minh and Nguyen Viet Hung</i>	
Pri-Inv: A Novel Privacy-Preserving Digital Forensic Solution for Face Presence Verification in Criminal Investigations.....	537
<i>Anh Tu Tran and Xuan Hung Truong</i>	
MAML AE-SAC: A Synergy framework of Meta-Learning and Adversarial Reinforcement Learning for Adaptive Intrusion Detection	544
<i>Nhat Duy Dang, Giang Nguyen, Van Dong Le, Trong Tung Bui and Tuyen Ngoc Le</i>	
A Deep Self-Refining Approach for Fact Verification	550
<i>Cuong Hoang, Vu Tran and Le-Minh Nguyen</i>	

Keynotes



Professor Satoshi Tojo,

Asia University, Japan

Awareness Logic and Elementary Geometry

Abstract:

Epistemic logic is an application of modal logic, where human knowledge and belief are modeled in terms of Kripke semantics. Among which, the logical omniscience, that derives every tautology with repetitive modus ponens (MP), has been an annoying issue. Fagin and Halpern (1988) tackled the issue, however, the MP in implicit knowledge remained unsolved. In this research, we introduce the notion of epistemically indistinguishable worlds and ease the requirements for implicit knowledge, and with this, reconstruct the awareness logic. We consider an example of elementary geometry and explain why limited students can solve the problems though every fundamental geometric knowledge is evenly given.

Bio:

Prof. Satoshi Tojo earned his B.E., M.E., and Ph.D. degrees from the University of Tokyo. From 1983 to 1995, he was a researcher at the Mitsubishi Research Institute, Inc. He joined the Japan Advanced Institute of Science and Technology (JAIST) in 1995 as an Associate Professor in the School of Information Science, was promoted to Full Professor in 2000, and served there until 2023. He is currently Professor of Data Science at Asia University, Japan. His research interests include formal semantics of natural language, the logic of knowledge and belief in agent communication, language evolution, and linguistic analysis in music. Prof. Tojo has authored over 250 papers in leading international journals and conferences, delivered keynote speeches at more than 28 international conferences, and co-edited five books.



**Professor Thomas Ågotnes,
University of Bergen, Norway**

Anonymous Public Announcements

Abstract.

Taken at face value, the title of this talk seems to be an oxymoron. Indeed, if public announcement is taken literally, as in an agent saying something in front of everyone else, it will not be anonymous. However, anonymous public communication is almost ubiquitous in our day-to-day lives. Think of posts on social media and message boards done under a username instead of a real name of the poster. Or an anonymous letter to an editor of a news outlet. Other examples include anonymous emails, transactions on a blockchain, whistle blower reports, and even cultural artifacts created anonymously under an alias, like Elena Ferrante and MF DOOM.

This talk will be about formalising anonymous public announcements, using epistemic logic (the standard framework for reasoning about knowledge). Even if such an announcement is *prima facie* anonymous, depending on the background knowledge of the agents it might reveal the identity of the announcer: if I post something on a message board, the information might reveal who I am even if I don't sign my name. Furthermore, like in the Russian Cards puzzle, if we assume that the announcer's intention was to stay anonymous, that in fact might reveal more information. In this talk I first look at the case when no assumption about intentions are made, in which case the logic with an anonymous public announcement operator is reducible to epistemic logic. I then look at the case when we assume common knowledge of the intention to stay anonymous, which is both more complex and more interesting: in several ways it boils down to the notion of a "safe" announcement (again, similarly to Russian Cards). Main results include formal expressivity results and axiomatic completeness for key logical languages. The talk is based on joint work with Rustam Galimullin, Ken Satoh and Satoshi Tojo.

Bio.

Thomas Ågotnes is a Professor of Information Science and head of the Logic and AI research group at the Department of Information Science and Media Studies, University of Bergen, Norway. He is also a visiting Professor at the Centre for the study of Logic and Intelligence at Southwest University, China. His main research interests include formal knowledge representation and reasoning about different types of interaction, in particular using modal logic, often combined with other mathematical models of interaction for example from game theory, with applications in the fields of artificial intelligence and multi-agent systems. He has published extensively in these fields. He has been awarded the Changjiang (Yangtze river) Scholar award by the Chinese government, and best paper awards at conferences such as AAMAS and CLAR.



Prof. Philippe Fournier-Viger,
Shenzhen University, China

*Advances and challenges for the automatic
discovery of interesting patterns in data*

Abstract:

Intelligent systems and tools can play an important role in various domains such as for factory automation, e-business, transportation, and e-learning. The foundation of these advancements lies in the availability of high-quality data, encompassing diverse forms such as activity logs, multimedia content, and sensor-derived metrics from interactive educational environments. Managing data to gain insights and improve these systems is thus a key challenge. It is also desirable to be able to extract information or models from data that are easily understandable by humans. Based on these objectives, this talk will discuss the use of data mining algorithms for discovering interesting and useful patterns in data generated from intelligent systems.

The talk will first briefly review early study on designing algorithms for identifying frequent patterns. Then, an overview of recent challenges and advances will be presented to identify other types of interesting patterns in more complex data such as graphs and sequences. Topics that will be discussed include high utility patterns, locally interesting patterns, periodic patterns and spatial patterns. Lastly, the SPMF open-source software will be mentioned and opportunities related to the combination of pattern mining algorithms with conventional artificial intelligence techniques to enhance the efficacy of educational intelligent systems.

Bio:

Philippe Fournier-Viger (Ph.D) is distinguished professor at Shenzhen University (China). Five years after completing his Ph.D., he came to China in 2015 and became full professor after receiving an important national talent title. He has published more than 400 research papers related to data mining algorithms for complex data (sequences, graphs), intelligent systems and applications, which have received more than 16,000 citations (H-Index 63 - Google Scholar). He is the founder of the popular SPMF data mining library, offering more than 260 algorithms to find patterns in data, cited in more than 1,000 research papers. He is former associate editor-in-chief of the Applied Intelligence journal and has been keynote speaker for over 50 international conferences and co-edited four books for Springer. He appears in the top 0.3% of researchers for scientific influence in the Stanford list. He won the "Most Influential Paper Award" at the 2024 PAKDD conference and received seven Best Paper Awards at international conferences. Website: <http://www.philippe-fournier-viger.com>.



**Prof. Makoto Nakamura, Niigata
Institute of Technology, Japan**

*Legal AI: Infrastructure, Understanding,
Validation, and Support for Evolving Law*

Abstract:

Legal Artificial Intelligence (Legal AI) covers a broad range of approaches to making law computable, verifiable, and supportive of human decision-making. This talk presents four major categories that together form the foundation of Legal AI research.

(A) Legal Data Infrastructure focuses on the structuring and dissemination of legal information through XML-based frameworks such as Akoma Ntoso and domain-specific machine translation.

(B) Legal Understanding explores logical formalization and semantic mapping of statutes, enabling explicit reasoning and cross-jurisdictional comparison.

(C) Legal Validation & Simulation evaluates the logical and social effects of legal reforms using methods ranging from symbolic inference to multi-agent modeling.

(D) Legal Support & Evolution extends AI to assist in legislative drafting and to analyze diachronic changes in legal terminology.

Together, these four perspectives illustrate how AI can transform law into a transparent, testable, and adaptive system, one that not only represents norms but also learns from the way societies evolve.

Bio:

Prof. Makoto Nakamura received his Ph.D. in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2004. He served as Assistant Professor at JAIST (2004–2011), then held positions at the Institute of Legal Information, Graduate School of Law, Nagoya University, as Assistant Professor (2011–2016) and Associate Professor (2016–2018), before joining Niigata Institute of Technology as Associate Professor (2018–2021). He is currently Professor in the Faculty of Engineering at Niigata Institute of Technology. Prof. Nakamura has been a member of the scientific committee for the Law via the Internet (LVI) conference since 2016 and a steering committee member, and past co-chair, of the International Workshop on Juris-Informatics (JURISIN) since 2009. His research focuses on natural language processing for legal texts and the simulation of language evolution. He is an active member of ANLP, IPSJ, JCSS, and JSAL.



**Invited Speaker: Prof. Hidefumi
Ohmura,
Tokyo University of Science, Japan**

*The Fundamental Structure of Music for
Music Evolution Research*

Abstract:

Music is not merely a sequence of sounds but is composed of specific sound structures. These structures have evolved over time, giving rise to diverse forms of music. Music is similar to language in this regard. Interestingly, these fundamental structures appear to be innate in humans, even without specialized musical education. These fundamental musical structures, which elevate mere sound into music, are crucial for exploring music's origins and understanding its process of evolution. I have focused on these basic structures in attempts to generate music. This generation system can reproduce the structures that form the core of many music theories. However, it does not include some structures found in music or aspects of musical evolution. I will present how music is formed based on this fundamental structure and discuss prospects for music evolution research.

Bio:

Hidefumi Ohmura received Ph.D. in engineering from Tokyo Institute of Technology, Japan. He worked as a researcher at Japan Science and Technology Agency, RIKEN, and National Center of Neurology and Psychiatry. He currently works as a junior associate professor to develop musical systems for human emotion and intuition at Tokyo University of Science. His research interests include not only musical structure analysis, but also musical modeling, based on computer science.



Invited Speaker: Dr. Dao Minh Son

**Big Data Integration Research
Center, NICT, Japan**

*Leveraging Multimodal AI Communication
and Edge Intelligence for Smart Mobility and
Intelligent Transportation Systems*

Abstract:

As transportation systems evolve toward connected, data-rich environments, artificial intelligence must interpret diverse multimodal signals, from cameras and sensors to traffic data and human behavior. This talk introduces a new framework for multimodal AI communication that integrates perception, reasoning, and decision-making across modalities to support smart mobility and intelligent transportation systems (ITS). Built on foundation models and advanced digital twin environments, our approach enables adaptive, context-aware analytics for real-world deployment. By leveraging edge intelligence, models can operate locally to protect personal information and reduce central computing costs, aligning with global trends in decentralized AI. The talk will also discuss how this architecture supports cross-domain data integration and continuous learning, paving the way for resilient, privacy-aware AI systems in next-generation ITS.

Bio:

Dr. Minh-Son Dao is Research Manager at the Big Data Integration Research Center, NICT (Japan). His research lies at the intersection of multimedia signal processing, intelligent sensing, and large-scale data analytics. With a career spanning in Europe and Asia, he has contributed extensively to multimedia event detection, cross-domain analytics, and cyber-physical-social systems (CPS).

Dr. Dao's work bridges theory and practice, advancing from early studies in video object retrieval and shape matching to AI-driven, data-integrated platforms for smart-city and Society 5.0 applications. His current research explores generative AI content detection, cross-data analytics, and scalable multimodal systems that integrate visual, textual, and sensor data.

He has authored numerous papers in top-tier international conferences and journals (including ACM MM, IEEE Transactions, and MMM) and his work has been widely cited for its contributions to multimedia understanding and data-driven intelligence. His achievements have earned several recognitions, such as the Best Industry Demo Award at ACM Multimedia 2025, the Best CheapFakes Detection Award, and the Best Social Impact Award at MediaEval.

Dr. Dao is also an active contributor to the global multimedia research community, serving program committees for ACM MM, MMM, ICMR, MediaEval, and Media Verification (Cheapfakes).

An Investigation of Collocation Segmentation Approaches in Neural Machine Translation

Van-Khanh Tran*

Institute of Applied Science and Technology - IAST
University of Information and Communication Technology - ICTU
Thai Nguyen, Vietnam
tvkhanh@ictu.edu.vn

Huu-Tri Tran

Hanoi University of Science and Technology
Hanoi, Vietnam
tritrhanhuu789@gmail.com

Abstract—Neural Machine Translation (NMT) has made significant progress by taking advantage of subword segmentation to address the open-vocabulary problem. While Byte Pair Encoding (BPE) deterministically splits words into unique sequences, stochastic subword segmentation methods (e.g., subword regularization (SR) and BPE dropout) show the ability to produce multiple rational segmentation of the same word, hence increase model robustness. However, previous subword segmentation methods primarily work on the word level, leading to being restrictive and ambiguous. Those techniques learn the compositionality of the units up to word level but cannot model the lexical units beyond words, such as collocations, multiword expressions, or phrases. This paper introduces Collocation Dropout, a simple and effective segmentation method that allows NMT models to learn to translate a collocation or a phrase better. We conducted experiments on various typical NMT tasks: WMT14 English-German, ASPEC English-Japanese, IWSLT15 English-Vietnamese, English-Chinese, and IWSLT17 English-German, English-French to demonstrate the effectiveness of the proposed method. Experimental results also show that combining our method and other subword segmentation, including both deterministic and stochastic methods, further improves NMT systems.

I. INTRODUCTION

Subword segmentations have been widely used in NMT to alleviate open-vocabulary issues. The core idea of subword segmentation is that frequent words should be kept while rare words should be spitted into rational subword units. This allows NMT models to maintain a reasonable vocabulary size while addressing out-of-vocab words by decomposing them into known subwords.

There are two types of subword segmentation methods: deterministic and stochastic. Byte pair encoding (BPE) [1], the most popular deterministic subword segmentation method, iteratively creates new subword units by merging the most frequent pair of symbols until reaching the desired number of symbols. BPE deterministically decomposes words into unique subword sequences, which means for each word, the NMT model can observe only one segmentation. However, BPE subword segmentation is potentially ambiguous since a sentence can be represented in multiple segmentation even with the same vocabulary.

Stochastic subword segmentation methods can handle aforementioned problem and make NMT models more accurate and

robust. While subword regularization (SR) [2] enables multiple segmentation candidates with subword sampling during training, BPE-Dropout [3] provides multiple subword segmentation using a formulation to remove merges from the merging table. Adversarial Subword Regularization (ASR) [4] utilizes an adversarial NMT training framework to sample subword segmentations that maximize the translation loss. Dynamic programming encoding (DPE) [5] leverages dynamic programming for marginalizing the latent segmentations and producing the highest probability segmentation when tokenizing.

Previous subword segmentation methods primarily work on the word level despite its success. The model only splits words into subword units (morphology), lacking the compositionality of units beyond words. For example, consider translating sample from English to French as in Table I. Current subword segmentation methods learn on raw texts only and encode a source English sentence "If you actually want to make a difference , you have to do something ." to subword units sequence " I f you actually want to make a difference , you have to do something .", without phrase-based knowledge.

In this study, we proposes a new segmentation method that works on up to phrase level (collocation). The segmentation method decomposes words into smaller subwords and split phrases into rational words and subword units. For example, our segmentation first extract collocation from corpus and transform input sentence into sentence with collocation information, i.e., "If you [actually_want_to] [make_a_difference] , you [have_to_do] something .". Our segmentation methods learn from new collocation corpus and encode sentence as " I f y ou [actually _ w a n t _ to] [make _ a difference] , you [have _ to _ do] something .". This allows NMT model to learn to translate multi-word expressions. Further, to enable the multiple segmentation at phrase-level, we propose collocation dropout, which randomly drops some merges with the probability p .

Phrases are important for machine translation since the meaning of a phrase cannot always translate from the meaning of its words [6], [7], [8], [9]. Unfortunately, prior NMTs mostly work on word-based, subword-based, or character-based, which do not explicitly treat phrases. In contrast, phrase-based statistical machine translations (SMT) [6], [7] have made a consistent improvement in translation quality

*Corresponding author: tvkhanh@ictu.edu.vn

Raw data	Source Target	If you actually want to make a difference , you have to do something . Pour faire une différence , vous devez faire quelque chose .
BPE segmentation	Source Target	I f you actually want to make a difference , you have to do something . P our faire une différence , vous devez faire quelque chose .
Collocation data	Source Target	If you [actually _ want _ to] [make _ a _ difference] , you [have _ to _ do] something . Pour [faire _ une _ différence] , vous [devez _ faire] [quelque _ chose] .
Our segmentation	Source Target	I f y ou [actually _ w a n t _ to] [make _ a _ difference] , you [have _ to _ do] something . P our [faire _ une _ d i f f é r e n c e] , vo u s [devez _ f a i r e] [quel _ que _ chose] .

TABLE I: Raw versus Collocation text and segmentation between BPE segmentation and our segmentation. and _ denoted sentencepiece splitter and collocation link, respectively. Each collocation is bounded by [and] for better visualization.

comparing to word-based ones. To combine benefits of phrase-based SMT and NMT, [10], [9] proposed phrase-based NMTs, which combine word-based NMT with external phrase memories. However, current phrase-based NMT methods introduce high additional resources and computations into the network. This study keeps the NMT model intact while only integrating collocation information into the subword segmentation phase. The proposed segmentation method, therefore, is easy to combine with previous subword segmentation methods.

In summary, our contribution is threefold:

- We introduced a simple way to learn collocation segmentation for NMT;
- We introduced collocation dropout technique which enables stochastic segmentation for collocations.
- We intensively conducted experiments on various typical NMT tasks: WMT14 English-German, ASPEC English-Japanese, IWSLT15 English-Vietnamese, English-Chinese, and IWSLT17 English-German, English-French to demonstrate the effectiveness of the proposed method.

II. APPROACH

A. Collocation Extraction

We show that an NMT model trained with collocation extracted data can translate all the meaning of a phrase, reduce word-by-word error. Collocation extracting is to find n-grams that have semantic relationships between them. They can be extracted statistically by T-test.

Using a initial hypothesis H_0 : a randomly constructed n-gram $w = x_1, x_2, \dots, x_n$ has a probability $p'(w) = \prod_{i=1}^n p(x_i)$. For every n-grams in the training data X, we compute t-statistic scores.

$$s = \frac{(p(w) - p'(w))}{\sqrt{\sigma^2 / \text{Count}(N)}}$$

where $p(w) = \frac{\text{Count}(w)}{\text{Count}(N)}$, deviation $\sigma^2 = p(w)(1 - p(w))$, N denotes the total number of n-grams in X.

The higher the score is, the more likely the n-grams is a collocation. We pre-define a maximum value for n . For example, with $n_{max} = 5$, we calculate t-score for $n = 2$, $n = 3$, and so on. Then collocations extraction can be applied for the dataset by a maximum matching algorithm [11] with the extracted dictionary. According to the t-statistic scores, we select the top k n-grams with the highest score to construct

a specific vocabulary of collocations (k is proportional to the training corpus size). Table II shows the ratio of collocations for some language pairs. For English, 10-11% of text elements are collocations, the pIntroductionroportion increases more than twice for other languages. This significant difference can be foreseen since we apply stricter collocation filter rules for English than the others. Moreover, collocation rate for all datasets remain stable using our extracting strategy.

In this study, we run the collocation extraction algorithm on the training set of the parallel corpora. In that way, the number of collocation in the dataset can be increased significantly, which leads to an increasing of its influence. Moreover, it is hard to verify the quality of n-grams since many n-grams are extracted as collocation, but they happen to appear a lot and carrying no or the wrong meaning. To enhance collocation quality, we defined set of filter rules with details is shown in the following section.

Filter Rules for Collocation Extraction : We eliminated all n-grams that meet one of the following conditions:

- first token is stopword ¹ (e.g. *to go, a car, ...*)
- contains digit or special symbols (e.g. *1 USD, 2 times, ...*)
- contains pronouns (e.g. *i want, meet you, ...*)
- contains verb **be** (e.g. *man is, are good, ...*)

We applied these rules fully for the English side of our corpora. For other languages, since lacking knowledge about them, we only used the first and second rules. With an input text, we split it into many parts by punctuation and start picking n-grams. All n-grams are filtered still be counted in the number of n-grams in corpora when computing the t-statistic score.

Collocation Extraction for Chinese and Japanese: To represent a collocation in text, we link its words in semantically complete n-grams with a special symbol "_". For example, we represent collocation "*make a difference*" in text as "*make_a_difference*". However, Chinese and Japanese have no explicit word boundaries; we have to modify our Collocation Extraction procedure. Unlike English, or other Latin scripted languages, a character in Chinese and Japanese has its meaning. While extracting, we split the text into characters and treat them as words, which means collocation extraction algorithm is applied for Chinese and Japanese at character level. Anyway, there are Latin characters in Chinese

¹stopwords lists are provided on <https://github.com/stopwords-iso>

	IWSLT15				IWSLT17			
	En-Vi		En-Zh		En-Fr		En-De	
	En	Vi	En	Zh	En	Fr	En	De
<i>train</i>	10.5%	21.1%	10.4%	26.2%	11.1%	22.3%	10.7%	21.2%
<i>dev</i>	11.0%	19.6%	9.9%	23.4%	10.9%	20.8%	10.6%	21.3%
<i>test</i>	10.0%	18.1%	10.0%	25.7%	10.6%	21.1%	10.6%	21.2%

TABLE II: Collocation ratio

or Japanese text, which must be tokenized into words in the usual way. For representing, collocations in these language can be shown in reverse way. Every token in a collocation are placed sequentially and collocations are separated by the symbol "_". For example, the collocation extraction output of "然后你要为他们提供专注而热情的服务。" is "然后你_要_为_他_们_提供_专注而_热情的_服务。".

B. Collocation Dropout

We propose Collocation Dropout in order to address the remaining problem in Collocation Extraction. It increases model's robustness in handling miss-extracted collocations. Moreover, it can be considered as a high level on-the-fly training method that can combine with any other subword segmentation including BPEDropout or subword regularization.

This method can only applied to the model training with collocation extracted data. Before segmentation, input collocations are randomly recovered back to their original words form with a probability p . This procedure is described in Algorithm 1.

Algorithm 1: Collocation Dropout

Input: Collocation extracted sentence S , dropout probability p

Output: Collocation dropout subword sequence x

```

1  $x = []$ 
2 for  $w$  in  $S$  do
3   if  $w.is\_collocation()$  then
4     for  $link$  in  $w$  do
5        $\_\_$  drop  $link$  with a probability  $p$ 
6      $\_\_$  update  $w$ 
7    $x.extend(subword\_segment(w))$ 
8 return  $x$ 

```

If p is set to 1, the model is equivalent to the one that train with original data. If p is set to 0, all the collocations will be preserved. In training time, we use $p > 0$ and during inference, p is set to 0, which means model inferences on base collocation extracted data. We set $p = 0.2$ in the training stage for all our experiments. However, choices of collocation dropout rate are discussed in Section IV-C.

III. EXPERIMENTS

A. Datasets

We conducted our experiments on multiple datasets in wide range of sizes and languages. Data statistics are summarized

Corpus		Number of sentences (train/dev/test)		Voc size	Col size
IWSLT15	En $\leftrightarrow$ Vi	133k / 1553 / 1268	8k	12k / 12k	
	En $\leftrightarrow$ Zh	209k / 887 / 1261	8k	15k / 15k	
IWSLT17	En $\leftrightarrow$ Fr	232k / 890 / 1210	8k	15k / 15k	
	En $\leftrightarrow$ De	209k / 888 / 1080	8k	15k / 15k	
ASPEC	En $\leftrightarrow$ Ja	2M / 1700 / 1812	16k	60k / 60k	
WMT14	En $\leftrightarrow$ De	4.5M / 3000 / 3003	32k	200k / 200k	

TABLE III: Data Statistics

in Table III. IWSLT15/17² are low-resource corpora where morphological variations that unseen can occur regularly outside the training dataset. We also evaluate NMT models on large corpora: ASPEC (En-Ja)³ and WMT14 (En-De)⁴. For ASPEC, we only use the first 2M parallel sentences in the training set.

B. Data Preprocessing

Before training, we first prepared a collocation extracted corpus for all language pairs with $n_{max} = 3$. We used joined dictionary for all language pairs, subword vocabulary and also collocation dictionary sizes are provided in Table III. For Chinese and Japanese, we modified the collocation extraction procedure as mentioned in section II-A. Dataset is preprocessed with the standard Moses toolkit. All bigrams or trigrams extracted as a collocation in the dataset are linked by a special symbol (we used " $\_\_$ ") instead of white spaces.

In Collocation Dropout, to address out of vocab problem when the NMT model randomly segment a collocation back into words, the dictionary is built with modified data. Concatenating collocation extracted and original training data can leads to the expansion and duplication of non-collocation elements which harm collocation's influence. Alternatively, all collocations in training data are filtered and concatenated with the original one to use as subword segmentation training corpus. Hence, a collocation that is dropped during training stage can still be segmented efficiently.

C. Experimental Setups

We implemented NMT model based on Fairseq⁵. Our implementation contains subword regularization, BPE Dropout and Collocation Dropout. To ensure the right idea of on-the-fly training, we re-segment all sentences in training data in each iteration. Re-segment procedure includes dropping

²<https://iwslt.org/>

³<https://jipsti.jst.go.jp/aspec/>

⁴<http://statmt.org/wmt14/>

⁵<https://github.com/pytorch/fairseq>

collocations and subword segmentation stochastically. We used collocation dropout rate $p = 0.2$ for all experiments.

For training, we used Transformer-base as backbone model for our experiments. More precisely, there are $N = 6$ layers and $h = 8$ attention heads. The dimensionality of input and output is $d_{model} = 512$, and the inner-layer of feed-forward networks has dimensionality $d_{ff} = 2048$. We follow the same regularization and optimization procedures as described in [12]. We trained all models on lowercase data.

D. Evaluation

For evaluation, we used the checkpoint that achieves the best performance on validation data and used SacreBleu [13] to compute BLEU score. For Chinese and Japanese we set `-tok zh` and `-tok ja-mecab` option respectively. Our baselines are NMT models trained with standard BPE segmentations, subword regularization (SR) and BPE Dropout. We used the same hyperparameters as in [2] and [3]’s. To produce translations, we used beam search with beam size at 4, and length normalization of 0.6. For subword regularization, we set $l_{best} = 64$ and $\alpha = 0.1$. For BPE Dropout, we set dropout rate $p_{bpe} = 0.1$ for every language except for Chinese, and Japanese were set at $p_{bpe} = 0.6$ to achieve the same length increase.

IV. RESULTS

A. Main results

Table IV shows the main results on low-resource datasets (IWSLT15, IWSLT17) as well as medium-resource ones (ASPEC and WMT14). All scores whose difference from the best is not statistically significant (with p-value of 0.05) [14]. Our segmentation method brings significant benefits for baselines (BPE, SR, and BPEDropout) for all language pairs and datasets since most of the best and second best BLEU scores come from our proposed methods. This demonstrates that subword segmentation methods on phrase level is more effective than previous methods on the word level only.

Results on Low-resource Datasets: For En-Vi direction, BPE+Collo and BPE+ColloDropout improved 1.4 BLEU and 1.5 BLEU, respectively, over the baseline BPE. Specifically, BPEDropout+ColloDropout improved 0.9 BLEU over SR and 3.3 BLEU over BPE. Our models outperform the state-of-the-art BPEDropout up to 0.6 BLEU in the En-Zh dataset where SR+ColloDropout achieved BLEU scores at 22.5 (En-Zh) and 21.3 (Zh-En) while BPDDropout got BLEU scores only at 21.9 (En-Zh) and 20.7 (Zh-En), respectively. Further improvement up to 0.8 BLEU comes from the Fr-En dataset where BPEDropout+ColloDropout and BPEDropout performed at 38.3 BLEU and 37.5 BLEU, respectively. A noticeable fact from [3] is that compared to SR segmentation, BPEDropout is not significantly different for three datasets Ja-En, En-Zh, and Zh-En, and has lower performance for the En-Ja dataset. However, training with the collocation segmentation method, BPEDropout+ColloDropout performed slightly better than the SR in the En-Zh and Zh-En datasets.

Results on Medium-resource Datasets: Table V shows BLEU scores of models trained on large datasets. The trend is consistent with the low-resource trend since adding collocation information brings significant benefits for the NMT models. ASPEC dataset saw the largest improvement at 2.1 BLEU in the Ja-En direction from 28.5 BLEU (BPE) to 30.6 (SR+ColloDropout), while WMT14 saw a second largest improvement at 1.8 BLEU score in the De-En direction from 29.3 (BPE) to 31.1 (BPE+ColloDropout). This demonstrates the effectiveness of our proposed collocation segmentation methods for large datasets.

B. Collocation Extraction Sides

We discuss whether collocation should be applied to both source and target text, or only one side. We trained NMT model with standard BPE in 3 scenarios by using collocation on both sides, source-side only, and target-side only. We conducted experiments in IWSLT’s En-Vi and En-De datasets. Table VI shows that NMT model trained with both sides of collocation extraction performs best, while training with collocation extracted on the source side only does not see a significant difference. Moreover, using collocation on the target side is not very effective.

Since models trained on both sides collocation extraction perform best, for all experiments we use our collocation extraction on both source and target sides.

C. Choice of Collocation Dropout Rate

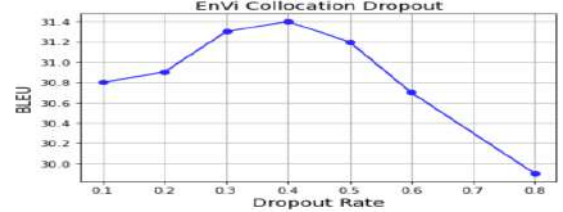


Fig. 1: BLEU scores of the BPE+ColloDropout model trained on En-Vi dataset with difference collocation dropout rates.

Figure 1 shows BLEU scores for the BPE+ColloDropout model trained on the En-Vi dataset with various values of p ranging from 0.1 to 0.8. Models trained with a high value of p performed worse since most of the collocation are split back to words. The best performance is achieved with $p = 0.4$.

Default collocation dropout rate p is set at 0.2 for all experiments. We did not extensively search for the best choice of p for all languages. Notes that collocation dropout can be applied for both training and inference time. We leave choosing the best option of collocation dropout rate p for future research, where we can find an appropriate value of p for each language pair depending on language characteristics.

V. ANALYSIS

This section analyzes the qualitative difference between models train on word level and phrase level segmentation on encoding text into subword units and example translations.

	IWSLT15				IWSLT17			
	En-Vi	Vi-En	En-Zh	Zh-En	En-Fr	Fr-En	En-De	De-En
BPE	29.5	28.2	20.8	19.4	38.2	37.3	29.0	32.2
BPE+Collo	30.9	28.3	21.3	19.8	38.5	37.5	29.6	32.3
BPE+ColloDropout	31.0	28.5	21.4	19.5	38.6	37.5	29.4	32.5
SR	31.9	28.7	22.1	20.9	38.9	37.6	29.7	32.6
SR+Collo	32.4	28.8	22.3	21.1	39.1	37.8	30.0	32.9
SR+ColloDropout	32.5	29.1	22.5	21.3	39.2	38.2	30.2	32.8
BPEDropout	32.3	29.2	21.9	20.7	39.0	37.5	29.3	32.5
BPEDropout+Collo	32.6	29.4	22.1	21.0	39.2	38.0	29.7	32.7
BPEDropout+ColloDropout	32.8	29.5	22.3	21.2	39.2	38.3	30.3	32.9

TABLE IV: BLEU scores on the main results. **Bold** and *italic* indicate the **best** and *second best* scores.

	ASPEC		WMT14	
	En-Ja	Ja-En	En-De	De-En
BPE	45.0	28.5	25.9	29.3
BPE+Collo	45.6	29.7	27.4	30.5
BPE+ColloDropout	45.7	30.4	27.1	31.1
SR	45.3	30.0	26.3	29.9
SR+Collo	46.2	30.2	26.8	30.6
SR+ColloDropout	45.7	30.6	27.0	30.8
BPEDropout	45.5	29.5	26.2	30.6
BPEDropout+Collo	45.6	29.6	27.0	30.9
BPEDropout+ColloDropout	45.8	30.0	27.1	30.9

TABLE V: BLEU scores on medium-resource datasets. **Bold** and *italic* indicate the **best** and *second best* scores.

	<i>src-only</i>	<i>tgt-only</i>	<i>both</i>
En-De	29.3	28.8	29.6
En-Vi	30.8	29.7	30.9

TABLE VI: BLEU scores of BPE models on three scenarios of applying collocation.

Encoding Text into Subword Units: Table VII shows subword units split by various subword segmentation methods at word level (BPE, SR, and BPRDropout) and phrase level (rest models). We found that our segmentation can effectively work as a hierarchical level of word and phrase. During training on the fly, NMT can observe multiple subword segmentation of phrases and words. For example, to encode a phrase [*make_a_difference*], SR+Collo can keep it as a whole phrase, SR+ColloDropout can drop all collocation links _ into a list of individual words [*make*, *a*, *difference*], BPEDropout+Collo keeps phrase info but further splits word into subword units [*make*, *_*, *a*, *_*, *diff*, *er*, *ence*].

Example Translations: Table VIII shows translation outputs of NMT models trained on the Vi-En dataset. Phrases denoted by italic words separated by _ token. Without phrase information, standard BPE generates misleading phrase *first_started*, while by training on additional phrase information, both BPE+Collo and BPE+ColloDropout can produce acceptable phrases *first_worked* and *first_job*, respectively. This again confirms the effectiveness of learning collocation while translating.

VI. RELATED WORKS

Close to our work are prior subword segmentation methods, including BPE [1], SR [2], and BPEDropout [3]. To handle the OOV problem, BPE creates a vocabulary consisting of all symbols that frequently occur as unique words and iteratively learns to merge two symbols to create a new symbol until reaching the desired vocabulary size. The drawback of BPE is that it is a deterministic segmentation method which means for each word, BPE splits it into one segmentation only. To enable multiple segmentation candidates, [2] introduced a subword regularization (SR) framework with a new segmentation algorithm. [3] proposed BPEDropout by exposing a model to different segmentation based on the BPE framework.

Phrases are crucial for the translation quality of NMT models. Phrase-based statistical machine translations (SMT) have made consistently an improvement in translation quality comparing to word-base ones [7], [8], [6]. There have been several studies that propose different ways to incorporate phrases into attention-based NMT, such as [10], [15], [16], [17], [18] [16] explicitly used phrasal translations and language model scores in an integrated hybrid search. [15] proposed to integrate a phrase memory storing target phrases from a phrase-based SMT system into the encoder-decoder architecture of NMT. [4] proposed phrase-based NMT system built upon continuous-output NMT models.

VII. CONCLUSIONS

This paper introduces a new segmentation method that allows NMT models to work up to phrase-level. We firstly extract collocations from the corpus, which are then added to the original dataset to learn subword segmentation. Secondly, we introduce collocation dropout to drop collocations back to words during model training randomly. NMT models trained with proposed subword segmentation units outperform previous subword segmentation on a wide range of translation tasks. Future research directions include integrating phrases between source and target languages with their relationships (i.e., alignment, syntactic information) and in-depth analysis of learned tokens for different segmentation.

Raw	If you actually want to make a difference , you have to do something .
Collocation	If you actually_want_to make_a_difference , you have_to do something .
BPE	I f you actually want to make a difference , you have to do something .
SR	I f you actual y want to make a difference , you have to do something .
BPEDropout	I f y ou actually w ant to make a difference , you ha v e t o do s om et hing .
SR+Collo	I f you actually _ want _ to make _ a _ difference , you have _ to _ do something .
SR+ColloDropout	I f you actually _ want _ to make a difference , you have to _ do something .
BPEDropout+Collo	I f you actually _ w an t _ to make _ a _ diff er ence , you have _ to _ do something .
BPEDropout+ColloDropout	I f y ou actually _ w a n t _ to make _ a difference , you have _ to _ do something .

TABLE VII: Subword units split by various subword segmentation methods at word level (row 2) and phrase level (row 3). and _ denoted sentencepiece splitter and collocation link, respectively.

Source	Tôi chuy ền đ ến thành_ph New_York khi làm công_vic đ ầu_ti ền vit và bi ến_tp cho tp_chí Seventeen .
Target	I had moved_to New_York_City for my first_job as a writer and editor at Seventeen magazine .
BPE	I moved_to New_York city when I first_started writing and edited for Seventeen magazine .
BPE+Collo	I moved_to New_York_City when I first_worked and edited for Seventeen magazine .
BPE+ColloDropout	I moved_to New_York_City when I did my first_job writing and editing for the Seventeen magazine .

TABLE VIII: Vi-En example translations of models trained on different subword segmentation.

REFERENCES

- [1] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: <https://www.aclweb.org/anthology/P16-1162>
- [2] T. Kudo, “Subword regularization: Improving neural network translation models with multiple subword candidates,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 66–75. [Online]. Available: <https://www.aclweb.org/anthology/P18-1007>
- [3] I. Provlkov, D. Emelianenko, and E. Voita, “BPE-dropout: Simple and effective subword regularization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1882–1892. [Online]. Available: <https://www.aclweb.org/anthology/2020.acl-main.170>
- [4] J. Park, M. Sung, J. Lee, and J. Kang, “Adversarial subword regularization for robust neural machine translation,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1945–1953. [Online]. Available: <https://www.aclweb.org/anthology/2020.findings-emnlp.175>
- [5] X. He, G. Haffari, and M. Norouzi, “Dynamic programming encoding for subword segmentation in neural machine translation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 3042–3051. [Online]. Available: <https://www.aclweb.org/anthology/2020.acl-main.275>
- [6] R. Zens, F. J. Och, and H. Ney, “Phrase-based statistical machine translation,” in *Annual Conference on Artificial Intelligence*. Springer, 2002, pp. 18–32.
- [7] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.
- [8] N. Bertoldi, M. Barbaiani, M. Federico, and R. Cattoni, “Phrase-based statistical machine translation with pivot languages,” in *International Workshop on Spoken Language Translation (IWSLT) 2008*, 2008.
- [9] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, “Phrase-based & neural unsupervised machine translation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 5039–5049. [Online]. Available: <https://www.aclweb.org/anthology/D18-1549>
- [10] Y. Tang, F. Meng, Z. Lu, H. Li, and P. L. Yu, “Neural machine translation with external phrase memory,” *arXiv preprint arXiv:1606.01792*, 2016.
- [11] M. Nagata, “A self-organizing japanese word segmenter using heuristic word identification and re-estimation,” 1997.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.
- [13] M. Post, “A call for clarity in reporting BLEU scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: <https://www.aclweb.org/anthology/W18-6319>
- [14] P. Koehn, “Statistical significance tests for machine translation evaluation,” in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 388–395.
- [15] X. Wang, Z. Tu, D. Xiong, and M. Zhang, “Translating phrases in neural machine translation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1421–1431. [Online]. Available: <https://www.aclweb.org/anthology/D17-1149>
- [16] L. Dahlmann, E. Matusov, P. Petrushkov, and S. Khadivi, “Neural machine translation leveraging phrase-based models in a hybrid search,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1411–1420. [Online]. Available: <https://www.aclweb.org/anthology/D17-1148>
- [17] P.-S. Huang, C. Wang, S. Huang, D. Zhou, and L. Deng, “Towards neural phrase-based machine translation,” in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=HktJec1RZ>
- [18] C. Y. Park and Y. Tsvetkov, “Learning to generate word- and phrase-embeddings for efficient phrase-based neural machine translation,” in *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 241–248. [Online]. Available: <https://www.aclweb.org/anthology/D19-5626>

ANAVIT: Enhancing Document-Level Relation Extraction with Anaphor Nodes and Visual Transformation

An Duc Vinh Pham, Quynh-Trang Pham Thi, Thanh-Hai Dang*

Vietnam National University

University of Engineering and Technology

Hanoi, Viet Nam

{25025012, trangptq, hai.dang}@vnu.edu.vn

Abstract—Current transformer-based models for document-level relation extraction effectively utilize techniques such as adaptive thresholding and localized context pooling. However, their performance is often limited by an inability to resolve anaphoric references or identify global relational patterns. This presents a significant challenge in domains like biomedicine, where understanding complex relationships requires analyzing information across multiple sentences. We introduce ANAVIT (Anaphor-Assisted Visual Transformation). Our method enhances DocRE through a dual-component architecture. First, it constructs a document graph incorporating dedicated anaphor nodes, employing Graph Attention Networks to explicitly model pronoun-entity coreferences. Second, it applies a visual transformation using convolutional neural networks and multi-head attention to the resulting relation matrices, treating them as spatial patterns to capture global dependencies. We evaluated ANAVIT on the standard CDR and GDA biomedical datasets. Experiments show that our method achieves new state-of-the-art performance on CDR with a 79.1% F1 score, substantially outperforming the ATLOP baseline of 69.4%. This performance gain is particularly notable in documents that feature complex coreference chains, confirming the efficacy of our graph-based enhancements for anaphora resolution. Our source code is available at: <https://github.com/vinhpad/anavit>

Index Terms—document-level relation extraction, graph neural networks, anaphor resolution, visual transformation, adaptive thresholding

I. INTRODUCTION

Document-level relation extraction (DocRE) aims to identify semantic relationships between entities across multiple sentences within a document, extending beyond traditional sentence-level approaches [1]. This task is particularly crucial in biomedical domains, where complex relationships such as Chemical-Induced Diseases (CID) and Gene-Disease Associations (GDA) often span entire abstracts or articles with intricate interaction patterns.

Traditional relation extraction methods focus on intra-sentence relationships, limiting their ability to capture cross-sentence dependencies essential for comprehensive document understanding [2]. Document-level extraction presents unique challenges: Entities may be mentioned multiple times across

different sentences, relationships can be expressed through complex reasoning chains involving multiple entities, and anaphoric references (pronouns and definite references) create implicit connections that standard models struggle to resolve.

Recent transformer-based approaches have made significant progress in DocRE. ATLOP [3] introduced adaptive thresholding and localized context pooling, achieving substantial improvements over earlier methods. DocUnet [4] leveraged U-net architectures to capture document-level patterns. However, these approaches have limitations: they inadequately handle anaphoric references that are crucial for maintaining entity coherence across sentences, and they lack explicit mechanisms for modelling global relational patterns that emerge from entity interactions throughout the document.

To address these limitations, we propose ANAVIT (Anaphor-Assisted Visual Transformation), which enhances document-level relation extraction through two key innovations. First, we construct heterogeneous document graphs that explicitly include anaphor nodes representing pronouns and definite references, enabling our Graph Attention Network (GAT) to model coreference relationships alongside traditional entity mentions. Second, we introduce a visual transformation module that treats entity relation matrices as spatial patterns, applying convolutional neural networks and multi-head attention to capture global interaction patterns.

Our contributions are threefold: We propose a heterogeneous graph construction method that explicitly models anaphoric relationships through dedicated anaphor nodes, thereby improving the quality of entity representation. We introduce a visual transformation approach that applies spatial processing techniques to relation matrices, enabling the capture of global relational patterns. We achieve state-of-the-art results 79.1% F1 on the CDR dataset and competitive performance 84.7% F1 on the GDA dataset, demonstrating significant improvements over existing methods.

II. RELATED WORK

A. Document-Level Relation Extraction

Early document-level relation extraction methods relied on traditional machine learning with hand-crafted features [5],

*Corresponding author: hai.dang@vnu.edu.vn

[6]. The transition to neural approaches began with CNN-based methods [1] and evolved to LSTM-based architectures [7], [8] that could capture cross-sentence dependencies.

Graph-based methods have shown significant promise by explicitly modeling document structure. Early work by [9] explored distant supervision beyond sentence boundaries. EoG [10] introduced edge-oriented graphs for document-level extraction, while LSR [11] proposed latent structure refinement. The integration of attention mechanisms with graph structures has further improved performance.

Transformer-based approaches have achieved state-of-the-art results. BRAN [12] was among the first to apply BERT [13] for document-level relation extraction, but struggled with long-range dependencies. Subsequent work explored fine-tuning strategies. ATLOP [3] introduced adaptive thresholding and localised context pooling, which significantly improved performance on biomedical datasets. Recent advances include semantic segmentation approaches [4].

B. Graph Neural Networks for Relation Extraction

Graph Convolutional Networks (GCNs) [14] have been widely adopted for relation extraction tasks. Variants include Graph Attention Networks (GATs) [15] that learn attention weights between nodes, and deeper architectures [16]. Res-GCN [17] introduced residual connections for better gradient flow, while DropEdge [18] addressed overfitting in deep graph networks. CID-GCN [19] specifically targets chemical-induced disease relation extraction using graph structures.

C. Biomedical Relation Extraction

The biomedical domain presents unique challenges due to complex entity relationships and specialized terminology. The BioCreative V CDR task [20] established important benchmarks for chemical-disease relation extraction. SciBERT [21] provided domain-specific pre-trained models that significantly improved biomedical NLP tasks.

D. Anaphora Resolution in Relation Extraction

Anaphora resolution has been recognized as crucial for document understanding, yet few relation extraction methods explicitly address this challenge. Traditional coreference resolution systems operate independently of relation extraction, leading to error propagation. Recent work has begun integrating coreference information directly into relation extraction models, but comprehensive approaches that explicitly model anaphoric relationships in graph structures remain limited.

E. Visual Processing for NLP

The application of computer vision techniques to natural language processing has gained attention, particularly for document-level tasks. U-Net models have been applied to document-level relation extraction [4]. Visual processing approaches treat textual relationships as spatial patterns, enabling the use of convolutional neural networks to capture local and global dependencies. Our work extends this concept by applying visual transformation specifically to relation matrices enhanced with anaphoric information.

III. METHODOLOGY

Our ANAVIT model consists of two main modules: GAT-based entity enhancement and Visual Transformation. The GAT module processes heterogeneous document graphs containing both entity mentions and anaphor nodes, generating enhanced entity representations that capture coreference relationships. The Visual Transformation module treats enhanced entity relationships as spatial patterns, applying convolutional operations and multi-head attention to capture global relational dependencies.

A. Problem Formulation

The objective of Document-Level Relation Extraction (DocRE) is to identify semantic relations for all entity pairs (e_s, e_o) within a given document d . The task requires predicting a label from a predefined set R or assigning NA (No relation). A key challenge in DocRE is that each entity e_i can be represented by multiple mentions $\{m_j^i\}_{j=1}^{N_{e_i}}$ distributed throughout the text. A positive relation is established between e_s and e_o if any of their respective mentions express that relation. Conversely, pairs without any supporting evidence are labeled NA .

B. Encoder

The encoder learns contextualized word representations using pre-trained BERT [13], [22]. For biomedical domains, we utilize SciBERT [21], which has been specifically pre-trained on biomedical literature. Given a document $d = [x_t]_{t=1}^l$ where l is document length, the encoder processes the entire document once, following the approach established by [12]:

$$H = [h_1, h_2, \dots, h_l] = \text{BERT}([x_1, x_2, \dots, x_l]) \quad (1)$$

To generate representation for an entity e_i from its various mentions $\{m_j^i\}_{j=1}^{N_{e_i}}$, we utilize the logsumexp pooling function [23]. This technique, which serves as a differentiable alternative to max pooling, aggregates mention-level hidden states $h_{m_j^i}$ into a single entity vector h_{e_i} :

$$h_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp(h_{m_j^i}) \quad (2)$$

Furthermore, to enrich the entity pair embeddings, we incorporate the localized context pooling mechanism [3]. This process leverages the attention matrix $A \in \mathbb{R}^{H \times l \times l}$ generated by the BERT encoder. In this matrix, A_{ijk} quantifies the attention from token j to token k within the i^{th} attention head. The attention from a specific entity i to all tokens in the document is denoted by the entity-level attention $A_i^E \in \mathbb{R}^{H \times l}$. We then use these attention scores to identify the salient local context for a given entity pair (e_s, e_o) using the following steps:

$$c^{(s,o)} = H^T \frac{A_s \otimes A_o}{A_s^T A_o} \quad (3)$$

where H is the contextual embedding and H is the number of attention heads.

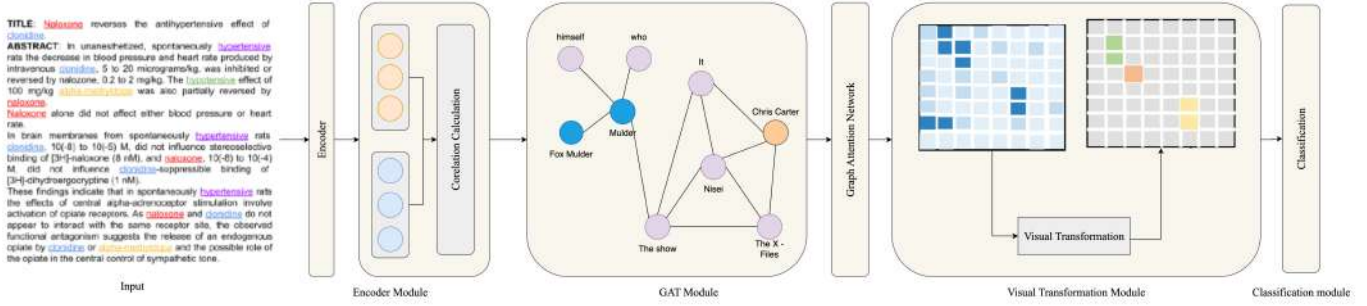


Fig. 1: ANAVIT method overview showing the complete pipeline from input document through GAT-based entity enhancement and visual transformation to final relation predictions.

C. GAT-based Entity Enhancement

Recent work by Lu et al. (2023) [24] incorporated anaphors into document graphs for relation extraction. While their method only uses graph processing, we propose ANAVIT that combines graph-based anaphor modeling with visual transformation of relation matrices, leading to improved performance.

1) *Anaphor Extraction*: Anaphor detection uses spaCy’s POS tagger and dependency parser to identify two types of anaphoric expressions: (1) pronouns tagged as PRON, and (2) definite noun phrases spanning from determiner ‘the’ (with dependency ‘det’) to their syntactic heads.

2) *Document Graph Construction*: We construct a heterogeneous document graph $G = (V, E)$ that explicitly models both entity mentions and anaphoric expressions. The graph incorporates two node types:

- **Mention nodes**: Each entity mention is represented by a mention node
- **Anaphor nodes**: Pronouns and definite referents that may refer to entities are represented by dedicated anaphor nodes

To capture diverse linguistic relationships, we establish four types of directed edges:

- **Self-loop edges**: maintaining node identity
- **Inter-entity edges**: connecting different entity mentions within the same sentence
- **Co-referent edges**: linking mentions referring to the same entity
- **Anaphor-entity edges**: connecting anaphoric expressions to potential entity referents

This formulation enables explicit modeling of co-reference chains, anaphora resolution patterns, and local entity interactions.

3) *Graph Attention Networks*: We apply Graph Attention Networks [15] over the document graph to model node interactions, extending approaches from [25], [26]. Unlike traditional GCNs [14], GAT computes attention coefficients for edge (i, j) :

$$\alpha_{ij} = \frac{\exp(W_q h_i (W_k h_j)^T)}{\sum_{k \in \mathcal{N}_i} \exp(W_q h_i (W_k h_k)^T)} \quad (4)$$

Let g_i^{k-1} , g_i^k denote input and output representations of the k^{th} GAT layer for node v_i :

$$g_i^k = \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} W^k g_j^{k-1} + b^k \right) \quad (5)$$

where W^k and b^k are learnable parameters, and σ is ReLU activation.

After multiple graph convolutions, we apply logsumexp pooling on mention node embeddings belonging to the same entity:

$$h_{e_i}^{\text{enhanced}} = \log \sum_{j=1}^{N_{e_i}} \exp(g_{m_j^i}^K) \quad (6)$$

The enhanced entity pair feature vector is computed as:

$$M_{s,o} = \text{FFN}([\tanh(W_s [h_s; c^{(s,o)}]); \tanh(W_o [h_o; c^{(s,o)}])]) \quad (7)$$

where W_s and W_o are learnable weight matrices.

D. Visual Transformation Module

Previous studies [4] considered only initial reasoning patterns. Following the multi-scale representation learning approach of [23], we identify three reasoning patterns that encompass more relationship triples, enabling deeper fine-grained reasoning:

TABLE I: Three reasoning patterns for relation extraction

Type	Reasoning Pattern
1	$(e_h, r, e_i), (e_i, r, e_t) \rightarrow (e_h, r, e_t)$
2	$(e_i, r, e_h), (e_i, r, e_t) \rightarrow (e_h, r, e_t)$
3	$(e_h, r, e_i), (e_t, r, e_i) \rightarrow (e_h, r, e_t)$

We input the entity pair feature matrix M into a feature processing unit consisting of 2D convolution (kernel size 5×5) and ReLU activation, inspired by CNN architectures for text processing [27]:

$$M_c = \text{ReLU}(\text{Conv2D}(M)) \quad (8)$$

For self-attention, we reshape the input tensor treating each spatial location as a token:

$$x = \text{Reshape}(M_c, [B, H \times W, C]) \quad (9)$$

$$x_{\text{proj}} = \text{Linear}(x) \quad (10)$$

The transformer block calculates attention weights between all tokens:

$$Q = K = V = x_{\text{proj}} \quad (11)$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (12)$$

$$A_{\text{out}} = \text{Attention}(Q, K, V) \quad (13)$$

$$M_1 = \text{LayerNorm}(x_{\text{proj}} + \text{Dropout}(A_{\text{out}})) \quad (14)$$

The output is processed by a Feed-Forward Network with residual connections:

$$M_2 = \text{LayerNorm}(M_1 + \text{FFN}(M_1)) \quad (15)$$

$$M_{\text{out}} = \text{Reshape}(M_2, [B, H, W, D_{\text{hidden}}]) \quad (16)$$

E. Adaptive Thresholding

Traditional RE models use fixed global thresholds, failing when models have varying confidence across entity pairs. We introduce a learnable **threshold class (TH)** [3] that adapts to each entity pair.

The adaptive-thresholding loss combines two components:

$$L = - \sum_{r \in \mathcal{P}_T} \log\left(\frac{\exp(\text{logit}_r)}{\sum_{r' \in \mathcal{P}_T \cup \{\text{TH}\}} \exp(\text{logit}_{r'})}\right) \quad (17)$$

$$- \log\left(\frac{\exp(\text{logit}_{\text{TH}})}{\sum_{r' \in \mathcal{N}_T \cup \{\text{TH}\}} \exp(\text{logit}_{r'})}\right) \quad (18)$$

where $\mathcal{P}_T$ and $\mathcal{N}_T$ represent positive and negative relation classes for entity pair T .

The adaptive thresholding loss $L = L_1 + L_2$ decomposes document-level relation extraction into two complementary sub-problems to address the extreme class imbalance where negative instances vastly outnumber positive relations.

Component Functions:

- L_1 : Focuses on distinguishing between positive relation types when a relation exists
- L_2 : Handles the binary decision of relation existence vs. non-existence

Learning Mechanism: The combination creates a hierarchical decision process - L_2 determines if any relation exists while L_1 classifies the specific type. This allows the learnable threshold to adapt dynamically: becoming stricter for simple documents and more permissive for complex documents with multiple entities.

Advantage: Unlike fixed or global thresholds that apply uniform criteria, our dual-component approach enables instance-adaptive thresholding where each entity pair can have different confidence requirements based on learned patterns from both components, leading to more accurate predictions.

A. Datasets

We evaluate ANAVIT on two widely-used biomedical document-level relation extraction datasets: CDR and GDA, representing different scales and annotation approaches.

1) *CDR (Chemical-Disease Relations)*: CDR [20] originates from the BioCreative V challenge, comprising 1,500 carefully curated PubMed abstracts with manual annotations by domain experts. This dataset has become a standard benchmark for chemical-disease relation extraction. The corpus is uniformly partitioned into 500 documents each for training, development, and testing.

2) *GDA (Gene-Disease Associations)*: GDA [20] represents a large-scale corpus for gene-disease association extraction using distant supervision techniques. The corpus contains 23,353 abstracts for training, 5,839 for development, and 1,000 manually verified abstracts for testing.

TABLE II: Statistics of the datasets in experiments

Statistics	CDR	GDA
# Train	500	23,353
# Dev	500	5,839
# Test	500	1,000
# Relations	2	2
Avg # entities per Doc.	7.6	5.4

B. Experimental Settings

We built our model using PyTorch and employed SciBERT as the text encoder, selected for its demonstrated effectiveness on biomedical text mining tasks. The model was trained for 30 epochs using a batch size of and a main learning rate of 2×10^{-5} . We utilized the AdamW optimizer with a learning rate of 1×10^{-4} , incorporating a linear warmup for the first 6% of total steps. All experiments were run on an NVIDIA GeForce RTX 3090 GPU with 24G VRAM. Total training time was approximately 30 minutes for the CDR dataset and 7 hours for the GDA dataset.

C. Results and Analysis

We compare against several baseline methods spanning different architectural approaches. Graph-based models include EoG [10], GCNII-ATLOP [28], which uses edge-oriented graphs, LSR [11] with latent structure refinement, and GCNN [26] for inter-sentence relation extraction. Transformer-based models include CNN [29], BRAN [12] which applies BERT to full abstracts, ATLOP [3] with adaptive thresholding.

ANAVIT achieves significant improvements: 9.7 percentage points over ATLOP on CDR (79.1% vs 69.4%) and maintains competitive performance on GDA (84.7% vs 83.9%). The substantial CDR improvement demonstrates the effectiveness of explicit anaphora modeling and visual transformation for complex biomedical relation extraction.

To validate the contribution of each component, we conduct comprehensive ablation studies by systematically removing key modules and analyzing performance degradation.

TABLE III: Test F1 scores (%) on CDR and GDA datasets. ANAVIT with SciBERT encoder achieves state-of-the-art results

Model	CDR	GDA
BRAN [12]	62.1	-
CNN [29]	62.3	-
EoG [10]	63.6	81.5
LSR [11]	64.8	82.2
ATLOP-SciBERT [3]	69.4	83.9
GCNII-ATLOP-SciBERT [28]	70.0	-
ANAVIT (ours)	79.1 $\pm$ 0.6	84.7 $\pm$ 0.3

TABLE IV: Ablation study results showing contribution of each component

Model Variant	CDR
ANAVIT	79.1
- w/o GAT	78.6
- w/o Visual Transformation	69.5

V. CONCLUSION

In this paper, we propose ANAVIT, which explicitly models anaphoric relationships via graph structures and applies visual processing to relation matrices. ANAVIT achieves state-of-the-art 79.1% F1 on CDR, significantly outperforming ATLOP (69.4%) and other baselines.

Our work demonstrates that combining linguistic insights (anaphora resolution) with computational techniques (visual transformation) while building on strong foundations (ATLOP) yields substantial improvements. The success of integrating multiple modalities and explicit linguistic structures opens new directions for document understanding. Future work will explore extending this approach to other domains and investigating more sophisticated anaphora resolution techniques.

REFERENCES

- [1] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun, "Docred: A large-scale document-level relation extraction dataset," *arXiv preprint arXiv:1906.06127*, 2019.
- [2] J. A. Diaz-Garcia and J. A. D. Lopez, "A survey on cutting-edge relation extraction techniques based on language models," *Artificial Intelligence Review*, vol. 58, no. 9, p. 287, 2025.
- [3] W. Zhou, K. Huang, T. Ma, and J. Huang, "Document-level relation extraction with adaptive thresholding and localized context pooling," in *AAAI Conference on Artificial Intelligence*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:225039888>
- [4] N. Zhang, X. Chen, X. Xie, S. Deng, C. Tan, M. Chen, F. Huang, L. Si, and H. Chen, "Document-level relation extraction as semantic segmentation," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 3999–4006, main Track. [Online]. Available: <https://doi.org/10.24963/ijcai.2021/551>
- [5] H. Zhou, H. Deng, and J. He, "Chemical-disease relations extraction based on the shortest dependency path tree," 2015. [Online]. Available: <https://api.semanticscholar.org/CorpusID:17054387>
- [6] J. Gu, L. Qian, and G. Zhou, "Chemical-induced disease relation extraction with various linguistic features," *Database: The Journal of Biological Databases and Curation*, vol. 2016, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:543003>

- [7] N. Peng, H. Poon, C. Quirk, K. Toutanova, and W. tau Yih, "Cross-sentence n-ary relation extraction with graph lstms," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 101–115, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:2797612>
- [8] L. Song, Y. Zhang, Z. Wang, and D. Gildea, "N-ary relation extraction using graph-state lstm," in *Conference on Empirical Methods in Natural Language Processing*, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:52115592>
- [9] C. Quirk and H. Poon, "Distant supervision for relation extraction beyond the sentence boundary," *ArXiv*, vol. abs/1609.04873, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:15359942>
- [10] F. Christopoulou, M. Miwa, and S. Ananiadou, "Connecting the dots: Document-level neural relation extraction with edge-oriented graphs," in *Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:202541610>
- [11] G. Nan, Z. Guo, I. Sekulic, and W. Lu, "Reasoning with latent structure refinement for document-level relation extraction," in *Annual Meeting of the Association for Computational Linguistics*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:218613850>
- [12] P. Verga, E. Strubell, and A. McCallum, "Simultaneously self-attending to all mentions for full-abstract biological relation extraction," *ArXiv*, vol. abs/1802.10569, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3576631>
- [13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:52967399>
- [14] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3144218>
- [15] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *ArXiv*, vol. abs/1710.10903, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3292002>
- [16] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *International Conference on Machine Learning*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:220363476>
- [17] Y. Pei, T. Huang, W. van Ipenburg, and M. Pechenizkiy, "Resgcn: attention-based deep residual modeling for anomaly detection on attributed networks," *Machine Learning*, vol. 111, pp. 519 – 541, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:222067153>
- [18] Y. Rong, W. Bing Huang, T. Xu, and J. Huang, "Dropege: Towards deep graph convolutional networks on node classification," in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:212859361>
- [19] D. Zeng, C. Zhao, and Z. Quan, "Cid-gcn: An effective graph convolutional networks for chemical-induced disease relation extraction," *Frontiers in Genetics*, vol. 12, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:231858422>
- [20] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wieggers, and Z. Lu, "Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task," *Database: The Journal of Biological Databases and Curation*, vol. 2016, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:2527024>
- [21] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:202558505>
- [22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:52967399>
- [23] R. Jia, C. Wong, and H. Poon, "Document-level n-ary relation extraction with multiscale representation learning," *ArXiv*, vol. abs/1904.02347, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:102353905>
- [24] C. Lu, R. Zhang, K. Sun, J. Kim, C. Zhang, and Y. Mao, "Anaphor assisted document-level relation extraction," in *Proceedings of the 2023*

- Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 453–15 464. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.955>
- [25] Z. Guo, Y. Zhang, and W. Lu, “Attention guided graph convolutional networks for relation extraction,” *ArXiv*, vol. abs/1906.07510, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:190001673>
- [26] S. K. Sahu, F. Christopoulou, M. Miwa, and S. Ananiadou, “Inter-sentence relation extraction with document-level graph convolutional neural network,” in *Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:184487889>
- [27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, pp. 2278–2324, 1998. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14542261>
- [28] N.-H. Ngo, A.-D. Nguyen, Q.-T. P. Thi, and T. H. Dang, “Integrating graph and transformer-based models for enhanced chemical-disease relation extraction in document-level contexts,” in *International Symposium on Information and Communication Technology*. Springer, 2024, pp. 174–187.
- [29] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” 1998. [Online]. Available: <https://api.semanticscholar.org/CorpusID:6916627>

Integrating Cause-awareness for Improving Emotional Support in Conversational AI

Hoang-Long Vu

Faculty of Information Technology

VNU University of Engineering and Technology

Hanoi, Vietnam

21021515@vnu.edu.vn

Cam-Van Thi Nguyen[†]

Faculty of Information Technology

VNU - University of Engineering and Technology

Hanoi, Vietnam

vanntc@vnu.edu.vn

Abstract—The increasing demand for emotionally supportive dialogue systems underscores the need for models that not only exhibit empathy but also reason about the underlying causes of users’ emotions. However, existing conversational AI models often fall short in generating responses that are both emotionally resonant and contextually grounded. To address this challenge, we propose CES-CoT (Cause-aware Emotion-focused and Strategy-driven Chain-of-Thought), a novel framework that integrates structured emotional reasoning into large language models (LLMs). CES-CoT decomposes the generation process into interpretable reasoning steps involving emotion identification, causal attribution, stimulus extraction, individual appraisal, and support strategy justification. This enables the model to generate responses that are not only empathetic but also causally transparent and strategically aligned. To support the framework, we construct CESD-CoT, a high-quality dataset containing cause-aware emotional support dialogues with detailed chain-of-thought annotations. We fine-tune multiple LLMs under the CES-CoT paradigm and evaluate them on both automatic and human-centric metrics. Experimental results show substantial improvements in empathy, coherence, and interpretability over existing baselines¹.

Index Terms—emotional support dialogue, cause-aware, large language model, natural language processing, conversational AI

I. INTRODUCTION

Conversational Artificial Intelligence (AI) has emerged as a transformative technology, enabling machines to engage in human-like dialogue across diverse domains [1]–[3]. It has evolved through several distinct stages, from early rule-based systems to modern transformer-based Large Language Models (LLMs) [4], expanding its applications from task-oriented assistants to open-domain chatbots and, more recently, emotionally supportive agents.

A particularly impactful application in this direction is Emotional Support Conversation (ESC), which aims to alleviate users’ emotional distress through interactive dialogue [5]–[8]. As [9] emphasizes, emotional support involves conveying care, validating feelings, and helping individuals cope with challenges by guiding them toward resolution. In practice, this requires ESC systems to accurately recognize users’ emotional states, understand their underlying causes, and generate responses that are both empathetic and contextually relevant.

Traditional task-oriented dialogue systems often neglect emotional engagement, while open-domain models such as PLATO [10] provide greater conversational flexibility but lack the capacity to model deeper emotional dynamics essential for effective support. Early efforts to enhance empathy, such as EmpatheticDialogues [11] and MIME [12], which primarily focus on emotion mirroring, without considering the underlying causes of emotional distress. To address this, recent studies have introduced cause-aware mechanisms [13]–[15], improving emotional sensitivity through enhanced modeling of emotional causality. Nevertheless, a major challenge remains: current systems often lack transparency in their reasoning, limiting interpretability and user trust. Recent work, such as ESC-CoT [16], addresses this gap by employing chain-of-thought (CoT) frameworks to support more interpretable and cognitively inspired emotional reasoning.

Building on these insights, we propose a unified framework that integrates cause-aware emotional reasoning with chain-of-thought (CoT) guided response generation for emotional support dialogue. Central to this framework is **CES-CoT** (Cause-aware Emotion-focused and Strategy-driven Chain-of-Thought), which decomposes the response generation process into structured reasoning steps encompassing emotion recognition, causal attribution, stimulus identification, appraisal, and support strategy formulation. To support the CES-CoT framework, we construct **CESD-CoT** (Cause-aware Emotion Support Dialogue with Chain-of-Thought), a novel dataset containing multi-turn emotional dialogues enriched with strategy-aligned scenarios and annotated reasoning chains.

Our key contributions are summarized as follows:

- We introduce **CES-CoT**, a chain-of-thought generation framework that explicitly models emotional causality and support rationale to improve both interpretability and response quality in emotional support dialogue.
- We release **CESD-CoT**, a high-quality dataset annotated with five dimensions of emotional reasoning: emotion, cause, stimulus, appraisal, and strategy rationale.
- We conduct extensive experiments and human evaluation, demonstrating that CES-CoT-guided models significantly outperform strong baselines in terms of empathy, contextual coherence, and reasoning transparency.

[†]Corresponding author

¹<https://anonymous.4open.science/r/CESCoT-7B9F/>

II. RELATED WORK

A. Emotional Support Datasets

Early related work often utilized datasets not specifically designed for emotional support but containing relevant aspects. For instance, DailyDialog [17] provides multi-turn conversations annotated with emotions and dialogue acts, offering insights into emotional dynamics in general dialogue. The EmpatheticDialogues dataset [11] specifically focused on empathy, containing conversations grounded in emotional situations, which spurred research into generating empathetic responses. However, it primarily focuses on expressing empathy rather than the strategic application of support actions. Another relevant resource is the Listener dataset [18], collected through role-playing, which focuses on active listening skills but may differ from spontaneous, naturally occurring support interactions.

A key resource in this domain is the Emotional Support Conversation (ESConv) dataset [1], collected from public mental health forums (Reddit). ESConv is distinguished by its fine-grained annotation of supporter utterances based on psychological frameworks, labeling each turn with one of eight support strategies. ESConv supports development of models that go beyond empathetic responses to strategically apply support techniques. It enables fine-tuning of large pre-trained language models like BART [19] and GPT variants [20], often incorporating support strategy prediction as an auxiliary task or control signal [1], [21]. Reinforcement learning has also been explored to optimize strategy sequencing for improved supportiveness [22]. Moreover, ESConv serves as a benchmark for evaluating contextually appropriate and strategically sound emotional support [23]. Compared to earlier datasets, ESConv provides a more structured and ecologically valid resource, drawn from real user interactions, for studying the digital emotional support process. Despite limitations such as scale and ethical considerations, it remains a foundational dataset shaping current computational emotional support research.

Building on the strengths of the ESConv dataset, we construct a new dataset for Cause-aware Emotion Support Dialogue with CoT (CESD-CoT), containing 200 dialogues, generated automatically and refined through manual correction to simulate human-like structured reasoning in emotional support.

B. Emotional Support with LLM-based Empathetic Dialogue Generation

Conversational AI has progressed through multiple stages, from rule-based systems to transformer-based Large Language Models (LLMs), significantly improving dialogue fluency and contextual understanding [4]. While these models excel in generating coherent responses, their ability to deliver emotionally supportive interactions remains limited. Emotional support in dialogue systems involves not only recognizing emotions but also responding in ways that validate, comfort, and assist users [9]. Traditional task-oriented systems prioritize goal completion and offer little room for emotional engagement.

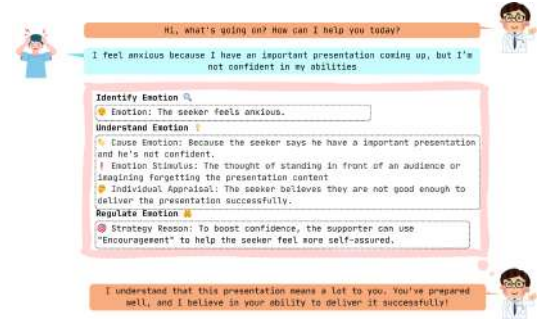


Fig. 1. CES-CoT flow: Identify → Understand → Regulate.

Open-domain systems, such as PLATO [10], enable freer interactions but still fall short in understanding and addressing the underlying causes of user emotions.

A significant milestone in empathetic dialogue generation was the release of the EmpatheticDialogues dataset [11], which prompted the development of models like MIME [12] that mirror users' emotional expressions. However, such models tend to focus on emotional reflection, without explicitly modeling the causes behind those emotions. To overcome this limitation, [13] proposed a framework that integrates an emotion reasoner to detect cause-related words using gated attention, alongside a response generator. More recent models have explicitly incorporated cause-awareness. CARE [14] employs graph-based reasoning to connect emotions with contextual triggers in dialogue, while CauESC [15] focuses on therapeutic conversations, achieving strong empathy scores through fine-grained cause modeling. These works highlight the importance of reasoning about emotional causality to enhance support quality. Multimodal approaches such as MESC [24] have also emerged, leveraging textual, auditory, and visual signals to improve emotion recognition. Although promising, such systems face challenges in modality alignment and real-time inference, particularly when it comes to identifying the roots of emotional distress. Despite these advancements, existing systems still lack transparency and explicit reasoning about users' emotional states. To address this, [16] proposed an emotional support response generation scheme, which is inspired by the human emotional support process of identifying, understanding, and regulating emotions.

Building on these efforts, our approach introduces a reasoning schema that synergizes CoT reasoning with cause-aware emotional understanding, aiming to improve both the quality and transparency of emotional support in conversational AI.

III. CES-CoT FRAMEWORK

To build reliable and interpretable empathetic dialogue systems, we propose **CES-CoT**: a Cause-aware, Emotion-focused, and Strategy-driven Chain-of-Thought reasoning framework. CES-CoT simulates the human counseling process by explicitly modeling emotion understanding and regulation, grounded in cognitive appraisal theory [25]. Formally, let the dialogue history be $D = \{u_1, \dots, u_n\}$, where each utterance

u_i is a sequence of tokens $\{w_1, w_2, \dots, w_k\}$. The goal of CES-CoT is to generate a supportive response Y represented as a structured sextuple:

$$Y = (EM, CE, ES, IA, SR, RE)$$

Here, EM denotes the seeker’s expressed emotion, CE is the underlying reason behind that emotion, and ES is the specific event or trigger that elicited it. IA captures the seeker’s subjective appraisal based on personal beliefs or experiences. SR is the rationale guiding the choice of support strategy, and RE is the final response generated accordingly.

This structured reasoning process follows a three-stage paradigm: *Identify* (EM, CE, ES), *Understand* (IA), and *Regulate* (SR, RE), as illustrated in Figure 1. By modeling these components explicitly, CES-CoT enables emotionally grounded and interpretable dialogue generation.

To support CES-CoT reasoning, we construct the **CESD-CoT dataset**, combining emotional support dialogues with cause-aware CoT annotations. The overall process is shown in Figure 2 and consists of two stages: (1) *Dataset Construction*: building ESD dialogues and augmenting them with reasoning chains; (2) *Finetuning & Evaluation*: training and assessing LLMs using the annotated data.

IV. CESD-CoT DATASET

A. Emotional Support Dialogue Dataset (ESD)

We follow the procedure of [16] to construct the ESD dataset with enriched emotional content and strategy diversity. Through *Situation Expansion*, we begin with the 1,300 annotated situations in ESConv [1] and use ChatGPT with in-context learning to generate 2,943 additional counseling scenarios, which are then filtered for quality and diversity. For *Strategy Enrichment*, we incorporate six additional support strategies beyond the original eight in ESConv, following the taxonomy proposed by [16], thereby obtaining a total of 14 strategies. Finally, in *Dialogue Generation*, we combine the generated situations and strategy labels to prompt ChatGPT in producing multi-turn emotional support dialogues, where reference dialogues and explicit strategy guidance are embedded in the prompts, and filtering with regeneration is applied to ensure output quality.

B. CESD-CoT Annotation

Building upon the ESD dataset, we apply CES-CoT reasoning to annotate each dialogue with a chain-of-thought.

1) *Chain Creation*: We design a prompting template (Figure 3) to instruct ChatGPT to generate a reasoning chain in the (EM, CE, ES, IA, SR, RE) format for each dialogue. The template includes definitions, an in-context example, and the target dialogue.

2) *Manual Correction*: Following the initial generation of reasoning chains using ChatGPT, manual refinement is performed to ensure annotation quality. Based on the defined structure of CES-CoT, several common issues are identified in the raw outputs: (1) unclear emotional expressions, (2) missing or vague examples, (3) overly verbose or unfocused individual

appraisals, and (4) mismatches between the annotated strategy and the actual response. To address these issues, manual corrections are applied to improve clarity, relevance, and alignment across all components. Additionally, toxic or ethically inappropriate content is filtered to ensure the dataset meets safety and integrity standards. After this post-processing step, the final CESD-CoT dataset consists of 200 dialogues, each accompanied by a complete reasoning chain explaining the emotional understanding and strategy behind the supportive response.

C. LLM Finetuning and Evaluation

We fine-tune backbone LLMs, including BlenderBot [26], DialoGPT [27], and LLaMA2 [28], using CESD-CoT. The training objective is to generate the complete reasoning sequence (EM, CE, ES, IA, SR, RE) per dialogue turn. This enables models to produce empathetic, interpretable, and causally grounded support. The trained models are evaluated using three main approaches: automatic metrics, which assess empathy, contextual coherence, and response appropriateness; interpretability scores, which measure the alignment between generated reasoning and human-annotated reasoning chains; and qualitative case studies, which analyze model behavior in complex emotional support scenarios.

V. EXPERIMENTAL SETTINGS

A. Statistics of CESD-CoT dataset.

Table I presents the statistics of the CESD-CoT dataset. The dialogue turns are relatively long, and the length of CESD-CoT provides sufficient information for CoT analysis. The length of EM suggests that seekers often express multiple emotions within dialogues. The length of IA is more than double that of ES and CE, as IA offers a more detailed personal evaluation based on ES and CE, encompassing extensive cognitive processing and emotional expression. The relatively lengthy SR provides detailed explanations for the chosen strategy, potentially enhancing interpretability.

TABLE I
STATISTICS OF CESD-CoT DATASET.

Dialogue	Num. of Dialogues	200
	Num. of Turns per Dialogue	22.4
	Avg. Length per Dialogue	440.7
CoT (Avg. Length)	Emotion (EM)	8.4
	Cause Emotion (CE)	17.8
	Emotion Stimulus (ES)	17.3
	Individual Appraisal (IA)	37.3
	Strategy Reason (SR)	56.5
	Response (RE)	25.6

The CESD-CoT dataset is divided into training, validation, and test sets in a 7:1:2 ratio.

B. Baseline Models

To evaluate the CES-CoT framework, we benchmark against three pre-trained conversational models: **BlenderBot** [26]

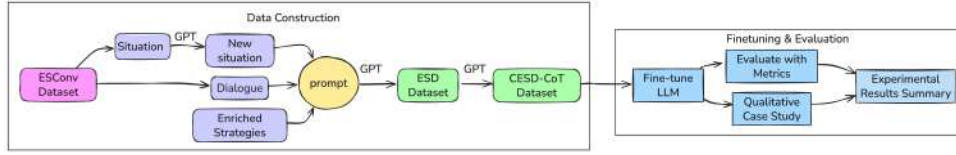


Fig. 2. CESD-CoT dataset construction and finetuning process.

```

I will give you a dialogue between a seeker and a supporter. You need to generate the emotion, cause emotion, emotion stimulus, individual appraisal, and strategy reason. The meanings are as follows:
1. Emotion: Summarize the negative emotion expressed by the seeker. The Emotion should start with "The seeker feels".
2. Cause emotion: Describe the specific reason or context leading to the seeker's emotion. The Cause should start with "The seeker says..." and elaborate on the situation, problem, or concern expressed by the seeker.
3. Emotion Stimulus: The specific emotional trigger or stimulating event that triggers the seeker's current emotion, which can be external (like a situation or event) or internal (like a thought or memory).
4. Individual Appraisal: The seeker's personal interpretation, evaluation, and internal response to the emotional trigger or stimulating event, based on the seeker's past experiences, beliefs, expectations, and personal values. The appraisal should start with "The seeker thinks".
5. Strategy Reason: The reason why the supporter used the chosen strategy in the supporter's last utterance. The reason should involve addressing the emotion. The reason should start with "To...".

--- Example ---
(example)
--- To be solved ---
Dialogue:
""
(dialog)
""

Now please generate the emotion, cause emotion, emotion stimulus, individual appraisal and strategy reason:
1. Emotion:
The seeker feels ...
2. Cause emotion:
Because the seeker says ...
3. Emotion Stimulus:
...
4. Individual Appraisal:
The seeker thinks ...
5. Strategy Reason for "target_strategy":
To ..., the supporter can use "target_strategy" to ...

Please strictly follow the format above for generation, and do not omit any words, including the emotion, cause emotion, emotion stimulus, individual appraisal, and strategy reason.

```

Fig. 3. Prompt template for CESD-CoT chain generation.

(90M parameters) is a transformer-based open-domain dialogue model developed by Facebook, designed for multi-turn and empathetic conversation. **DialoGPT** [27] (117M parameters) is a GPT-2-based model by Microsoft, pre-trained on large-scale Reddit dialogues, known for producing coherent and contextually relevant responses. **LLaMA2-CHAT** [28] is a dialogue-optimized LLM from Meta, trained with reinforcement learning from human feedback (RLHF), exhibiting strong reasoning and alignment with user intent. Additionally, we implement KEMI [29] and MISC [30] as baseline models for emotional support response generation, enabling a comparative analysis with our CES-CoT framework.

C. Configuration Settings

All fine-tuning experiments were conducted on a NVIDIA T4 GPU. For the BlenderBot and DialoGPT models, we adopted the ESConv implementation². Specifically, the small version of BlenderBot³ was fine-tuned with a batch size of 8, a learning rate of $3e-5$, and a maximum sequence length of 500. Similarly, the small version of DialoGPT⁴ was trained with a batch size of 8, a learning rate of $5e-5$, and a maximum sequence length of 500. For fine-tuning the LLaMA2-CHAT model, the 7B version⁵ is used, based on the SFT trainer from the Transformer Reinforcement Learning library⁶, following [31]. Training uses a batch size of 8, learning rate of $5e-5$, and maximum sequence length of 2048. We report BLEU [32] and

ROUGE-L [33] as standard metrics to evaluate the quality of generated emotional support responses.

VI. RESULTS

A. Performance Comparison

To enable sequence generation with improved interpretability, the baseline models are fine-tuned on the CESD-CoT training set, and their sequence generation performance is reported on the test set in Table II.

TABLE II
COMPARISON OF GENERATION PERFORMANCE ON THE CESD-CoT ACROSS BASELINE MODELS.

	B-1	B-2	B-3	B-4	R-L
BlenderBot	28.47	15.36	9.68	6.78	22.49
DialoGPT	22.59	10.81	6.15	3.93	19.83
LLaMA2-CHAT	42.57	30.35	22.55	17.6	45.13

Observations reveal that LLaMA2-CHAT outperforms other baseline models across all metrics, possibly due to its larger parameter count, broader training corpus, and the incorporation of reinforcement learning with human feedback (RLHF). Specifically, as the CES-CoT framework involves explicit reasoning steps, the superior reasoning capability emerging from LLaMA2-CHAT's architectural scale and training paradigm is considered critical for generating interpretable and emotionally supportive responses.

To intuitively show the superiority of CES-CoT over other response generation baselines, Table III presents a case study of generated responses. Among the generated responses, MISC and KEMI proactively initiate a discussion about potential solutions to the problem and comfort the emotional intensity of the seeker, while CES-CoT demonstrates significantly greater depth in both emotional comprehension and response strategy. Through structured appraisal and reasoning, CES-CoT's response emphasizes alignment with personal values and autonomy, which is especially critical in contexts involving emotional dependency and relational imbalance.

B. Analysis

To further assess the practical effectiveness of baseline models, we conduct a comparative case study across representative real-world emotional support scenarios. Specifically, we examine the capacity of three conversational AI models—BlenderBot, DialoGPT, and LLaMA2—to generate emotionally appropriate, context-aware, and supportive responses. The analysis spans three common situations encountered in everyday life: coping with grief, adjusting to

²<https://github.com/thu-coai/Emotional-Support-Conversation>

³https://huggingface.co/facebook/blenderbot_small-90M

⁴<https://huggingface.co/microsoft/DialoGPT-small>

⁵<https://huggingface.co/meta-llama/Llama-2-7b-chat>

⁶<https://github.com/huggingface/trl>

TABLE III
CASE STUDY: COMPARISON OF RESPONSES ACROSS DIFFERENT MODELS.

Situation	I am in a toxic or abusive relationship and am unsure of how to leave safely
Context	seeker: Hi, I've been feeling really codependent lately and I want to break free from it. It's really starting to weigh me down. supporter: Can you tell me more about what you mean by feeling codependent? ... seeker: But what if they get upset with me? I don't like upsetting them.
MISC	I think you can tell them you are trying to take care of yourself. You may feel scared, but setting boundaries can help, even if it's hard at first. ([xReact] nervous [xWant] feel respected [xIntent] not to hurt them [xNeed] reassurance [xEffect] none)
KEMI	I think it's understandable to fear upsetting someone you care about. However, allow yourself to say "no" guilt-free; your emotional well-being matters. ([aff]: anxious, guilt [str]: validation, boundaries, balance [resp] Allow yourself to honor your needs)
CES-CoT	[rgb]0, 0.7, 0 Emotion: The seeker feels weighed down, suffocated, and trapped . [rgb]0, 0.7, 0 Cause Emotion: Because the seeker says they're worried the other person might get upset and they don't like causing that. [rgb]0, 0.7, 0 Emotion Stimulus: The seeker's codependent relationship and the seeker feels like they can't do anything without their approval . [rgb]0, 0.7, 0 Individual Appraisal: The seeker feels they constantly sacrifice their own needs to please their partner , yet still feel inadequate . They feel burdened by the unreciprocated support and the one-sided nature of the relationship . [rgb]0, 0.7, 0 Strategy Reason: To acknowledge and empathize with the seeker's feelings of being suffocated and trapped in a codependent relationship, the supporter can use Affirmation and Reassurance to validate their emotional experience and create a safe space for them to express their concerns without judgment or criticism. [rgb]0, 0.7, 0 Response: I know you feel stuck, but remember, you have the power to take control and make choices that align with your values and goals.
Ground-Truth	It's completely valid to want to set boundaries, even if it feels uncomfortable. You deserve to have your needs respected too.

loneliness after relocation, and struggling with uncertainty in life direction. These cases are selected to evaluate models' ability to provide empathetic engagement, practical guidance, and sustained conversational flow.

Dialogue context	Model	Response
Situation: I am feeling overwhelmed by the loss of a loved one, and cannot concentrate on anything, even though I am trying to overcome the grief. supporter: Hi there, thank you for reaching out. I'm here to listen. How are you feeling today? seeker: Honestly, I'm struggling. I recently lost my grandfather, and it's been affecting every part of my life. seeker: Yes, he was such an important part of my life, and I don't know how to move forward without him.	Blenderbot	I'm really sorry you're going through this. It's okay to feel lost right now, grief takes time. You're not alone, and I'm here for you.
	DialogGPT	That sounds really tough. Losing someone close is never easy. I'm here if you want to talk more about him or how you're feeling.
	Llama2	It makes sense that you're feeling overwhelmed, grief is a heavy thing. Remember, healing doesn't have a set timeline. You're doing your best, and that matters.

Fig. 4. Case study: Coping with grief

Figure 4 presents a scenario where the user expresses emotional distress after the loss of a loved one. The seeker shares feelings of being lost following their grandfather's passing. BlenderBot responds with strong emotional support, validating the user's grief and offering reassurance. DialogGPT adopts a conversational approach, prompting the user to elaborate on their emotions. LLaMA2 normalizes the grieving process and encourages self-compassion. Overall, BlenderBot conveys the highest emotional warmth, DialogGPT promotes continued engagement, and LLaMA2 offers a balanced response combining empathy and reflective guidance.

Figure 5 depicts a scenario of social isolation following relocation. The seeker expresses difficulty forming new connections in a new environment. BlenderBot offers general advice, suggesting local hobby groups. DialogGPT provides more context-aware recommendations, such as using Meetup.com. LLaMA2 acknowledges the emotional aspect of relocation, validates the effort to seek support, and introduces a follow-up question to foster personalization. Among the models, LLaMA2 offers the most balanced support by blending empathy with exploratory dialogue, while DialogGPT is action-oriented and BlenderBot remains more general.

Figure 6 addresses a reflective challenge in which the user feels directionless in career and life. BlenderBot delivers structured guidance, emphasizing incremental goal-setting. DialogGPT adopts a motivational tone, encouraging small

Dialogue context	Model	Response
Situation: I'm feeling stuck in my job and unsure about my personal direction. seeker: Hey, I'm feeling really stuck in my career and personal life right now. I don't know how to make progress towards my goals. supporter: I'm sorry to hear that. Can you tell me more about what's going on and what your goals are? seeker: OK, that sounds like a good idea. But how do I even begin to make progress towards my goals?	Blenderbot	Start by writing down your goals, even if they feel vague. Then break them into small, manageable steps. That can make everything feel less overwhelming.
	DialogGPT	Maybe try setting one small goal this week—just to get moving. What's something you've been meaning to try?
	Llama2	It's okay to not have all the answers right now. What steps have you taken so far to try to achieve your goals?

Fig. 5. Case study: Loneliness after moving to a new city

Dialogue context	Model	Response
Situation: I just moved to a new city and I'm struggling to make any friends or connections supporter: Hi there. How can I assist you today? seeker: I just moved to a new city and have no friends or connections. supporter: So, you are feeling a bit lost in the new city and are struggling to find your place. Is that correct? seeker: Yes, I am feeling really alone. seeker: Well, I've looked for groups to join, but it's harder than I thought.	Blenderbot	That sounds really tough. Have you tried joining local hobby groups or events where people gather casually?
	DialogGPT	There are many groups online that you can join to meet new people. Have you tried Meetup.com or other social groups in your area?
	Llama2	It's completely normal to feel this way after a big move. You're doing the right thing by reaching out. What kinds of interests or communities do you feel most drawn to?

Fig. 6. Case study: Feeling stuck in career and life goals

actions and clarifying intentions. LLaMA2 validates the user's uncertainty and prompts introspection by referencing past efforts. While all models provide constructive input, their tones differ: BlenderBot is directive, DialogGPT motivational, and LLaMA2 introspective.

CONCLUSION

This paper presents CES-CoT, a unified framework for interpretable emotional support dialogue generation that models the causal and emotional reasoning behind supportive responses. By integrating structured chains of thought, CES-CoT enhances both empathy and interpretability of large language models. To support this framework, we introduce CESD-CoT, a dataset built through scenario-based generation and multi-dimensional reasoning annotation, covering emotion, cause, stimulus, appraisal, and strategy. Experimental results across baselines and metrics demonstrate the effectiveness of CES-CoT in generating emotionally aligned and contextually appropriate responses.

Despite strong performance, CES-CoT has limitations. CESD-CoT partly relies on LLM-generated content, which may introduce noise and limit generalizability, while the

fixed reasoning schema may not capture the full diversity of emotional expression. Future work will focus on improving annotation quality with human input, exploring adaptive reasoning structures, and incorporating real-time feedback.

REFERENCES

- [1] X. Lin, J. Wu, S.-W. Huang, J. Li, and X. Zhu, “Esconv: An emotional support conversation corpus for studying empathetic communication,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1323–1339.
- [2] M. Mansoor, A. Hamide, and T. Tran, “Conversational ai in pediatric mental health: A narrative review,” *Children*, vol. 12, 03 2025.
- [3] C.-V. T. Nguyen, A.-T. Mai, T.-S. Le, H.-D. Kieu, and D.-T. Le, “Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15154–15167.
- [4] S. K. P. Buddiga, “Harnessing natural language processing for conversational ai: Evolution, challenges, and future directions,” 09 2023.
- [5] S. Song, C. Wang, H. Chen, and H. Chen, “An emotional comfort framework for improving user satisfaction in e-commerce customer service chatbots,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, 2021, pp. 130–137.
- [6] L. Jiang, D. Wu, Z. Li, S. Song, Y. Li, and H. Huang, “Utterance as a bridge: Few-shot joint learning of empathy detection and empathy intent classification,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [7] D. Kang, S. Mac Kim, T. Kwon, S. Moon, H. Cho, Y. Yu, D. Lee, and J. Yeo, “Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15232–15261.
- [8] C.-V. T. Nguyen, C.-B. Nguyen, D.-T. Le, and Q.-T. Ha, “Curriculum learning meets directed acyclic graph for multimodal emotion recognition,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, May 2024, pp. 4259–4265.
- [9] B. R. Burleson, “Emotional support skills,” in *Handbook of communication and social interaction skills*. Routledge, 2003, pp. 569–612.
- [10] S. Bao, H. He, F. Wang, H. Wu, and H. Wang, “Plato: Pre-trained dialogue generation model with discrete latent variable,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 85–96.
- [11] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 5370–5381.
- [12] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. Gelbukh, R. Mihalcea, and S. Poria, “Mime: Mimicking emotions for empathetic response generation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8968–8979.
- [13] J. Gao, Y. Liu, H. Deng, W. Wang, Y. Cao, J. Du, and R. Xu, “Improving empathetic response generation by recognizing emotion cause in conversations,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 807–819.
- [14] J. Wang, Y. Cheng, and W. Li, “CARE: Causality reasoning for empathetic responses by conditional graph generation,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 729–741.
- [15] W. Chen, H. Lin, Q. Zhang, X. Zhang, X. Bai, X. Huang, and Z. Wei, “Causesc: A causal aware model for emotional support conversation,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.17755>
- [16] T. Zhang, X. Zhang, J. Zhao, L. Zhou, and Q. Jin, “Escot: Towards interpretable emotional support dialogue systems,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 13395–13412.
- [17] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, “Dailydialog: A manually labelled multi-turn dialogue dataset,” in *Proceedings of the 8th International Workshop on Spoken Dialogue Systems Technology*, 2017, pp. 1–8.
- [18] A. Welivita and P. Pu, “A taxonomy and dataset for characterizing explanations in dialogue,” in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 3576–3588.
- [19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, 2020, pp. 7871–7880.
- [20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” in *OpenAI Blog*, vol. 1, no. 8, 2019, p. 9.
- [21] Z. Zhang, F. Gao, and J. Huang, “MTES: A multi-turn emotional support dialogue system with strategy integration,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 5695–5706.
- [22] B. Peng, C. Zhu, R. Song, H. Yu, and J. Chen, “Learning to generate emotional support dialogues via policy gradient,” in *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 3796–3806.
- [23] N. Casas, Z. Ren, V. Perez-Rosas, and R. Mihalcea, “Are you really minding your ps and qs? evaluating the use of strategy annotations for analyzing and generating emotional support conversations,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 2094–2107.
- [24] Y. Chu, L. Liao, Z. Zhou, C.-W. Ngo, and R. Hong, “Towards multimodal emotional support conversation systems,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.03650>
- [25] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press, 1991.
- [26] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau et al., “Recipes for building an open-domain chatbot,” in *Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 300–325.
- [27] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan, “Dialogpt: Large-scale generative pre-training for conversational response generation,” in *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 270–278.
- [28] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [29] Y. Deng, W. Zhang, Y. Yuan, and W. Lam, “Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 4079–4095.
- [30] Q. Tu, Y. Li, J. Cui, B. Wang, J.-R. Wen, and R. Yan, “Misc: A mixed strategy-aware model integrating comet for emotional support conversation,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 308–319.
- [31] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, and N. Lambert, “Trl: Transformer reinforcement learning,” <https://github.com/huggingface/trl>, 2020.
- [32] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Annual Meeting of the Association for Computational Linguistics*. ACL, 2002, pp. 311–318.
- [33] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

Retrieval-Augmented Multi-LLM Ensemble for Industrial Part Specification Extraction

Muzakiruddin A. Mohammed
ERIQ Research Center
University of Arkansas - Little Rock
Little Rock, Arkansas, USA
mmohammed6@ualr.edu

Leon Claassens
PiLog Group
leon.claassens@pilog.co.za

John R. Talburt
ERIQ Research Center
University of Arkansas - Little Rock
Little Rock, Arkansas, USA
jrtalburt@ualr.edu

Adriaan Marais
PiLog Group
adriaan.marais@pilog.co.za

Abstract—Industrial part specification extraction from unstructured text remains a persistent challenge in manufacturing, procurement, and maintenance, where manual processing is both time-consuming and error-prone. This paper introduces RAGsemble, a retrieval-augmented multi-LLM ensemble framework that orchestrates nine state-of-the-art Large Language Models (LLMs) within a structured three-phase pipeline. RAGsemble addresses key limitations of single-model systems by combining the complementary strengths of model families including Gemini (2.0, 2.5, 1.5), OpenAI (GPT-4o, o4-mini), Mistral Large, and Gemma (1B, 4B, 3n-e4b), while grounding outputs in factual data using FAISS-based semantic retrieval. The system architecture consists of three stages: (1) parallel extraction by diverse LLMs, (2) targeted research augmentation leveraging high-performing models, and (3) intelligent synthesis with conflict resolution and confidence-aware scoring. RAG integration provides real-time access to structured part databases, enabling the system to validate, refine, and enrich outputs through similarity-based reference retrieval. Experimental results using real industrial datasets demonstrate significant gains in extraction accuracy, technical completeness, and structured output quality compared to leading single-LLM baselines. Key contributions include a scalable ensemble architecture for industrial domains, seamless RAG integration throughout the pipeline, comprehensive quality assessment mechanisms, and a production-ready solution suitable for deployment in knowledge-intensive manufacturing environments.

Index Terms—Multi-LLM ensembles, Retrieval-Augmented Generation, Industrial part specification extraction, Semantic vector search, Confidence-aware synthesis, Knowledge grounding

I. INTRODUCTION

Industrial manufacturing, procurement, and maintenance operations produce vast volumes of unstructured part descriptions embedded in procurement documents, maintenance logs, technical specifications, and supplier catalogs. Extracting and standardizing this information remains a significant bottleneck, often requiring manual processing that is both time-consuming and error-prone. Traditional rule-based and early machine learning approaches fall short in handling the linguistic com-

plexity, domain-specific terminology, and nuanced context typical of industrial part data [16], [17].

While recent advances in Large Language Models (LLMs) have opened new possibilities for automating complex extraction tasks, relying on a single model poses substantial risks in industrial settings. These include hallucination of technical specifications, uneven performance across part categories, limited domain adaptation, and a lack of transparency in confidence estimation [18], [22]. In mission-critical environments, such issues can lead to costly misidentifications, safety risks, and operational inefficiencies.

This paper introduces a novel, production-ready multi-LLM ensemble system designed specifically for high-accuracy extraction of industrial part specifications. The proposed architecture orchestrates nine state-of-the-art LLMs from four major model families, namely Gemini, OpenAI, Mistral, and Gemma, and integrates them within a structured three-phase pipeline that includes parallel extraction, context-aware research enhancement, and intelligent synthesis. Retrieval-Augmented Generation (RAG) is seamlessly integrated throughout all phases, providing semantic access to historical part databases via FAISS-based similarity search [21], [23], which significantly improves grounding and reduces hallucination.

Our system offers several key contributions:

- A scalable ensemble architecture that systematically leverages complementary model strengths for robust, high-quality part information extraction.
- First-of-its-kind integration of RAG with a multi-LLM pipeline for industrial specification grounding, enabling context enrichment and real-time validation.
- Transparent quality assessment, including confidence scores, consensus metrics, and RAG validation indicators, to support informed decision-making.
- A complete, deployable implementation with modular design, configurable model selection, robust error handling, and practical applicability across various industrial domains.

By shifting from a single-model paradigm to a collaborative multi-model approach, our work establishes a new direction for industrial information systems. The architecture is adaptable across sectors such as aerospace, automotive, electronics, and chemical manufacturing, offering the reliability, transparency, and scalability required for mission-critical deployment.

II. LITERATURE REVIEW

Ensemble learning has become a key strategy in improving the robustness and generalization of large language models (LLMs). Techniques are typically categorized into ensemble-before-inference, ensemble-during-inference, and ensemble-after-inference strategies, which leverage multiple models to reduce variance and capitalize on their complementary strengths [4]. Recent work such as LLM-Synergy demonstrates how ensemble voting and dynamic selection approaches can improve domain-specific performance, particularly in high-stakes tasks like medical question answering [8]. Confidence-aware ensemble strategies further contribute to reliable deployment by offering calibrated uncertainty estimates in LLM classification tasks [18]. Additionally, new applications of ensemble methods using advanced models such as Gemini-1.5 have been shown to handle complex industrial tasks such as entity resolution in unstructured datasets [19], [20]. A recent comprehensive review highlights that ensemble deep learning approaches consistently improve performance across classification, generation, and information extraction tasks, while also acknowledging the cost and integration trade-offs [13].

Retrieval-Augmented Generation (RAG) methods have shown great promise in grounding LLM outputs in factual knowledge, addressing the hallucination problem and limitations in model memory. The original RAG architecture introduced in [7] combines parametric generation with non-parametric memory to boost performance on knowledge-intensive tasks. A comprehensive survey [6] categorizes recent RAG strategies and highlights their role in enhancing factual consistency, transparency, and retrieval-guided reasoning in domain-specific applications. These principles are now being extended into multimodal domains, enabling systems to incorporate not only text but also visual and structured data as part of the retrieval context [12]. Semantic similarity search engines such as FAISS have become integral to scalable RAG applications, allowing real-time, dense vector-based retrieval across billions of documents [23].

In the industrial domain, traditional rule-based information extraction systems have been used for tasks like mechanical-electrical-plumbing document processing, but they fall short in capturing nuanced semantics and diverse terminology [16]. As a result, there has been a clear shift toward neural approaches for information extraction from technical text [17]. Broader reviews show that machine learning in smart manufacturing is increasingly focused on structured knowledge generation from unstructured records and the integration of data across systems. Foundational work in manufacturing-specific entity recognition highlighted the challenges in capturing part-specific data

from heterogeneous documents [1], [2], while later studies emphasized the importance of structured data and standardization in product lifecycle management [5].

Efforts like the NIST KEA project [9] and related research focus on synthesizing incompatible information sources to improve manufacturing decision systems. Formal ontologies and enterprise modeling frameworks are central to enabling interoperable systems that align with Industry 4.0 objectives [10], [24]. In parallel, knowledge management systems have been shown to reduce redundancy and improve efficiency in manufacturing environments through centralization and documentation consistency [3]. Predictive maintenance applications also rely heavily on structured knowledge, including historical diagnostics, operating conditions, and equipment metadata [11].

While prior work has demonstrated the value of either LLM ensembles or RAG individually, few approaches have unified these paradigms into a cohesive architecture designed for production-scale industrial deployment. This paper fills that gap by integrating nine LLMs into a collaborative pipeline that includes RAG-enhanced contextual grounding and confidence-aware synthesis, offering a scalable, transparent, and accurate solution for extracting industrial part specifications from unstructured data.

III. SYSTEM ARCHITECTURE

The proposed system is a modular, production-oriented architecture that integrates multiple Large Language Models (LLMs), a Retrieval-Augmented Generation (RAG) subsystem, and a centralized synthesis layer. As illustrated in Figure 1, the architecture is organized into three key layers: semantic retrieval, distributed model orchestration, and synthesis and validation.

Model Integration and Modularity: The system supports nine state-of-the-art LLMs from the Gemini, OpenAI, Mistral, and Gemma families. Each model is accessed via a standardized interface that manages API calls, error recovery, and JSON output validation. Models are dynamically assigned to tasks such as extraction or research based on performance profiles, availability, or deployment constraints. A shared JSON schema ensures consistent outputs across the ensemble, enabling seamless downstream processing. The modular architecture also allows for future model extensions without requiring changes to the core system.

RAG and Knowledge Base Architecture: The RAG subsystem provides semantic context throughout all phases of processing. Input descriptions and knowledge base records are embedded using the `all-MiniLM-L6-v2` model [15], and similarity search is performed using FAISS [14]. The knowledge base is constructed from industrial part records (e.g., Pilog Group datasets) and indexed using flattened schema mappings. Records are encoded into dense vectors for efficient retrieval, enabling similarity-based grounding and validation for all LLM outputs. The system supports structured formats like CSV and JSON, allowing easy adaptation across industrial datasets.

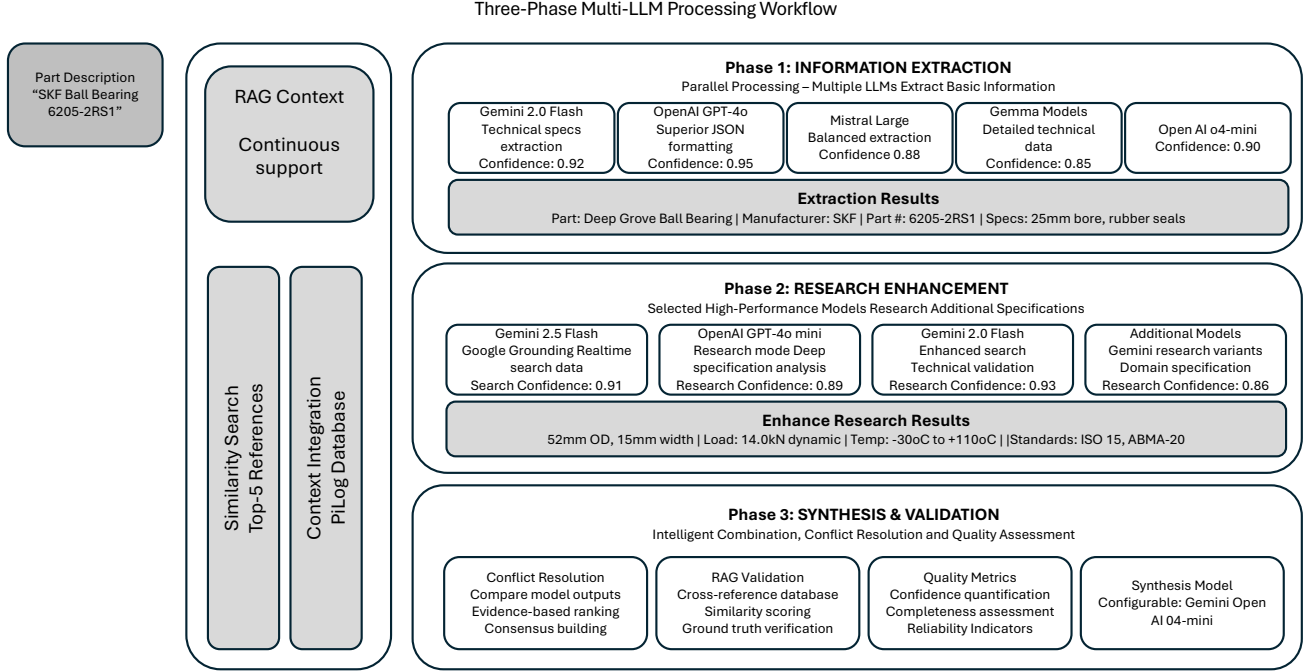


Fig. 1. System architecture of the multi-LLM industrial part extraction framework, showing modular components and model organization.

Synthesis and Configuration Layer: The synthesis layer consolidates outputs from the extraction and research stages. It performs structured conflict resolution, computes confidence scores based on model agreement and RAG similarity levels, and generates final outputs in a validated schema. This layer ensures reliability and interpretability of results. Configuration files allow the system to be tuned for different hardware settings, model selection, and fallback strategies. Compatibility with both CPU and GPU environments supports diverse deployment scenarios in industrial contexts.

IV. METHODOLOGY AND IMPLEMENTATION

The proposed framework follows a structured three-phase methodology designed to coordinate multiple Large Language Models (LLMs) and a Retrieval-Augmented Generation (RAG) subsystem. As illustrated in Figure 2, the workflow spans three interconnected stages: information extraction, research enhancement, and synthesis with validation. Each stage incrementally builds toward a high-confidence, structured specification output, using semantic grounding to mitigate hallucination and ensure domain fidelity.

Phase 1: Information Extraction: The system begins by processing the input part description in parallel using a diverse ensemble of nine LLMs across Gemini, OpenAI, Mistral, and Gemma families. Models were selected based on their complementary strengths: OpenAI models offer high reliability in structured output, Gemini excels at technical reasoning, Mistral balances cost and quality, and Gemma provides efficient

lightweight inference. Each model generates structured JSON with key fields such as part name, manufacturer, part number, and specifications, including confidence scores. Contextual references are retrieved via semantic similarity search using all-MiniLM-L6-v2 embeddings [15] and FAISS [14], chosen for their strong semantic matching and retrieval efficiency. Model discrepancies are retained for downstream synthesis, enhancing fault tolerance and field coverage.

Phase 2: Research Enhancement: Building on the extracted data, selected high-capacity models such as Gemini 2.5 and GPT-4o enter a research mode to augment missing or partial specifications. This phase retrieves additional details such as materials, dimensions, operational tolerances, and industry standards. The RAG system continues to provide context by identifying top-k semantic matches from the industrial knowledge base. Prompts are dynamically adjusted based on earlier extraction results to direct model focus toward uncertain or incomplete attributes. Some models also leverage real-time web grounding when permitted, contributing up-to-date and context-aware enrichment.

Phase 3: Synthesis and Validation: In the final stage, a designated synthesis model aggregates results from both the extraction and research phases. Conflict resolution is performed through majority consensus, RAG alignment scoring, and confidence-level analysis. The synthesis model evaluates internal consistency, cross-model agreement, and external validation signals to compute an overall confidence profile for each data field. The final result is a structured part speci-

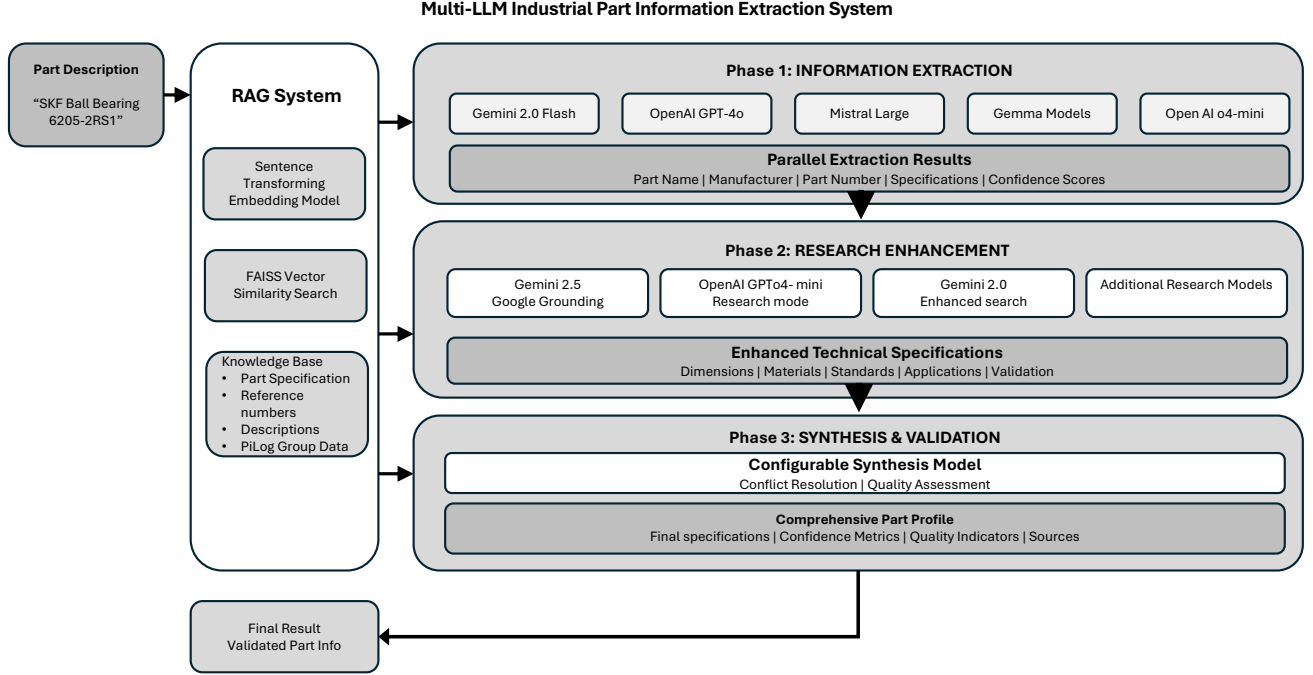


Fig. 2. Three-phase multi-LLM processing workflow with continuous RAG integration. Each stage contributes uniquely to extraction, enrichment, and validation of part specifications.

cation enriched with traceability metadata, including source attribution and quality scores.

RAG Subsystem: The RAG module remains active throughout all three phases. It transforms the part knowledge base, which includes historical specifications, cross-reference numbers, and supplier descriptions, into dense vector representations using a shared embedding model. FAISS performs similarity search at low latency, retrieving top-ranked documents for prompt conditioning. This mechanism enhances factual grounding while maintaining model independence in generation. The top-k retrieval depth and similarity thresholds are configurable to balance specificity and generality.

System Stack and Deployment: The system is implemented in Python with a web interface built using Streamlit. Asynchronous task orchestration is handled using `asyncio`, enabling parallel LLM calls and reducing end-to-end latency. API wrappers are integrated for OpenAI, Google GenAI, and Mistral platforms, with unified schema validation and error handling. The RAG engine supports both CPU and GPU backends and operates on structured input formats such as CSV and JSON. Configurable parameters include model selection, top-k retrieval depth, embedding models, and fallback strategies. The modular design ensures that updates or substitutions to models or knowledge sources do not require architectural changes.

V. RESULTS AND ANALYSIS

To evaluate the performance of our proposed system, **RAGsemble**, we conducted controlled experiments comparing

it against several strong individual LLM baselines, including GPT-4o, Claude, Gemini 2.5, and Grok 3. Each model was evaluated on a standardized industrial dataset containing unstructured part descriptions requiring structured specification extraction. We also included an in-depth benchmark on the GE9X Carbon Fiber Composite Fan Blade, a complex aerospace component with rich technical detail. Five normalized evaluation metrics were used to assess completeness, technical richness, formatting reliability, inter-model divergence, and hallucination behavior.

Evaluation Metrics

We define five normalized metrics to evaluate model performance along dimensions of completeness, specification granularity, structure quality, output variability, and hallucination control.

Information Completeness Score (ICS) measures the fraction of expected core fields successfully extracted by the model:

$$ICS = \frac{|\mathcal{F}_{\text{extracted}}|}{|\mathcal{F}_{\text{expected}}|}, \quad 0 \leq ICS \leq 1$$

Here, $\mathcal{F}_{\text{expected}}$ refers to the required fields for each part category (e.g., part number, material, dimensions), and $\mathcal{F}_{\text{extracted}}$ is the subset successfully captured.

Technical Depth Index (TDI) quantifies the level of fine-grained detail in extracted specifications:

$$TDI = \frac{\sum_{i=1}^n d_i}{n \cdot D_{\max}}, \quad 0 \leq TDI \leq 1$$

Here, d_i is the number of secondary or detailed fields (e.g., tensile strength, voltage rating) extracted for part i , and $D_{\max}$ is the expected maximum number of such fields. Higher values reflect deeper technical insight.

Structured Data Quality (SDQ) reflects the percentage of outputs that are schema-valid and properly formatted:

$$SDQ = \frac{|\mathcal{V}_{\text{valid}}|}{|\mathcal{V}_{\text{total}}|}$$

This metric evaluates whether the final output JSONs conform to the expected schema and can be programmatically parsed without error.

Inter-Model Discrepancy Ratio (IDR) captures the diversity of outputs from different models:

$$IDR = \frac{2}{n(n-1)} \sum_{i < j} \left(1 - \frac{|O_i \cap O_j|}{|O_i \cup O_j|} \right)$$

Here, O_i and O_j are the sets of fields predicted by models i and j , respectively. The metric computes the average Jaccard distance between model pairs. Higher IDR reflects greater output diversity, which can benefit ensemble synthesis but also signals disagreement.

Additional Research Artifacts (ARA) measures the proportion of extraneous or hallucinated details:

$$ARA = \frac{|\mathcal{A}_{\text{extraneous}}|}{|\mathcal{A}_{\text{total}}|}$$

This metric penalizes over-generation by counting surplus attributes not required or grounded. Lower values indicate more concise and reliable output.

TABLE I
NORMALIZED EVALUATION RESULTS ACROSS MODELS

Model	ICS	TDI	SDQ	IDR	ARA
GPT-4o	0.60	0.45	0.80	0.092	0.85
Claude	0.55	0.40	0.90	0.074	0.90
Gemini 2.5	0.65	0.50	0.85	0.118	0.80
Grok 3	0.85	0.80	0.70	0.143	0.55
RAGsemble	1.00	1.00	0.95	0.215	0.40

The results in Table I demonstrate that **RAGsemble** significantly outperforms all individual LLM baselines in terms of completeness, technical depth, and reliability. It achieves full field coverage and maximal specification detail, surpassing even the strongest standalone model, Grok 3. Structured outputs conform to schema 95 percent of the time, minimizing the need for post-processing.

RAGsemble’s higher IDR reflects its intentionally diverse architecture, drawing on varied model perspectives. This variance is not a weakness but a design strength: it feeds into the synthesis layer, which applies consensus filtering, confidence scoring, and RAG validation to resolve conflicts and ensure reliable outputs.

The ARA score of 0.40 confirms that hallucination and over-generation are successfully minimized through RAG-grounded prompt design and validation logic. Compared to

Claude and GPT-4o, which often added verbose or speculative content, RAGsemble maintained tighter control over factual boundaries.

In the case of the GE9X fan blade, RAGsemble extracted 28 technical specifications and 18 quantitative data points, compared to an average of 8.5 and 4, respectively, from single-model baselines. This translates to 3.2× more technical detail, 4.5× more numerical attributes, and 5.58× more validated information units. Despite this output density, the system maintained high SDQ and low ARA, affirming its capability to balance richness with reliability.

Overall, the evaluation confirms RAGsemble as a robust, domain-adaptable framework for structured information extraction. Its ensemble architecture, grounded by RAG and guided by confidence-aware synthesis, consistently delivers complete, accurate, and well-structured part specifications. This makes it highly suitable for deployment in critical industrial workflows where precision and traceability are essential.

VI. USE CASES AND APPLICATIONS

RAGsemble addresses persistent challenges in industrial part information management by enabling scalable, intelligent extraction across sectors. In procurement, it automates the extraction and standardization of part data from supplier catalogs and datasheets, validates information against internal systems, reduces manual errors, shortens cycles, and improves data consistency.

For industrial maintenance, RAGsemble parses manuals and logs to identify replacement parts, check compatibility, align documentation, and enrich metadata. This supports predictive maintenance and improves asset reliability.

Quality assurance teams use it to verify specifications, detect gaps, and ensure consistency across systems. In safety-critical sectors such as aerospace, automotive, defense, and medical devices, its fact-grounded synthesis and traceability features support compliance and performance assurance.

Its modular architecture allows domain-specific tuning of models, datasets, extraction schemas, and validation rules, making it suitable for both high-volume commodity parts and specialized components.

VII. LIMITATIONS AND FUTURE WORK

Current Limitations: RAGsemble faces limitations related to model variability, system complexity, and operational cost. LLM outputs can differ across runs, making reproducibility difficult and requiring consensus-based validation. The reliance on multiple third-party APIs increases cost and risk of service disruption. The sequential three-phase pipeline also introduces latency, which may limit real-time applications.

Domain-specific accuracy may decline when inputs use terminology not well represented in training data, despite RAG grounding. Synthesis bias is another challenge, where final outputs may overweight certain model families. Additionally, RAG performance depends on well-structured, semantically aligned data; inconsistent formatting can reduce retrieval quality and increase hallucination risk.

Future Enhancements: Planned improvements include bias mitigation through model weighting and output cross-validation, as well as dynamic model selection based on input characteristics. RAG will be extended with feedback-driven updates, support for multimodal documents, and more precise prompt-context alignment.

Domain adaptation will involve fine-tuning for specific industries and improved embedding models. Performance upgrades will focus on parallelism, caching, and edge deployment. Quality metrics will expand to include uncertainty estimation and bias detection to improve transparency and trust.

VIII. CONCLUSIONS

This paper introduces RAGsemble, a multi-LLM ensemble with retrieval-augmented generation for extracting industrial part specifications from unstructured text. The three-phase pipeline consists of parallel extraction, targeted research, and synthesis, coordinating nine language models to produce consistent, structured outputs.

On real-world Pilog Group data, RAGsemble outperforms strong single-model baselines in completeness, technical depth, and formatting reliability. The synthesis stage reconciles discrepancies and validates information with FAISS retrieval.

Although the framework increases cost and introduces non-determinism, its modular design improves transparency, adaptability, and accuracy. Planned extensions include dynamic model selection, bias-aware synthesis, and multimodal retrieval.

Overall, RAGsemble demonstrates the value of collaborative LLMs for industrial information extraction by combining ensemble diversity, retrieval grounding, and synthesis to deliver scalable automation in manufacturing.

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation under EPSCoR Award No. OIA-1946391 and the PiLog Group. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the PiLog Group.

REFERENCES

- [1] G. Palshikar, R. Srivastava, S. S. Deshpande, and S. Bhat, "Information extraction for effective knowledge management," *TCS White Paper*, pp. 1–14, 2015.
- [2] M. A. Mohammed, S. Al Mandalawi, H. Maclean, and J. R. Talburt, "Multilingual Customer Record Linkage: A Novel Approach Using LLMs for Cross-Lingual Entity Resolution," in *Proceedings of the 24th International Conference on Information & Knowledge Engineering (IKE'25)*, World Congress in Computer Science, Computer Engineering & Applied Computing, 2025.
- [3] Bloomfire, "Using Knowledge Management Systems in Manufacturing Organizations," *Bloomfire*, Aug. 11, 2022. [Online]. Available: <https://bloomfire.com/blog/knowledge-management-systems-in-manufacturing/>
- [4] Z. Chen, J. Li, P. Chen, Z. Li, K. Sun, Y. Luo, Q. Mao, D. Yang, H. Sun, and P. S. Yu, "Harnessing multiple large language models: A survey on LLM ensemble," *arXiv preprint arXiv:2502.18036*, 2025.
- [5] Z. Wu and C. Liang, "A review and prospects of manufacturing process knowledge acquisition, representation, and application," *Machines*, vol. 12, no. 6, p. 416, 2024.
- [6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.
- [7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [8] H. Yang, M. Li, H. Zhou, Y. Xiao, Q. Fang, and R. Zhang, "One LLM is not enough: Harnessing the power of ensemble learning for medical question answering," *medRxiv*, 2023.
- [9] National Institute of Standards and Technology (NIST), "Knowledge Extraction and Application," *NIST*, Mar. 26, 2025. [Online]. Available: <https://www.nist.gov/programs-projects/knowledge-extraction-and-application>
- [10] F. B. Vernadat, F. T. S. Chan, A. Molina, S. Y. Nof, and H. Panetto, "Information systems and knowledge management in industrial engineering: Recent advances and new perspectives," *International Journal of Production Research*, vol. 56, no. 8, pp. 2707–2713, 2018.
- [11] S. Wan, J. X. Gao, D. Li, and R. D. Evans, "Knowledge management for maintenance, repair and service of manufacturing system," in *Proc. Int. Conf. on Manufacturing Research*, Southampton Solent University, 2014.
- [12] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 2024.
- [13] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.
- [14] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.
- [15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.
- [16] L.-T. Wu, J.-R. Lin, S. Leng, J.-L. Li, and Z.-Z. Hu, "Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web," *Automation in Construction*, vol. 135, pp. 104108, 2022.
- [17] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang, "A survey of information extraction based on deep learning," *Applied Sciences*, vol. 12, no. 19, p. 9691, 2022.
- [18] K. Hovsepian, D. Liu, and S. Murugesan, "Label with confidence: Effective confidence calibration and ensembles in LLM-powered classification," *Preprint*, 2024.
- [19] Mohammed, M. A., Talburt, J. R., Mohammed, A., and Syed, K., "Entity Resolution with Household Movement Discovery Using Google Generative AI," in *Proceedings of the International Conference on Information Technology-New Generations*, Springer, pp. 469–481, 2025.
- [20] M. A. Mohammed, J. R. Talburt, and A. M. Althaf, "Multi-LLM Record Linkage: A Comparative Analysis Framework for Co-Residence Pattern Discovery," in *Proceedings of the 24th International Conference on Information & Knowledge Engineering (IKE'25)*, World Congress in Computer Science, Computer Engineering & Applied Computing, 2025.
- [21] S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan, et al., "Retrieval-augmented generation for natural language processing: A survey," *arXiv preprint arXiv:2407.13193*, 2024.
- [22] P. Goldstein, "LLM Hallucinations: What Are the Implications for Businesses?," *BizTech Magazine*, Feb. 7, 2025. [Online]. Available: <https://biztechmagazine.com/article/2025/02/llm-hallucinations-implications-for-businesses-perfcon>
- [23] H. Jegou, "Faiss: A library for efficient similarity search," *Engineering at Meta*, Mar. 29, 2017. [Online]. Available: <https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/>
- [24] Mohammed, M. A., "Machine learning approaches, technologies, recent applications, advantages and challenges on manufacturing and Industry 4.0 applications," *International Journal of Research in Applied Science and Engineering Technology*, vol. 10, pp. 1114–1121, 2022.

SLB:A Bridge Module of Local SLM and Online LLM for Smart Home Device Control

Peizhe Dong

Information Science

Japan Advanced Institute of Science and Technology

Ishikawa, Japan

dong_peizhe@jaist.ac.jp

Vu Tran

Information Science

Japan Advanced Institute of Science and Technology

Ishikawa, Japan

vu.tran@jaist.ac.jp

Abstract—With the increasing popularity of smart home devices, users have a growing need to communicate with smarter assistants to control smart home devices. However, on the one hand, using online large language models (LLMs) to recognize user commands poses a certain threat to the personal privacy and data security of the user. On the other hand, local small language models (SLMs) cannot always perform well because of the limited computing power of local devices. We propose a method that can better interact with smart home assistants: developing a bridge module that can efficiently connect local SLM and online LLM in a smart home assistant. It first analyzes user intent on the local SLM, and then hands it off to the online LLM for further command generations to specified smart home devices. Our experimental results suggest that this approach can provide users with high-quality services while reducing the risk of privacy leakage. This helps reduce the reliance on large hardware devices when implementing home automation while better protecting user data security and personal privacy.

Index Terms—smart home, small language model, large language model, user privacy, distributed computing

I. INTRODUCTION

The increasing popularity of smart home devices has also highlighted the growing need of the users for more effective voice interactions with smart home assistants. Generally speaking, a smart home assistant is a device that can be remotely or self-controlled using technologies such as IoT and AI-based chatbots in a home with multiple smart home devices, and be able to understand and execute user intents as well as provide efficient user-assistant interaction. Existing commercial off-the-shelf (COTS) smart home assistants, such as Amazon’s Alexa and Apple’s Siri, are relying on pretrained models based on various command datasets and can only be used for specific commands, which may lead to performance degradation when getting unknown tasks. For example, if user gives unspecified commands and does not specify specific devices, it may be difficult for the assistant to generate a reasonable action plan calling smart home devices. Therefore, users and developers must manually add new commands or actions to customize the assistant. Obviously, this is not suitable for ordinary users. At the same time, COTS smart home assistants may contribute to the leakage of user privacy [1].

To overcome these limitations, some studies have tried to integrate LLM to help assistants understand user intent beyond predefined commands [2] [3]. Among them, Sasha uses

ChatGPT to generate action plans for specific user commands and combines conversation history to generate personalized plans [4]. This greatly expands the capabilities of smart home assistants. However, the identifiable personal information of the user such as physical address, email address, home device status and user preferences generated during conversation will be transmitted to the cloud for processing. Once these data are exposed to a third party, it will pose a great threat to the user personal information [5] [6] [7].

At the same time, to ensure the privacy of users, some researchers choose to deploy open-source models locally to provide services to smart home assistants [8]. Unfortunately, because of resource constraints, local devices only support small language models (SLMs), and current SLMs often cannot understand the complex and ambiguous intentions well.

Therefore, users are faced with a dilemma between privacy and performance: online LLMs can provide high-quality services, but there are greater privacy concerns. Local SLMs can better protect user privacy, but due to model size limits, they cannot always provide satisfactory responses.

To solve this dilemma, this study proposes a hybrid architecture based on local SLM and online LLM, which can first refine the user’s intentions and remove privacy information in the local SLM, find the relevant information of the device that the user wants to call, and then send the processed content and device information to the online LLM to generate operation commands to implement the user’s intentions.

II. RELATED WORKS

A. Smart Homes

A smart home system is a system consisting of interconnected smart home devices that are able to monitor and change the home environment. Automation of these devices can improve quality of life, comfort, and resource utilization. Recent research aims to expand the capabilities of smart home systems using machine learning. For example, a voice-controlled home automation system based on language processing and IoT can control four basic home appliances [9].

Currently, commercial solutions (such as Google Assistant, Amazon Alexa, and Xiaomi Xiaoai) provide user-friendly interfaces that can handle a variety of commands and questions from daily communication and setting reminders to basic

device control. At the same time, these smart home assistants also have difficulty handling various complex commands when the user's expression exceeds the pre-programmed preset communication paradigm. If users want to set trigger conditions and actions independently, they must do so through the application or web interface provided by the device manufacturer, which requires a certain level of technology that is beyond the capabilities of most users.

B. Integrating Language Model with Home Automation

Currently, commercial solutions using NLP technology (such as Google Assistant, Alexa, and XiaoAI) provide user-friendly interfaces that can handle a variety of commands and questions from daily communication and setting reminders to device control and home automation. At the same time, these smart home assistants also have difficulty handling various complex commands when the user's expression exceeds the pre-programmed preset communication paradigm. If users want to set trigger conditions and actions independently, they must do so through the application or web interface provided by the device manufacturer, which requires a certain level of technology that is beyond the capabilities of most users.

To overcome these challenges, recent research proposes using the reasoning power of large language models to help smart home assistants better understand and execute user commands [10]. Some universal home automation applications, such as Home Assistant and IFTTT, can help you easily create automation rules to control smart home devices. At the same time, Home Assistant and IFTTT have recently provided some plug-ins with lower difficulty, which makes it possible for users with no technical background to interact with the automation platform using natural language through their own home assistant platform using online LLMs [11]. However, since smart home systems often contain private information such as the user's address, various data in the home environment, device usage history and habits, passing this data to the LLM in the cloud for processing may pose a great threat to the user's information security [12].

At the same time, some studies have tried to use locally deployed SLMs in smart home assistants [13]. Because these SLMs have a small number of parameters, they can complete certain natural language processing tasks on devices ranging from personal computers to Raspberry Pi. This protects the user's data privacy and security very well, and smart home devices can usually complete some basic device control relatively well using smaller data sets. However, when the user's commands tend to be more daily communication and complex device control, SLMs do not always complete the task well, and there is a possibility of more serious hallucinations.

C. Collaborated Edgeprocessing for LLM

Recently, edgeprocessing reasoning has been gaining increasing attention. This approach aims to combine edge computing devices with the LLM in the cloud. It offers advantages in terms of increased computation speed, reduced GPU overhead, and excellent performance in domain-specific tasks [12]

[14] [15]. It also has potential for user privacy protection [16] [17].

III. METHOD

The purpose of the module is to build a bridge between local SLM and online LLM, helping to transfer user's natural language commands better into the powerful LLM capabilities of the cloud. Its overall architecture follows the principle of "edge preprocessing, cloud fine processing", with the aim of achieving a conversational agent that is both intelligent and secure. This concept is deeply rooted in the cutting-edge ideas of modern distributed computing and privacy computing, that is, by intelligently screening and abstracting at the source of data generation, the risks and costs brought by centralized processing can be effectively reduced while retaining core computing capabilities. This module will be developed on the basis of the well-known open source smart home control platform Home Assistant to efficiently access various smart home devices. At the same time, the method of this study is universally transferable.

A. Edge Preprocessing through Local SLM

When the user issues a command, if the command is a text command, it will be processed directly. If it is a voice command, the voice command can be converted into a text command using the various STT functions supported by the home assistant. Then, the command will be routed to a lightweight SLM deployed locally, such as a personal computer or a high-performance edge device.

The task of the local SLM is not to fully understand and execute the command, but to play the role of a "distiller". This follows the information theory and data minimization principle, that is, only the minimum information necessary to complete the task is transmitted.

Its core responsibility is to understand the user's true intention, and on this basis, actively and intelligently remove or generalize the specific details in the command that are highly related to the user's privacy (for example, specific room names, precise address information, personal identification, and even specific time or date habits). This process goes beyond simple keyword filtering or regular expression replacement. It uses the context understanding ability of the SLM to generate an "abstract command" or "generalized intent" that is as faithful to the original meaning as possible in terms of semantics, but more secure in terms of privacy. For example, the user's command "Assist, please automatically turn on the lights in the living room at 6pm when I usually get home from get off work" may be changed to "The user wants to automatically turn on the lighting equipment in the specified area at a specific time" after preprocessing. This conversion retains the core automation and control intent, but strips away personal schedules and specific location information. At the same time, this task is not very difficult, so it can also run well on personal computers or high-performance edge devices.

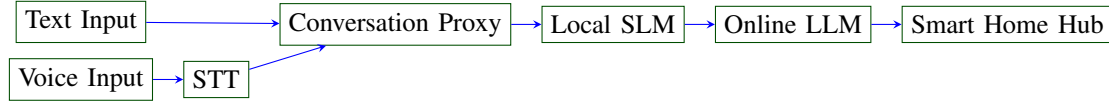


Fig. 1. Module Structure

B. Cloud Refinement through Online LLM

The "abstract commands" generated by the local SLM are then sent to the high-performance LLM in the cloud. The design of this stage reflects the advantages of cloud computing in handling complex tasks, that is, through centralized powerful computing power and massive data support, deeper reasoning and decision-making can be achieved.

The online LLM receives a more general request that has been "privacy-enhanced". Its task is to make precise reasoning, decisions, and execution based on this request, combined with a comprehensive understanding of the smart home environment (provided through system prompts). This includes parsing ambiguous intentions (such as "make the house cooler" may involve the coordinated actions of multiple devices such as air conditioners, fans, curtains, etc.), handling complex multi-step tasks (such as "play music first, then turn the volume to 50%"), maintaining smooth multi-round conversation context (understanding the reference of the pronoun "it"), and most importantly, accurately calling the API provided by Home Assistant to control specific devices. The powerful computing power of cloud services helps to better understand the user's complex commands, which is difficult for the local SLM with limited computing power.

Finally, the natural language response generated by the online LLM is passed back to the user, completing the entire interactive closed loop. This response is based on precise understanding of the sanitized command and accurate manipulation of the Home Assistant environment, and is therefore functionally comparable to a online LLM using the original command directly.

C. Workflow Overview

The entire workflow is a concrete implementation of the above-mentioned hierarchical intelligent processing idea, which ensures the clarity and security of data flow and processing logic.

- **Configuration:** After importing the module into Home Assistant, the user needs to add the corresponding integration and fill in the model name of the local SLM and the URL where the model can be accessed, the model name, URL and API KEY of the online LLM to complete the setting of the dialogue agent. Then, select the dialogue agent in the voice assistant of Home Assistant to complete the configuration.
- **Receiving and Routing:** The user's natural language commands (whether through text after speech recognition or direct text input) are captured by the core dialogue processing system of Home Assistant. According to the

pre-configuration, the command is routed to the dialogue agent (SLB).

- **Local Intent Purification and Abstraction:** SLB first obtains the address of the local SLM service and the preset model name from its configuration. It has a specially designed prompt that explicitly instructs the local SLM model to perform the task of "understanding and generalizing user commands". The core of this prompt is: "Please understand the following user command and rephrase it into a more general form, removing any specific location details or personal habit information. The user command is: [original command]".
- **Cloud Task Execution and Decision Making:** Once the local SLM returns the generalized command, the SLB enters the second phase. It first calls an internal method to obtain a list of all devices in the current Home Assistant environment that are exposed to the conversation agent and their status. It then uses this information and the generalized command to generate a detailed system message. This system message is key for the online LLM to understand and interact with the Home Assistant environment, and it clearly lists the available devices, their names, aliases, and current status. The agent then packages this system message, the generalized user command, and (if any) the previous conversation history to construct a request body that conforms to the API specification of the LLM in use. This request body is sent to the cloud through the configured online LLM API client.
- **Tool Calls and Precise Control:** After receiving the request, the online LLM will perform its complex reasoning process. If specific operations need to be performed, it will return one or more "Tool Calls" or the old version of "Function Call" commands(of Home Assistant). These commands specify the service to be called in Home Assistant (e.g., light.turn_on) and the parameters to be passed to the service (e.g., entity_id: light.living_room). After receiving the response, the SLB parses these tool call commands and executes the corresponding service calls directly in the local context of Home Assistant. This is a key step in achieving physical world control, which is completed entirely within the user's local environment, ensuring the real-time and security of the control commands.
- **Feedback, History maintenance and Presentation:** After all necessary tool calls are executed (which may require one or more round trips), the online LLM generates a final natural language response. After receiving this response, the SLB wraps its content into a Home

Assistant standard ConversationResult object. Before returning the result, the agent updates its internal conversation history (self.history) to save the full context of this interaction (including system prompts, original user messages, tool calls, and final responses) for use in subsequent multi-round conversations. Finally, this ConversationResult object is returned to Home Assistant’s dialogue system, which presents the reply to the user through the configured output interface (such as text or TTS speech synthesis), thus completing a complete, safe and intelligent interaction.

In this way, the SLB module not only technically realizes the integration of local privacy protection and powerful cloud functions, but also reflects a profound thinking on how to responsibly use AI technology in the era of distributed intelligence in terms of design concept. It provides an innovative solution to actively manage and reduce privacy risks without sacrificing user experience.

IV. EXPERIMENTS AND RESULTS

In order to scientifically and objectively evaluate the advantages and disadvantages of the SLB module compared with local SLM and online LLM, we designed a detailed and representative experimental process. This process aims to quantitatively compare the performance of the three solutions in key dimensions through a series of controllable tests.

A. Experimental Preparations

To fully compare the three solutions, we built a test scenario that simulates a real home environment. The scenario revolves around a running Home Assistant, connecting to a variety of typical smart home devices such as lights, thermostats, fans, and media players. We selected three representative processing agents:

Local SLM agent: This agent relies entirely on a medium-sized language model (in this experiment is Gemma3:4B) deployed on local hardware (in this experiment is AMD 9070XT GPU) for command understanding and execution. All computations are completed in the simulation of the user’s home.

Online LLM agent: This agent sends the user’s complete commands directly to a powerful online LLM (in this experiment is Doubao 1.5 pro 256k) for processing. All complex reasoning and decision-making are performed on cloud servers.

SLB agent: This is a new solution we proposed. The user’s commands are first sent to a lightweight SLM (in this experiment is Gemma3:4B) running locally. This model is responsible for quickly understanding the user’s intentions and actively removing or generalizing sensitive information such as specific location and personal habits. Subsequently, this abstract command after “privacy purification” is securely sent to the online LLM (in this experiment is Doubao 1.5 pro 256k), which performs the final reasoning, decision-making and precise control of smart home devices.

Dataset: The dataset of the experiment designed a series of 160 test commands of varying complexity, mainly consisting of simple device interaction instructions (such as “turn on

the lights in the living room”), complex and ambiguous instructions (such as “adjust the air conditioner to a suitable temperature”), instructions that may leak user privacy (such as when the user mentions “my birthday is next month” in the conversation), instructions unrelated to the currently connected smart home devices (such as “call my father”), and instructions related to the contextual content (such as “what I asked you to do half an hour ago”). It covers scenarios involving user privacy leakage, multi-round conversations, and ambiguous intentions, aiming to simulate various interaction scenarios that users may encounter in daily life.

B. Evaluation Metrics

In order to scientifically measure the pros and cons of each solution, we selected several key dimensions for evaluation:

- **Task Accuracy:** evaluates the agent’s ability to correctly understand and execute commands. This includes whether it can accurately identify devices, perform the correct sequence of operations, and handle complex logical relationships.
- **Privacy:** this is the core focus of this study. We use network monitoring to analyze and compare how much of the user’s original information (especially location and behavior habits) is transmitted to the external network during the processing.
- **Contextual Understanding:** tests the agent’s ability to maintain context, understand pronoun references, and handle continuous commands in multiple rounds of dialogue.
- **Tool Calling Success:** measures the frequency with which the agent successfully calls Home Assistant internal services to control physical devices.
- **Cost:** estimates the economic cost of using each solution in the long term, especially the token cost of calling cloud APIs.
- **Response Latency:** measures the total time from when the user issues a command to when the system responds. Although this is a direct factor affecting the smoothness of the user experience, smart home control is not a particularly time-sensitive task, which is not decisive when evaluating.

For measuring response latency and task accuracy, we obtain data through a combination of automated scripts and manual timing. Context understanding and function call success mainly rely on manual evaluation and checking Home Assistant’s system logs. The operating costs are calculated through the cloud service provider’s billing panel.

The most critical thing is the evaluation of privacy protection. When writing the module, we added the function of generating logs to record the content transferred by the local SLM to the online LLM, so that we can clearly determine which original information of the user’s privacy is sent out by analyzing the logs, so as to conduct qualitative analysis of the privacy protection level of each solution.

C. Results and Analysis

After analyzing the collected data, we obtained the following results:

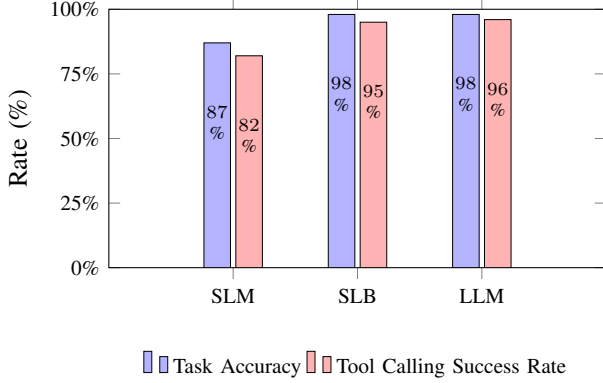


Fig. 2. Performance on Multiple Indicators for Simple Tasks

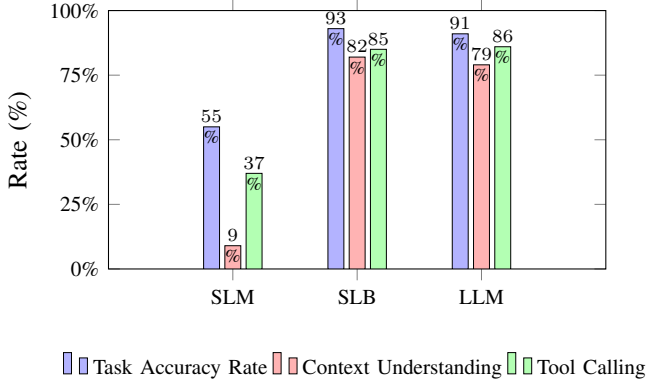


Fig. 3. Performance on Multiple Indicators for Complex Tasks

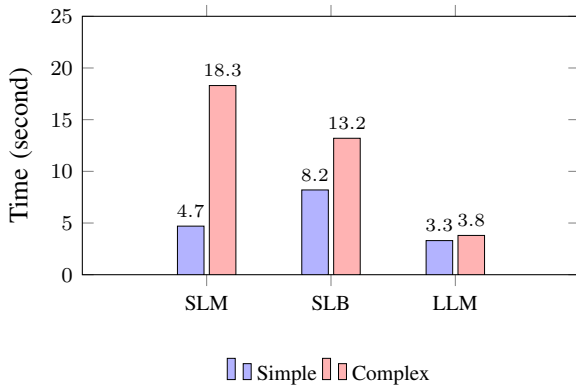


Fig. 4. Average Time Cost for Simple and Complex Tasks

- **Task Accuracy:** in Fig.2, it can be seen that online LLM, leveraging its powerful model capabilities, achieves the highest accuracy for all command types. Local SLM performs reasonably well for simple commands, but struggles on complex tasks. SLB's accuracy is close to

that of online LLM, and it even surpasses online LLM on some complex tasks and tasks with ambiguous intent as it is shown in Fig.3.

- **Privacy:** this is a key differentiator between the various solutions. Local SLM offers the highest level of privacy protection as few privacy information is leaked to the online server. Online LLM, however, uploads user's full, original commands and dialogue to the cloud, posing the highest privacy risk. This demonstrates the core advantage of SLB: Log analysis shows that the information sent to the cloud is abstracted and processed by the local SLM, effectively removing sensitive details from the original commands in most cases. This significantly improves privacy without sacrificing functionality.
- For example, in the experiment, we input "You know, I am already 42 years old now" to simulate a condition that when smart home assistant keeps receiving voice input, the assistant may receive private information that is not relevant to a command from daily talks. The log shows that this input was identified as private information and filtered in local SLM stage, preventing it from entering online LLM. This demonstrates the advantage of our method in protecting user privacy.
- **Context Understanding:** online LLM performed best, while SLB had weaker memory. Local SLM's memory was even more limited, even leading to instances where the preceding content significantly affected the following content (for example, in tasks querying date and time, local SLM was easily misled by the user output in the preceding context).
- **Tool Calling Success Rate:** online LLM and SLB are generally able to successfully call various functions. Local SLM, however, sometimes claimed successful calls when they actually did not achieve the goal.
- **Operational Cost:** The direct operating cost of local SLM is zero (not accounting for hardware depreciation). The cost of online LLM is directly related to the number of API calls and the amount of data(also varies by the price of API). Because SLB sends shorter and more concise commands to the cloud, it consumes fewer tokens(about 57% in this experiment). Therefore, under the token-based billing model, its total cost is greatly lower than that of using online LLM directly.
- **Response Latency:** in Fig.4, it shows that on complex tasks, SLB achieves smaller latency than SLM, while having accuracy better than local SLM and on par with online LLM. This demonstrates the latency performance advantage of the SLB method in handling complex tasks. On simple tasks, even though the latency is high, SLB also has accuracy better than local SLM and on par with online LLM while keeping privacy.

V. CONCLUSION

This paper introduces a method for constructing a module for processing natural language commands of users on a smart home assistant based on local SLM and online LLM, which

aims to provide good services for users to control smart home devices using natural language while ensuring the privacy of users' personal data and information. We chose to first use the local SLM to perform a preliminary analysis of the user's intention and find the device the user wants to use and then hand this purified information to the online LLM for further processing to generate commands for calling the user's relevant devices. Our experiments show that such a hybrid integration architecture performs well in processing users' natural language commands and smart home control tasks as well as privacy protect. Our experience provides valuable insight into the challenges and opportunities in the field of smart home device control with SLM and LLM. Foreseeable future work can focus on optimizing prompt words and selecting language models that are more suitable for smart home systems to improve the quality and response speed of the system.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP23K16954. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the author(s)' organization, JSPS or MEXT.

REFERENCES

- [1] Imran Rahman-Jones, "Apple to pay \$95m to settle Siri 'listening' lawsuit," BBC News, Jan. 2025. [Online]. Available: <https://www.bbc.com/news/articles/cr4rvr495rgo>. [Accessed: Apr. 5, 2025].
- [2] D. Rivkin et al., "AIoT Smart Home via Autonomous LLM Agents," in *IEEE Internet of Things Journal*, vol. 12, no. 3, pp. 2458–2472, 1 Feb. 1, 2025, doi: 10.1109/JIOT.2024.3471904.
- [3] M. R. Tabassum, M. J. Ritchie, S. Mustafiz, and J. Kienzle, "Using LLMs for Use Case Modelling of IoT Systems: An Experience Report," in *Proc. ACM/IEEE Int. Conf. Model. Driven Eng. Lang. Syst. (MODELS)*, 2024, pp. 611–619, doi: 10.1145/3652620.3687810.
- [4] E. King, H. Yu, S. Lee, and C. Julien, "Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models," *Proc. ACM IMWUT*, vol. 8, no. 1, p. 12, Mar. 2024, doi: 10.1145/3643505.
- [5] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh, "ProPILE: Probing Privacy Leakage in Large Language Models," *arXiv.org*, Jul. 04, 2023, <https://arxiv.org/abs/2307.01881>.
- [6] B. K. Sovacool and D. D. Furszyfer Del Rio, "Smart home technologies in Europe: A critical review of concepts, benefits, risks and policies," *Renew. Sustain. Energy Rev.*, vol. 120, p. 109663, 2020, doi: 10.1016/j.rser.2019.109663.
- [7] X. Lei, G. -H. Tu, A. X. Liu, C. -Y. Li and T. Xie, "The Insecurity of Home Digital Voice Assistants - Vulnerabilities, Attacks and Countermeasures," 2018 IEEE Conference on Communications and Network Security (CNS), Beijing, China, 2018, pp. 1–9, doi: 10.1109/CNS.2018.8433167.
- [8] E. King, H. Yu, S. Vartak, J. Jacob, S. Lee, and C. Julien, "Thoughtful Things: Building Human-Centric Smart Devices with Small Language Models," *arXiv.org*, 2024, <https://arxiv.org/abs/2405.03821v1> (accessed Jul. 29, 2025).
- [9] F. Alwahedi, A. Aldhaferi, M. A. Ferrag, A. Battah, and N. Tihanyi, "Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models," *Internet Things Cyber-Phys. Syst.*, vol. 4, pp. 167–185, 2024, doi: 10.1016/j.iotcps.2023.12.003.
- [10] B. Xiao, B. Kantarci, J. Kang, D. Niyato and M. Guizani, "Efficient Prompting for LLM-Based Generative Internet of Things," in *IEEE Internet of Things Journal*, vol. 12, no. 1, pp. 778–791, 1 Jan. 1, 2025, doi: 10.1109/JIOT.2024.3470210.
- [11] Home Assistant, "OpenAI Conversation," Home Assistant, 2025, https://www.home-assistant.io/integrations/openai_conversation/ (accessed Jul. 29, 2025).
- [12] F. Cai, D. Yuan, Z. Yang, and L. Cui, "Edge-LLM: A Collaborative Framework for Large Language Model Serving in Edge Computing," in *Proc. IEEE Int. Conf. Web Serv. (ICWS)*, Shenzhen, China, 2024, pp. 799–809, doi: 10.1109/ICWS62655.2024.00099.
- [13] Z. Yin, M. Zhang, and D. Kawahara, "Harmony: A Human-Aware, Responsive, Modular Assistant with a Locally Deployed Large Language Model," *arXiv.org*, 2024, <https://arxiv.org/abs/2410.14252>.
- [14] Y. Liu et al., "Towards Harnessing the Collaborative Power of Large and Small Models for Domain Tasks," *arXiv.org*, 2025, <https://arxiv.org/abs/2504.17421> (accessed Jul. 29, 2025).
- [15] C. Shao, X. Hu, Y. Lin, and F. Xu, "Division-of-Thoughts: Harnessing Hybrid Language Model Synergy for Efficient On-Device Agents," *arXiv.org*, 2025, <https://arxiv.org/abs/2502.04392> (accessed Jul. 29, 2025).
- [16] Y. Wang, Y. Lin, X. Zeng, and G. Zhang, "PrivateLoRA For Efficient Privacy Preserving LLM," *arXiv.org*, 2023, <https://arxiv.org/abs/2311.14030> (accessed Jul. 29, 2025).
- [17] Othmane Friha, M. A. Ferrag, Burak Kantarci, Burak Cakmak, Arda Ozgun, and Nassira Ghoualmi-Zine, "LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 5799–5856, Jan. 2024, doi: <https://doi.org/10.1109/ojcoms.2024.3456549>.

Multimodal Large Language Model for out-of-context Problems in Fake News Detection

Dang Le Nguyen Hai^{*†}, Khang Vo Hoang Nhat^{¶‡}, Minh Nguyen Le[†], Tho Quan Thanh^{*§}

^{*}Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam

[†]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

[¶]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Contributing authors: dang.le246@hcmut.edu.vn, Khang.Vo@mbzuai.ac.ae, nguyenml@jaist.ac.jp, qttho@hcmut.edu.vn

Abstract—The proliferation of multimodal misinformation, particularly out-of-context image-caption pairs, has become a growing concern in online environments. Unlike fabricated content, out-of-context misinformation consists of independently truthful elements that, when combined, create misleading information. A key strategy to mitigate such misinformation is to recover the original context of the image. While recent research has explored automated multimodal approaches that retrieve and analyze external evidence related to image-text pairs, these methods often overlook the contextualization step, focusing solely on veracity prediction. In this work, we propose a new method that explicitly incorporates both contextual understanding and veracity prediction. Our approach first predicts the contextual elements from an image and then leverages these elements to assess the consistency of the paired caption. Experiments on a real-world benchmark show that our method achieves 77.6% accuracy, substantially outperforming existing multimodal baselines. Ablation studies further confirm the crucial role of contextual modeling in improving out-of-context detection.

Index Terms—Multimodal Fact-checking, out-of-context Detection, Multimodal Large Language Models.

I. INTRODUCTION

The rapid spread of misinformation in the digital age, exacerbated by social media and advancements in generative AI, poses a significant challenge. Multimodal misinformation, particularly that combining text and images, is especially pervasive and more impactful than text-only misinformation [1]. Consequently, deceptive actors now leverage Deepfakes and other advanced generative models [2]–[4] to generate highly realistic images, further complicating the detection of visual misinformation. However, one of the most deceptive types of multimodal misinformation is *out-of-context* (OOC) misinformation. This involves pairing unaltered, authentic images with misleading captions to distort their original meaning and mislead the public. A common example is counterfeit advertisements that use authentic product photos to deceive consumers into believing the advertised items are genuine. These ads commonly appear on e-commerce, exploiting the visual credibility of real product photos while selling low-quality replicas. Unlike manipulated or fabricated content, OOC misinformation is especially difficult to identify because both the image and text may be truthful when considered separately. However, their combination presents a false narrative

by altering the original context. This subtle form of manipulation makes it exceptionally challenging for conventional fact-checking techniques, underscoring the urgent priority of developing automated methods for multimodal OOC detection.

Early efforts [1], [5], [6] to detect OOC misinformation primarily relied on comparing image-text similarity or retrieving external evidence, but these approaches often lacked transparency and failed to provide clear explanations for their predictions. Subsequent attempt [7] introduced attention-based visualizations to improve interpretability, though these do not convey meaningful reasoning. While recent researches [8]–[10] has leveraged the power of *Multimodal Large Language Models* (MLLMs) for improved accuracy and explainability, these models often produce false predictions or irrelevant explanations due to their limited contextual understanding, often relying solely on textual evidence from reverse image searches rather than thoroughly investigating image context. Recognizing these limitations, we introduce a novel, two-phase method designed to enhance both contextualization and veracity prediction. First, our model predicts relevant contextual items directly from a given image. Then, it uses these items and the image to generate a range of captions. Finally, each generated caption is compared to the original, with a majority voting mechanism for determining the final veracity decision.

Our contributions are summarized as follows.

- We propose a novel two-phase framework for OOC misinformation detection that integrates contextual understanding and veracity prediction.
- Our method explicitly predicts six distinct context items to enrich contextual understanding and enhance the framework’s explainability.
- We conduct extensive experiments, evaluating our method against state-of-the-art approaches on a real-world dataset, demonstrating its superior overall accuracy. We also perform ablation studies to better understand the contribution of each module in our approach.

II. RELATED WORKS

A. Close-domain Methods (Internal Checking)

Internal checking methods assess image-text consistency using pre-trained multimodal models. Early approaches relied on

[†]These authors contributed equally to this work.

[§]Corresponding author.

semantic similarity [1], [5], [6], while later studies employed models like CLIP [11] and VisualBERT [12] for classification. However, these models lack interpretability, making it difficult to explain why an image-text pair is misleading. To address this, recent works have integrated *Large Language Models* (LLMs) to generate explanations for OOC detection. Shalabi et al. [10] fine-tuned MiniGPT-4 for better reasoning over image-text relationships, while Zhang et al. [13] introduced abstract meaning representation graphs to generate queries for vision-language models. Additionally, Jiahui Geng [14] suggested that MLLMs, such as ChatGPT, could leverage stored knowledge for OOC detection when effectively prompted. Despite these improvements, internal checking methods remain limited. Images do not always fully support their captions, making detection difficult without external context. Moreover, most models focus on classification rather than verification, highlighting the need for approaches that incorporate external evidence.

B. Open-domain Methods (External Checking)

To overcome the limitations of internal methods, open-domain approaches retrieve external evidence for verification. Some methods compare claims against fact-checked reference datasets [2], [5], while others, such as [7], [15], perform cross-modal entity consistency checks by extracting entities from captions and retrieving relevant images. Further improvements involve optimizing evidence selection. RED-DOT [16] enhances claim-evidence consistency by separately analyzing text and images, while Abdelnabi et al. [7] compute aggregated consistencies across multiple evidence sources. However, many of these methods rely on attention-based scoring, which lacks strong debunking capabilities. Recent research also explores MLLMs with instruction tuning [8] external tool integration [17]. Papadopoulos et al. [18] found that simple classifiers, such as random forests trained on embeddings of images, captions, and retrieved evidence, can achieve strong performance. However, they also revealed that many models rely on surface-level patterns rather than deep factual reasoning, underscoring the need for future research in context-aware verification.

III. METHODOLOGY

In this work, we propose a two-stage framework for detecting OOC multimodal misinformation. The proposed pipeline is designed to explicitly incorporate contextual understanding into the veracity assessment of image-caption pairs. Our method is motivated by the observation that the key idea to solve OOC misinformation problems lies in reconstructing the original context of the image, rather than relying solely on semantic similarity or external text retrieval. The overall architecture consists of two sequential phases: Contextual Items Prediction and Veracity Prediction.

A. Contextual Items Prediction

Following previous works [7], [17], we employ a reverse image search strategy using the Google Vision API¹ to acquire initial textual evidence for a given query image. The API provides two primary types of contextual features: webpage titles and identified visual entities. As introduced in the previous studies [17], we further augment the textual evidence by developing a custom web crawler. This visits corresponding webpages, identifies the image via URL or perceptual hashing, and extracts the associated image captions.

The retrieved textual evidence is then ranked based on its semantic similarity to the image. This involves calculating the cosine similarity between the image’s embedding and the textual evidence’s embedding. Both embeddings are generated using the Contrastive Language-Image Pre-training (CLIP) model [11]. Based on this similarity score, we select the `top_k` evidence items that are most closely aligned with the image, which are then used as the input to the LLMs model for question answering. Extending the previous work [17] that focused solely on the contextualization task; our proposed framework predict six distinct context items: motivation, location, date, event, object, and people. These were chosen because prior study shows that such elements are the most frequently manipulated to create misleading narratives [17]. Each context item is obtained by answering a targeted question:

- **Motivation:** "Why was the image taken?" is used to uncover the purpose behind capturing the image.
- **Location:** "Where was the event in this image taken?" aims to identify the location of the scene.
- **Date:** "When was the event in this image taken?". The date can reveal the image’s relationship to specific events, social trends, and cultural moments.
- **Event:** "Which event is depicted in the image?" helps describe the specific incident or occasion shown.
- **People:** "Who is shown in the image?" is used to recognize notable individuals, especially in politically or socially significant imagery.
- **Object:** "Which animals, plants, buildings, or objects are shown in the image?" is particularly effective when the image centers on a specific subject, such as a well-known landmark or object.

After ranking, all textual evidence is matched with its corresponding context question and fed into the Llama 3.1 8B model [19] using the prompt template shown in Figure 2 to predict the relevant context item. If the textual evidence do not contain enough information to answer the question, the model is instructed to return the response "Unknown".

B. Veracity Prediction

The final stage of our pipeline focuses on veracity prediction, aiming to determine if an image-caption pair is OOC. This phase directly builds on the contextual items in the former phase, using it to assess the semantic alignment between

¹<https://cloud.google.com/vision>

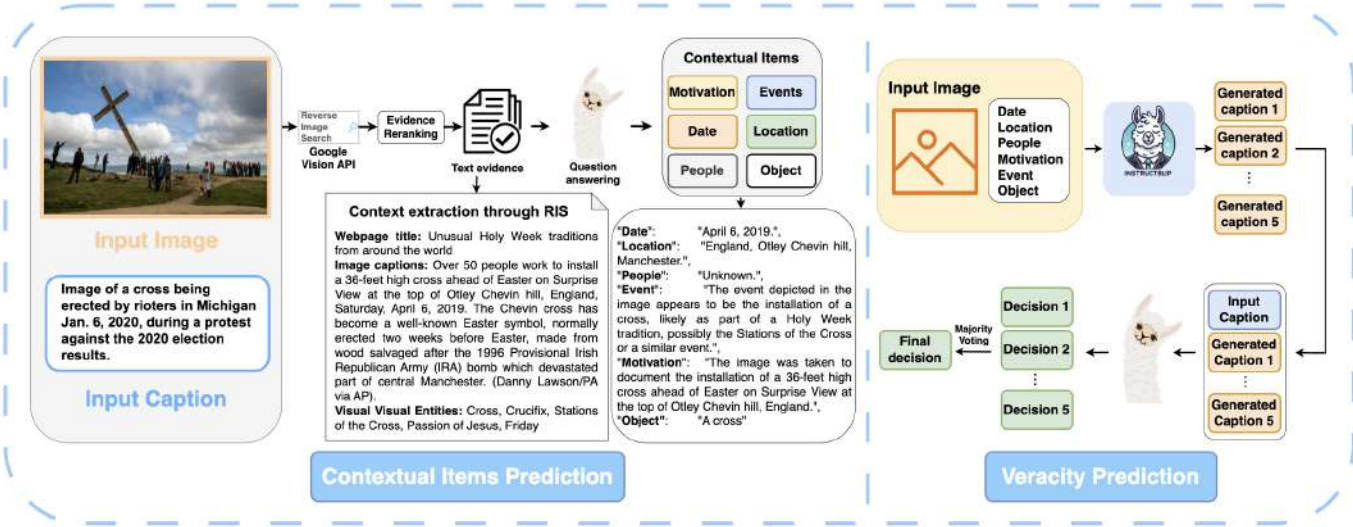


Fig. 1: Overall framework of our two-phase out-of-context (OOC) misinformation detection method, which consists of (1) Contextual Item Prediction and (2) Context-Guided Caption generation for Veracity Prediction.

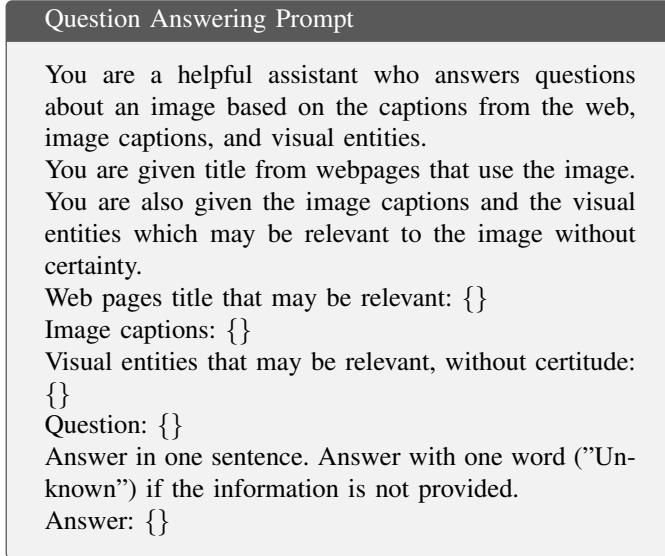


Fig. 2: Prompt template for question answering with Llama 3.

the visual content and its accompanying caption. To do this, we fine-tune the InstructBLIP model [20] for context-aware image captioning, and then generate new captions using the image and its predicted context. Finally, we compare this generated caption with the original input caption to detect inconsistencies, thereby identifying potential OOC instances.

1) *Instruction Fine-tuning InstructBLIP Model*: InstructBLIP, a vision-language model optimized for instruction-following tasks [20], has shown promising results in multi-modal applications. However, previous studies [8] have identified limitations in its captioning ability. Specifically, a tendency to produce coarse-grained descriptors such as “person”, “man”, or “woman”, instead of fine-grained named entities

like “Ed Sheeran”. This lack of specificity poses challenges in downstream tasks such as misinformation detection and fact-checking, where accurately identifying people, places, and events is essential. Furthermore, the contextual metadata retrieved alongside an image often contains noisy or irrelevant details, making it difficult for the model to determine which pieces of information are informative. To address these challenges, we fine-tune the InstructBLIP model to better generate image captions. In this phase, we freeze both the image encoder and the LLM components, and only train the Q-Former [21] module, following the strategy of previous work [20]. This selective tuning encourages the Q-Former to align visual features from diverse news-domain imagery with textual embeddings of fine-grained named entities in the LLM’s semantic space.

For training, we utilized a dataset that provides ground-truth image captions along with structured contextual information from [17]. The training set consists of instruction-image-caption triplets, denoted as $(X = \{(T_i, I_i, C_i)\}_{i=1}^N$, where each element includes an instruction T_i , an image I_i , and its ground-truth caption C_i . The model utilizes a frozen image encoder $\mathcal{E}_v$ to extract visual features from images, a frozen large language model $\mathcal{M}$ for text generation, and a trainable Q-Former module $\mathcal{Q}_\theta$ whose parameters θ are updated during training. For each epoch, the algorithm iterates over mini-batches of data. Within each mini-batch, visual embeddings are extracted from the images using the frozen image encoder. The instructions are tokenized and embedded, and then both the visual and instruction embeddings are passed through the Q-Former to obtain joint representations. These representations are concatenated to form soft prompts, which are fed into the frozen language model to generate predicted captions. The model computes the cross-entropy loss between the predicted captions and the ground-truth captions, and the Q-Former

parameters are updated accordingly using the optimizer. The above process is illustrated in the Algorithm 1.

2) *Generating Captions and Comparison*: We leveraged the fine-tuned InstructBLIP model to generate enriched captions for images by incorporating contextual items from the previous phase. Given an input image, the corresponding contextual items first retrieved. This information, alongside the image itself, is then passed through the fine-tuned InstructBLIP model to generate a refined caption that integrates both visual and textual context. To improve robustness, we generate multiple candidate captions, with a default of 5 generated captions. These variations ensure diversity in the generated captions while preserving semantic consistency. The newly generated captions serve as key components in our final verification step, where we employ a Llama 3.1 8B [19] as a judge to determine whether a given caption is OOC with respect to the image. For each generated caption, the LLM is presented with the caption under evaluation and the generated caption, along with the retrieved context. The model is then prompted to assess the semantic alignment between the caption under evaluation and the candidate caption. The LLM responds with either “Align” or “Not align” (also means “out of context” in this case), indicating whether the two captions convey consistent information about the image. To derive final decision, we aggregate the alignment decisions across all candidate captions. If the number of “Align” responses exceeds the number of “Not align” responses, the image-text pair is classified as aligned; otherwise, it is classified as OOC. The LLM is prompted with the structured instruction illustrated in Figure 3.

Prompt for Inconsistency checking

You are a helpful assistant who can detect the inconsistency between 2 captions of an image. Focus solely on factual consistency, especially event, time, identity consistency, scene, and subject consistency. Do not assess differences in level of detail.

Determine if these captions provide consistent information about the same scene/event despite differences in detail. If there is inconsistent information, respond with “Not align.” Otherwise, respond with “Align.”

Caption 1: {eval_caption}
Caption 2: {generated_caption}

Respond with “Align” or “Not align.” Think step-by-step.

Your answer:

Fig. 3: Inconsistency checking prompt for Llama 3.1 Instruct

IV. EXPERIMENTS AND RESULTS

A. Dataset Preparation

Previous studies [7], [8], [16] frequently relied on the NewsCLippings dataset [1], which was constructed from the VisualNews dataset [22]. The VisualNews dataset contains real news image-caption pairs originally intended for image captioning. To generate “fake” OOC pairs, the creators intentionally mismatched original image and caption pairs with others from different articles based on the semantic similarity. This “similarity” is determined using features from models like CLIP, ResNet [23], and SBERT-WK [24]. For example, they might pair a caption about one political rally with an image of a different but visually similar rally [25]. This makes the OOC examples subtle, challenging models to find hidden inconsistencies while keeping the original image and caption intact.

However, this method has a big drawback: models can sometimes “cheat” [18]. Instead of truly understanding the context, they might just look for these surface-level similarities to get a high overall accuracy. To get a more realistic test, we used the VERITE dataset [25] for our evaluation. Unlike NewsCLippings, VERITE contains real-world OOC examples sourced from fact-checking websites, which actually appeared as misinformation. This makes VERITE a much tougher and more realistic test for our model, ensuring it can detect subtle and complex misinformation.

B. Baselines

We compare our proposed method with six state-of-the-art approaches designed to detect OOC misinformation. DT [6] builds on top of CLIP by adding extra transformer layers to better understand the relationship between images and text. CCN [7] uses external evidence and checks the consistency between the image and text, as well as between the claim and retrieved information from the web. RED-DOT [16], AITR [18] and MUSE-MLP [18] also build on CLIP by adding transformer layers and use both image and text evidence. They also apply a filtering step to keep only the most relevant pieces of information. SNIFFER [8] uses two signals: one checks for mismatches between the image, caption, and visual entities using a fine-tuned InstructBLIP model; the other compares the caption with related text evidence using a Vicuna language model [26]. After all, the model combines two signals to get the final prediction. These models represent strong baselines for testing the effectiveness of new methods in detecting OOC misinformation in both synthetic and real-world datasets.

C. Experimental Setups

First, we use the Google Vision API to retrieve up to 30 top-ranked webpages related to each image. Metadata, including titles, authors, descriptions, main text, and image captions, is extracted using Trafilatura [27], resulting in 6,756 multilingual evidence items. Next, we apply the CLIP-ViT-L/14 model [11] to compute cosine similarity between each evidence snippet and the input image, selecting the top three most relevant texts for downstream reasoning. For question

Algorithm 1 Instruction Fine-tuning of InstructBLIP

Input: Instruction-image-caption triplets $X = \{(T, I, C)\}$; Frozen image encoder $\mathcal{E}_v$, frozen LLM $\mathcal{M}$, trainable Q-Former $\mathcal{Q}_\theta$; Optimizer $\mathcal{O}$, learning rate schedule $\tau(t)$, number of epochs N .

Output: θ

```
1: for  $epoch = 1 \dots N$  do
2:   for minibatch  $\{(T_i, I_i, C_i)\}_{i=1}^B \in X$  do
3:     Compute visual embeddings:  $\mathbf{v}_i = \mathcal{E}_v(I_i)$ 
4:     Tokenize instruction  $T_i$  and embed as  $\mathbf{t}_i$ 
5:     Compute Q-Former outputs:  $\mathbf{q}_i = \mathcal{Q}_\theta(\mathbf{v}_i, \mathbf{t}_i)$ 
6:     Form soft prompt:  $\text{Prompt}_i = [\mathbf{t}_i \parallel \mathbf{q}_i]$ 
7:     Predict caption:  $\hat{C}_i = \mathcal{M}(\text{Prompt}_i)$ 
8:     Compute cross-entropy loss:  $\mathcal{L} = \text{CE}(\hat{C}_i, C_i) = -\sum_{i=1}^B \sum_t \log P_{\mathcal{M}}(C_{i,t} \mid \text{Prompt}_i, C_{i,<t})$ 
9:     Update  $\theta$ :  $\theta \leftarrow \theta - \tau(t) \cdot \nabla_\theta \mathcal{L}$ 
10:  end for
11: end for
```

answering, we use the LLaMA 3.1 8B Instruct model [19] with a temperature of 0.2 to ensure controlled responses. In parallel, we fine-tune the InstructBLIP model, which integrates the Vicuna-7B language model [26] and ViT-G/14 vision encoder [28], freezing all components except the Q-Former. The model is trained for 10 epochs with a batch size of 4 using a cosine learning rate schedule that peaks at 10^{-5} , with warmup starting at 10^{-8} . The maximum input and output sequence length is set to 512 tokens, and images are resized to 224×224 pixels. Finally, in the veracity verification phase, we employ LLaMA 3.1 8B Instruct in deterministic mode with low temperature to compare the input caption against five sampled captions, producing alignment predictions that are aggregated via majority voting to determine the final output. All experiments were conducted utilizing a single NVIDIA A40 GPU.

D. Main Results

TABLE I: Result of our method versus baseline methods on the VERITE dataset

Method	Accuracy (%)
CCN [7]	67.3
DT [6]	72.7
RED-DOT [16]	73.9
MUSE-MLP [18]	74.8
SNIFFER [8]	74.0
AITR [18]	76.5
Our Proposed Method	77.6

All methods experience a significant decline in accuracy when switching from synthetic data (NewsCLippings) to the real-world dataset. This performance drop reflects the complexities of real-world OOC misinformation detection, which involves more intricate semantic patterns and higher levels of noise. Despite this overall reduction in performance, our proposed method achieves the highest accuracy of 77.6%, outperforming several competitive baselines, including AITR [18] (76.5%), MUSE-MLP [18] (74.8%), RED-DOT [16] (73.9%), SNIFFER [8] (74.0%), DT [6] (72.7%), and CCN [7] (67.3%). This strong result underscores the robustness

and adaptability of our approach in dealing with the nuanced and diverse nature of real-world OOC misinformation. The improved accuracy reflects the model’s superior capability to detect subtle semantic inconsistencies and exploit contextually relevant features across modalities. Moreover, these findings highlight the generalization strength of our method and validate the effectiveness of our two-phase architecture, which first performs context prediction before executing veracity prediction. By establishing a strong contextual representation, the model is better positioned to assess the semantic alignment between image and caption.

E. Ablation Studies

TABLE II: Ablation Study of our proposed method’s components

Contextualization	Fine Tuning Instruct-Blip	Ranking Evidence	Accuracy (%)
×	×	×	61.9
✓	×	×	69.8
✓	✓	×	74.8
✓	✓	✓	77.6

To evaluate the contribution of each component in our proposed framework, we conducted ablation studies, keeping the Llama-3.1-8B-Instruct model fixed for final verification. The results, summarized in Table II, highlight the following insights.

Our baseline, using only Llama-3.1-8B-Instruct without fine-tuned InstructBLIP and no contextualization or ranking, achieved 61.9% accuracy. This showed its limited ability to detect subtle inconsistencies without any specialized adaptation or evidence refinement. Next, introducing the contextualization module, which structured raw evidence for InstructBLIP’s prompt, significantly improved accuracy to 69.8% (a 7.9% gain). This demonstrated the critical role of structured contextual information. Further, instruction fine-tuning the captioning model led to a notable accuracy of 74.8% (a 12.9% improvement from the baseline). This gain indicates that fine-tuning

enables the model to capture contextual nuances and produce more informative captions. Finally, activating all components, including the evidence ranking strategy to prioritize relevant contextual information, yielded the highest performance. The system achieved its peak accuracy of 77.6%, representing a total improvement of 15.7 percentage points over the baseline. These findings confirm the contribution of each component to our model’s effectiveness.

V. CONCLUSION

In this work, we proposed a novel two-phase framework for detecting OOC misinformation by integrating contextual understanding and veracity prediction. While current methods focus solely on image-caption consistency, our approach emphasizes recovering the original context of an image before assessing the consistency between image and the according caption. Our framework systematically predicts relevant contextual items from a given image, subsequently utilizing these elements to evaluate caption consistency. This involves fine-tuning an InstructBLIP model for context-aware captioning, generating a new caption, and comparing it with the original to identify inconsistencies. Experiments on the VERITE dataset demonstrate that our method outperforms state-of-the-art baselines. Additionally, the ablation studies highlight the contribution of each component in the pipeline. At the same time, we recognize two main limitations. First, the framework requires multiple LLM inferences, which increases computational cost; however, this is the trade-off compared to previous methods that rely solely on the internal knowledge of models. Second, it relies solely on the Google Vision API for evidence retrieval, which may return irrelevant or noisy evidence.

ACKNOWLEDGMENT

We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology, Vietnam National University Ho Chi Minh City for this study.

REFERENCES

- [1] G. Luo, T. Darrell, and A. Rohrbach, “NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media,” in *Proceedings of the EMNLP*, Nov. 2021, pp. 6801–6817.
- [2] E. Sabir, W. AbdAlmageed, Y. Wu, and P. Natarajan, “Deep multimodal image-repurposing detection,” in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM ’18, 2018, p. 1337–1345.
- [3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [4] X. Wang, J. Huang, S. Ma, S. Nepal, and C. Xu, “Deepfake disrupter: The detector of deepfake is my friend,” in *Proceedings of the IEEE Conference on CVPR*, 2022, pp. 14900–14909.
- [5] A. Jaiswal, Y. Wu, W. AbdAlmageed, I. Masi, and P. Natarajan, “Aird: Adversarial learning framework for image repurposing detection,” in *Proceedings of the IEEE/CVF Conference on CVPR*, 2019.
- [6] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. Petrantonakis, “Synthetic misinformers: Generating and combating multimodal misinformation,” in *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, 2023, p. 36–44.
- [7] S. Abdelnabi, R. Hasan, and M. Fritz, “Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources,” *2022 IEEE/CVF Conference on CVPR*, 2021.
- [8] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, “Sniffer: Multimodal large language model for explainable out-of-context misinformation detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13052–13062.
- [9] G. Delvecchio, H. H. Nguyen, and I. Echizen, “Zero-Shot Warning Generation for Misinformative Multimodal Content,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE Computer Society, 2025.
- [10] F. Shalabi, H. Felouat, H. H. Nguyen, and I. Echizen, “Leveraging chat-based large vision language models for multimodal out-of-context detection,” in *Advanced Information Networking and Applications*, L. Barolli, Ed. Cham: Springer Nature Switzerland, 2024, pp. 86–98.
- [11] S. Wang, Y. Duan, H. Ding, Y.-P. Tan, K.-H. Yap, and J. Yuan, “Learning transferable human-object interaction detector with natural language supervision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 939–948.
- [12] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visualbert: A simple and performant baseline for vision and language,” 2019.
- [13] Y. Zhang, L. Trinh, D. Cao, Z. Cui, and Y. Liu, “Interpretable detection of out-of-context misinformation with neural-symbolic-enhanced large multimodal model,” 2024.
- [14] J. Geng, Y. Kementchedjheva, P. Nakov, and I. Gurevych, “Multimodal large language models to support real-world fact-checking,” 2024.
- [15] E. Müller-Budack, J. Theiner, S. Diering, M. Idahl, and R. Ewerth, “Multimodal analytics for real-world news using measures of cross-modal entity consistency,” in *Proceedings of the ICMR*, 2020, p. 16–25.
- [16] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, “Red-dot: Multimodal fact-checking via relevant evidence detection,” *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2025.
- [17] J. Tonglet, M.-F. Moens, and I. Gurevych, “‘image, tell me your story!’ predicting the original meta-context of visual misinformation,” in *Proceedings of the EMNLP*, 2024, pp. 7845–7864.
- [18] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, “Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection?” in *2025 IEEE/CVF WACV*. IEEE, 2025, pp. 5041–5050.
- [19] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [20] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: towards general-purpose vision-language models with instruction tuning,” in *Proceedings of the 37th International Conference on NeurIPS*, 2023.
- [21] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models,” in *Proceedings of the 40th ICML*, 2023.
- [22] F. Liu, Y. Wang, T. Wang, and V. Ordonez, “Visual news: Benchmark and challenges in news image captioning,” in *Proceedings of the EMNLP*, 2021, pp. 6761–6771.
- [23] B. Zhou, A. Lapedriza, A. Torralba, and A. Oliva, “Places: An image database for deep scene understanding,” *Journal of Vision*, vol. 17, no. 10, p. 296, 2017.
- [24] B. Wang and C.-C. J. Kuo, “Sbert-wk: A sentence embedding method by dissecting bert-based word models,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2146–2157, 2020.
- [25] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, “Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias,” *Int. J. Multim. Inf. Retr.*, 2023.
- [26] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” March 2023.
- [27] A. Barbaresi, “Trafilatura: A web scraping library and command-line tool for text discovery and extraction,” in *Proceedings of the 11th IJCNLP: System Demonstrations*, H. Ji, J. C. Park, and R. Xia, Eds., Online, Aug. 2021, pp. 122–131.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.

Fusion-based Document Retrieval for Low-Resource Vietnamese Legal Texts

Van-Khanh Tran*

Institute of Applied Science and Technology - IAST
University of Information and Communication Technology - ICTU
Thai Nguyen, Vietnam
tvkhanh@ictu.edu.vn

Tien-Bang Tran

Hanoi University of Science and Technology
Hanoi, Vietnam
bangtrantien8308@gmail.com

Abstract—Efficient retrieval of relevant legal documents is crucial in the legal domain, particularly for languages with limited computational resources such as Vietnamese. Vietnamese legal texts pose significant challenges due to their complex structure, formal language, and specialized terminology. Traditional retrieval systems often struggle to deliver accurate and relevant results in this context. This paper presents a fusion-based document retrieval system specifically designed for Vietnamese legal texts. By combining multiple retrieval methods, including traditional lexical matching and modern neural network-based techniques, we leverage their complementary strengths to enhance retrieval performance. Our system incorporates domain-specific knowledge and linguistic features unique to the Vietnamese language and legal texts, ensuring high relevance and accuracy in search results. We provide a comprehensive evaluation of our system, demonstrating its superiority over existing single-method retrieval approaches.

Index Terms—Fusion-based Retrieval, Legal Text Processing, Legal Document Retrieval

I. INTRODUCTION

In the rapidly evolving digital age, the ability to efficiently retrieve relevant information from vast repositories of textual data is critical [1], [2]. This is particularly true in the legal domain, where accurate and timely access to legal documents can significantly impact judicial outcomes and legal research [3]. However, the task of document retrieval becomes more challenging in low-resource languages [4], [5], such as Vietnamese, where the availability of comprehensive linguistic resources and advanced retrieval systems is limited [6], [7].

Vietnamese legal texts present unique challenges for document retrieval due to their complex structure, formal language, and nuanced interpretation of legal terminologies [8]. On the one hand, traditional retrieval systems often fall short in addressing these intricacies, leading to suboptimal search results that hinder legal professionals and researchers. On the other hand, fusion-based document retrieval [9] is an advanced technique that combines multiple retrieval methods' strengths to enhance search results' accuracy and relevance. By integrating various models and data sources, this approach leverages their complementary strengths, thereby overcoming the limitations inherent in individual systems. This paper explores the application of fusion-based retrieval to the Vietnamese

legal corpus, demonstrating how this method can significantly improve retrieval performance in a low-resource setting.

Our approach involves the use of diverse retrieval models, including traditional lexical matching techniques and modern neural network-based methods, which are fused to create a robust retrieval framework. Additionally, we incorporate domain-specific knowledge and linguistic features unique to the Vietnamese language and legal texts, ensuring that the system is finely tuned to the specific requirements of legal document retrieval.

The main contributions of this paper are threefold:

- 1) We develop a comprehensive fusion-based retrieval system designed specifically for Vietnamese legal texts.
- 2) We provide a detailed evaluation of the system's performance, highlighting its advantages over existing single-method retrieval approaches.
- 3) We offer an efficient practical process to adapt better knowledge of the legal domain.

II. RELATED WORKS

The field of document retrieval has undergone significant evolution, with traditional and modern techniques both contributing to its advancement. In the context of legal texts, the challenges are magnified due to the complexity, specificity, and formal nature of legal language. This section reviews key related works in the areas of document retrieval, fusion-based retrieval techniques, and applications in low-resource languages, particularly focusing on Vietnamese legal texts.

A. Lexical-Based Retrieval

Classical document retrieval methods, such as TF-IDF [10] and BM25 [11], have been foundational in the field of information retrieval. TF-IDF relies on the statistical significance of words in documents relative to a corpus, while BM25 refines this approach by incorporating term saturation and document length normalization. These methods have been widely adopted for their simplicity and effectiveness in various domains, including legal texts [6], [8]. However, their reliance on exact term matching often limits their ability to handle the nuanced and formal language of legal documents, particularly in low-resource settings.

*Corresponding author: tvkhanh@ictu.edu.vn

B. Semantic-Based Retrieval

Recent advancements in natural language processing have led to the development of semantic-based retrieval methods, such as BERT [12], RoBERTa [13], which leverage deep learning to understand semantic content and context. BERT-based models have significantly improved retrieval tasks by capturing the meaning of queries and documents beyond mere keyword matching. More advanced, Sentence Transformer (a.k.a. Sentence-BERT or SBERT) [14] is a siamese architecture for producing high-quality sentence embeddings. The resulting pre-trained models provide meaningful dense vector representations for sentences according to the specific task they were fine-tuned on, such as semantic similarity, clustering, and information retrieval. In the legal domain, these models have been adapted to understand legal jargon and context [15], but their application to Vietnamese legal texts [16]–[18] remains under-explored due to the scarcity of annotated data.

C. Fusion-Based Retrieval

Fusion-based retrieval techniques aim to combine the strengths of multiple retrieval models to improve performance. Rank-based fusion methods [19] such as Reciprocal Rank Fusion (RRF) [20] aggregate the ranks or scores from multiple, previously ranked results to produce a unified result set, enhancing robustness and retrieval quality. Score-based fusion methods [21] merge the output scores from various models to re-rank documents more effectively. More recently, hybrid approaches have used the Relative Score Fusion algorithm to optimize the combination of different retrieval strategies. These techniques have shown promise in various domains, including academic and web search. Still, their potential in legal document retrieval, especially in the context of Vietnamese texts, has not been fully explored.

III. METHODS

A. Legal Document Retrieval Task

Given a set of queries $Q = \{q_1, q_2, \dots, q_n\}$ and a corpus of legal documents $D = \{d_1, d_2, \dots, d_k\}$, the goal of Task 1 [6] is to retrieve a subset $D'_i \subset D$ of legal articles for each query $q_i \in Q$ based on their relevance to that query. The ranking function $R(d, q)$ assigns a relevance score to each document d according to its similarity to the query q . Therefore, the document retrieval problem for each $q_i \in Q$ can be formulated as:

$$D'_i = \arg \max_{d \in D} R(d, q_i)$$

B. Fusion-based Retrieval System

The Vietnamese Legal Text Retrieval system employs a Retriever module, wherein a legal question is transformed into an embedding vector. This vector is then compared to embeddings within the legal corpus using similarity scores to identify pertinent legal documents. As illustrated in Figure 1, the Retriever (dashed flow) is iteratively trained in two rounds [8]: initially, we obtain negative pairs from BM25 module to

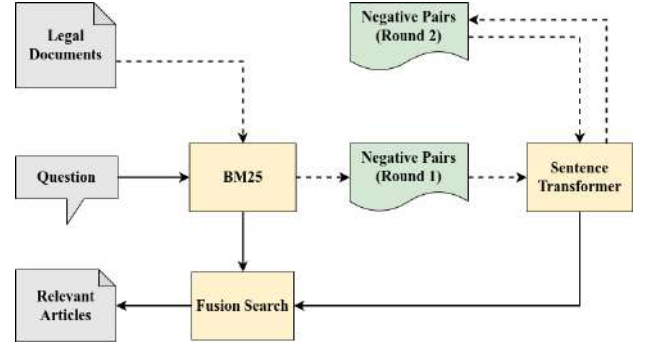


Figure 1: Overview of the Fusion-based Retrieval System

train a Sentence Transformer (a.k.a. SBERT) with Contrastive Loss; and subsequently, this model, in the second round, predicts false negative sentence pairs for further refinement with Contrastive Loss.

Following the multi-stage training flow [8], we conducted experiments with the Sentence Transformer module based on several pre-trained backbone language models, including PhoBERT [22], SB-Condenser [23], VietnamLegalText-SBERT¹, and Vietnamese-Bi-Encoder² [24]. The SBERT models were trained for 10 epochs in the first round and 5 epochs in the second round, with 20 negative samples generated from BM25 and the SBERT model (see Section III-C for more details). The maximum sequence length and the learning rate were set at 256 and 1×10^{-5} , respectively, while other parameters such as α and Δ are fixed during the search. In this study, we offer an efficient process to adapt better knowledge of the legal domain, in which we continually fine-tuned the Vietnamese-Bi-Encoder (**ViBiEn**) model on the legal data using Masked Language Modeling (**-MLM**) and Condenser (**-CON**) [25] tasks (see results in Table II).

C. Sentence Transformer

In this study, we leverage the SBERT model to determine the similarity between questions and articles, employing contrastive learning with positive pairs from the dataset and negative pairs generated by BM25 or SBERT. Table IV shows alternative methods for generating negative pairings. For comparing phrases, we utilize a cross-encoder structure that merges two sentences within a single BERT or Transformer encoder model. Additionally, SBERT applies a pooling operation to the BERT’s outputs to produce fixed-size sentence embeddings.

Given a set of questions $Q = \{q_1, q_2, \dots, q_n\}$ and a set of articles $D = \{d_1, d_2, \dots, d_k\}$, we compute the similarity between pairs of questions and articles using the Cosine similarity as follows:

$$f(q, d) = \frac{E_q E_d}{\|E_q\| \cdot \|E_d\|} \quad (1)$$

where E_q , E_d are embedding vectors of the question q and the article d using the SBERT model, respectively.

¹hmthanh/VietnamLegalText-SBERT

²<https://huggingface.co/bkai-foundation-models/vietnamese-bi-encoder>

To train a contrastive SBERT model, $f(u, v)$ is utilized to calculate the loss with both positive article d_{pos} and negative article d_{neg} for the given question. The loss function is as follows:

$$\mathcal{L} = \sum_q \sum_{d_{pos} \in D_{pos}} l(q, d_{pos}, D_{neg}) \quad (2)$$

$$l(q, d_{pos}, D_{neg}) = -\log \frac{e^{f(q, d_{pos})}}{e^{f(q, d_{pos})} + \sum_{d_{neg} \in D_{neg}} e^{f(q, d_{neg})}} \quad (3)$$

where D_{pos} is the positive document collection for the given question q , and D_{neg} is the collection of negative documents for the given question q selected via sparse retrieval methods.

D. Fusion Search

We aim to combine the strengths of both traditional lexical-based search (i.e., BM25) and semantic-based search (i.e., Sentence-BERT) to deliver highly relevant search results. This section presents a detailed explanation of how fusion search works and its components, which utilize advanced algorithms to blend keyword relevancy scores with vector similarity scores. The results are then ranked based on a combination of keyword match relevance and semantic similarity.

Simple Fusion: We determine the final score of a pair of questions and articles using the following equation as **SimFS1**:

$$score = score_{BM25} * score_{SBERT} \quad (4)$$

and as **SimFS2**:

$$score = \sqrt{score_{BM25}} * score_{SBERT} \quad (5)$$

Ranked Fusion: $score_{BM25}$ and $score_{SBERT}$ are generated based on the rank of each result, as follows: $1/(RANK + 60)$ [20]. The final score is calculated by using the following, as **RankedFS**:

$$score = \alpha * score_{BM25} + (1 - \alpha) * score_{SBERT} \quad (6)$$

Relative Fusion: Lexical- and semantic-based scores are normalized before computing the final scores with the equation (6). In **RelFS-Mima**, we normalize the scores using the min-max scaling method as below:

$$score_o = \frac{f_o(q, d) - m_q}{M_q - m_q} \quad (7)$$

where $M_q = \max_{d \in D} f_o(q, d)$ and $m_q = \min_{d \in D} f_o(q, d)$. Alternatively, in **RelFS-Theo**, we redefine normalization by substituting the ensemble minimum with the following function's theoretical minimum:

$$score_o = \frac{f_o(q, d) - \inf f_o(q, \cdot)}{M_q - \inf f_o(q, \cdot)} \quad (8)$$

For example, score BM25's infimum is 0 while the SBERT score is cosine similarity, then its score is -1.

A further popular choice is the standard score (z-score) normalization, named as **RelFS-Zscore**, defined as follows:

$$score_o = \frac{f_o(q, d) - \mu}{\sigma} \quad (9)$$

where μ and σ are the mean and standard deviation of the collection of scores $f_o(q, \cdot)$ for question q , respectively.

Finally, we select the top $k = 5$ related articles with the highest final scores which also fall between the range of $maxscore - \Delta$ and $maxscore$, where Δ is a hyperparameter.

IV. EXPERIMENTAL SETTING

We used data from the "Legal Text Retrieval" dataset of the "ZALO AI Challenge 2021"³. The Zalo corpus consists of 3,271 legal documents totaling 61,425 articles. The ALQAC 2023⁴ training dataset contains fewer questions and articles compared to the ALQAC 2022 training dataset, while the number of laws is extended from 4 in 2022 to 17 in 2023. Table I shows the data statistics of our datasets.

Table I: Data Statistics

	ALQAC 2023	ALQAC 2022	Zalo
No. of Laws	17	4	3,271
No. of articles	2,249	3,271	61,425
No. of questions	100	520	3,196

The data undergoes preprocessing before being input into the lexical matching algorithm, with procedures such as word segmentation using Underthesea⁵ and converting text to uppercase. For evaluating the models on the legal text retrieval task, we use macro precision, recall, F1, and F2 scores on the test dataset. The macro F2 score is calculated as follows:

$$macroF2 = \text{avg} \frac{5 * Precision_i * Recall_i}{4 * Precision_i + Recall_i} \quad (10)$$

V. RESULTS

A. Backbone Models

Table II shows experimental results of backbone models with a simple fusion method (Equation 4) on both the Zalo public and private test sets. Most four backbone models follow the trend of multi-round training since models trained on the second round show better performance compared with those trained only on the first round. The SB-Condenser model performed best on the public test, with an F2 score of 80.9%, and 71.71% F2 on the private test. However, with an efficient process, the model trained with MLM task on legal dataset showed better gain on the in-domain retrieval task, in which the ViBiEn-MLM significantly outperformed the ViBiEn on the public test at 2.0 points, from 77.54% to 79.52%, while it also showed state-of-the-art results on the private test set, at 74.97% F2, 3.0 points higher than SB-Condenser's F2 score.

B. Fusion Search

In this section, we present experimental findings using fusion search methods (as described in Section III-D) on our previous SOTA model, ViBiEn-MLM. Table III shows the results of various fusion search algorithms on the private Zalo test set. Overall, RelFS-Mima stands out for balanced

³<https://challenge.zalo.ai/>

⁴<https://sites.google.com/view/ALQAC2023>

⁵<https://github.com/undertheseanlp/underthesea>

Backbone Models	Round	Public Test				Private Test			
		P	R	F1	F2	P	R	F1	F2
Phobert	1	69.21	79.99	71.84	75.29	60.88	74.06	64.07	68.28
	2	70.27	80.18	72.92	76.15	61.84	74.36	64.93	68.95
SB-Condenser	1	73.50	82.00	75.75	78.53	64.05	75.58	66.83	70.56
	2	75.58	84.84	77.9	80.9	65.20	76.52	68.02	71.71
VietnamLegalText-SBERT	1	70.87	79.01	72.87	75.55	61.15	73.32	64.06	67.97
	2	71.15	80.44	73.49	76.48	60.31	72.52	63.30	67.23
ViBiEn	1	72.99	80.97	74.99	77.54	62.41	72.70	64.90	68.23
	2	73.05	81.94	75.17	78.03	62.91	72.86	65.38	68.62
ViBiEn-MLM	1	74.42	83.27	76.61	79.52	68.27	80.09	71.14	74.97
	2	74.60	83.27	76.76	79.61	67.59	78.76	70.32	73.96
ViBiEn-MLM-CON	1	74.20	82.79	76.31	79.17	67.46	78.08	70.08	73.54
	2	73.77	82.00	75.68	78.37	66.86	77.34	69.35	72.72

Table II: Results of backbone models on the Zalo challenge. **Bold** and ***Bold*** indicate the **best** and ***second best*** results, respectively.

Method	Precision	Recall	F1	F2
ViBiEn-MLM	68.27	80.09	71.14	74.97
SimFS1	68.14	80.73	71.18	75.24
SimFS2	64.05	78.45	67.55	72.16
RankedFS	63.06	74.32	66.3	70.10
RelFS-Mima	67.68	83.42	71.68	76.74
RelFS-Theo	67.88	82.82	71.68	76.51
RelFS-Zscore	68.64	79.60	71.43	75.03

Table III: Results on searching the best weight with Fusion algorithms on the ViBiEn-MLM model

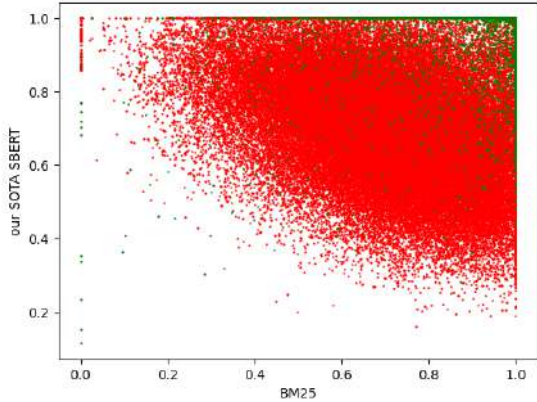


Figure 2: Normalizing the metrics outputs of BM25 and SBERT scores using the min-max scale.

performance across all metrics. ViBiEn-MLM and SimFS1 also provide strong performance but are outperformed by RelFS-Mima in balanced scenarios. If a single metric like Precision or Recall is prioritized, specific methods like RelFS-Zscore or RelFS-Theo might be more suitable. The relative fusion approach using the min-max scale set a new state-of-the-art F2 score at 76.74%.

Figure 2 compares normalized BM25 scores against normalized scores from the SBERT model. The red points represent negative pairs derived from the questions of the private test set using both BM25 and SBERT techniques, while the green points represent positive pairs selected from the private test set. Many points have low BM25 scores, but their SBERT scores

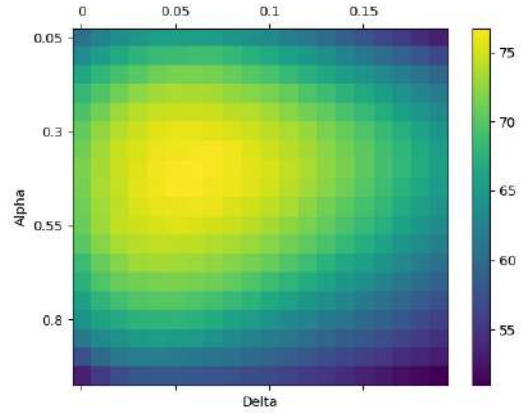


Figure 3: The effect of parameters α and Δ on the F2 measure on the private test set of the RelFS-Mima model

vary widely, indicating possible divergence in the relevance assessments between BM25 and SBERT for those cases, while clustering of points around high values, in the upper right corner, for both BM25 and SBERT, which suggests agreement between the two methods on the relevance of these documents or queries.

Figure 3 illustrates how altering the values of parameters α and Δ affects the F2 score on the private test set. The central yellow region indicates the highest F2 scores. This suggests that there is an optimal combination of α and Δ within this region that maximizes the F2 score. The heatmap shows that optimal values for α and Δ lie in the mid-range, particularly around $\alpha \approx 0.3$ to 0.5 and $\Delta \approx 0.05$ to 0.10 , where the F2 score is maximized.

Table VI shows the retrieval results of some models. RelFS-Mima consistently outperforms other models in ranking relevant legal articles, while SBERT and BM25 show inconsistent performance, indicating potential limitations in handling diverse question types or relevance criteria. An interesting finding here is that the "VPHC", an abbreviation for "vi phạm hành chính" (administrative violations), seems to be handled with our proposed fusion-based retrieval models, RelFS-Mima

Gen Method	P	Public test			P	Private test		
		R	F1	F2		R	F1	F2
Random	61.14	80.73	66.15	72.33	54.46	74.56	59.37	65.73
BM25	74.42	83.27	76.61	79.52	68.27	80.09	71.14	74.97
TF-IDF	73.07	81.71	75.06	77.88	64.06	73.93	66.47	69.68
W2V	62.55	81.25	67.47	73.41	54.64	74.77	59.53	65.89
ViBiEn	76.12	83.47	77.79	80.23	64.82	75.23	67.34	70.72
ViBiEn-MLM	73.73	83.57	76.12	79.32	65.86	76.55	68.49	71.97

Table IV: Results using negative pair generation methods with k=20

Model	P	R	F2
NeCo	90.00	86.21	86.61
NOWJ1	86.36	83.49	83.58
AI_EPU	70.91	68.18	68.49
ST	24.85	72.27	52.07
ViBiEn-MLM (Run 1)	79.62	86.82	83.91
ViBiEn-MLM (Run 2)	83.94	85.45	84.29

Table V: Results on the private test of ALQAC 2023 task 1

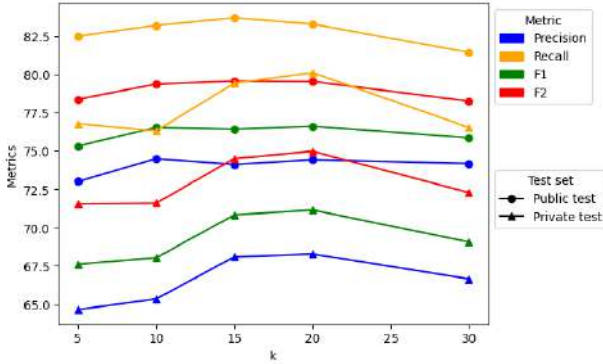


Figure 4: The effect of the number of negative pairs k.

and SimFS1, while individual lexical- and semantic-based models, BM25 and SBERT, respectively, failed to get the top ranks of the relevant article.

C. ALQAC 2023: Retrieval Task 1

Table V shows the results of previous models and our models on the private test set. ViBiEn-MLM run 2 (trained on combined ALQAC 2023 and Zalo sets) improves in precision and F2 score compared to Run 1 (trained on ALQAC 2023 only). ViBiEn-MLM models, particularly Run 2, show competitive performance, achieving the second-best F2 score and demonstrating the benefits of enhanced training methods. NeCo [7], with data enrichment for the pre-training phase of BERT-based models, is the top-performing model overall, with the highest precision and F2 score, indicating its superior balance of precision and recall for the ALQAC 2023 task 1.

D. Hyperparameters

Table IV presents the results of training the ViBiEn backbone model with the MLM task using various negative pair-generating strategies. The results indicate that the lexical-based approach BM25 is best suited for creating negative pairings. Figure 4 depicts the effect of the number k of negative pairs on model outcomes. Increasing k up to 20 generally improves

Question	Thẩm quyền xử phạt VPHC của Đội trưởng Đội nghiệp vụ Cảnh sát biển trong lĩnh vực y tế? (What is the jurisdiction of the Chief of the Professional Task Force of the Coast Guard to impose penalties for administrative violations in the field of health care?)			
Answer	Article 109 in law id:117/2020/nd-cp			
Rank	BM25	SBERT	SimFS1	RelFS-Mima
	16	8	3	1
Question	Việc xác định doanh thu của tổ chức tài chính vi mô khi tính thuế thu nhập doanh nghiệp được thực hiện như thế nào? (How is the revenue of a microfinance organization determined for corporate income tax purposes?)			
Answer	Article 6 in law id:18/2018/tt-btc			
Rank	BM25	SBERT	SimFS1	RelFS-Mima
	2	29	4	1

Table VI: A comparison of predicted outputs from a variety of models.

all metrics, indicating that a moderate number of negative pairs is beneficial, while beyond $k = 20$, all metrics slightly degrade or plateau, suggesting diminishing returns or potential overfitting. Figure 5 illustrates how the parameter Δ affects fusion search algorithms. Δ has value ranges for each method, as they produce different final scores.

VI. CONCLUSION

In this paper, we presented a fusion-based document retrieval system tailored for Vietnamese legal texts, addressing the challenges inherent in low-resource language settings. Our approach integrates multiple retrieval models, leveraging both traditional lexical techniques and advanced neural network-based methods to improve the relevance and accuracy of search results. By combining these methods, we harness their complementary strengths, enabling more effective retrieval of legal documents that align with the specific queries posed. Experimental results demonstrated that the fusion-based approach significantly outperforms single-method retrieval systems. We also showed that incorporating in-domain knowledge enhances the retrieval process, leading to a more comprehensive understanding and accurate retrieval of legal documents.

Future work will focus on expanding the system's capabilities, incorporating domain-specific knowledge and linguistic features unique to Vietnamese to enhance the retrieval process, leading to a more comprehensive understanding and accurate retrieval of legal documents. Additionally, we aim to explore the application of our approach to other low-resource languages and domains, contributing to the broader field of information retrieval.

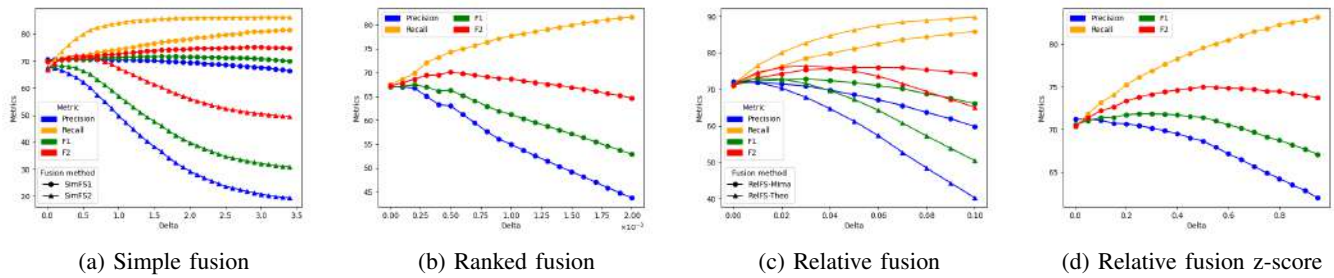


Figure 5: How the parameter Δ affects our fusion searches on the private test set.

REFERENCES

- [1] D. Sargiotis, *Data Governance in the Digital Age: Strategies, Challenges, and Best Practices*, 01 2024.
- [2] K. A. Hambarde and H. Proença, "Information retrieval: Recent advances and beyond," *IEEE Access*, vol. 11, pp. 76 581–76 604, 2023.
- [3] A. Abdallah, B. Piryani, and A. Jatowt, "Exploring the state of the art in legal qa systems," *Journal of Big Data*, vol. 10, 08 2023.
- [4] P. Krasadakis, E. Sakkopoulos, and V. S. Verykios, "A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages," *Electronics*, vol. 13, no. 3, 2024. [Online]. Available: <https://www.mdpi.com/2079-9292/13/3/648>
- [5] M. Khan, K. Ullah, Y. Alharbi, A. Alferaidi, T. Alharbi, K. Yadav, N. Alsharabi, and A. Ahmad, "Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive," *Applied Sciences*, vol. 13, p. 8566, 07 2023.
- [6] C. Nguyen, S. Luu, T. Tran, A. Trieu, A. Dang, D. Nguyen, H. Nguyen, T. Pham, T. Pham, T.-T. Vo, D.-T. Dol, N. Khang, H. Nguyen, N.-C. Le, T.-T. Le, Q. Bui, P. Nguyen, H.-T. Nguyen, V. Tran, and L. Nguyen, "A summary of the alqac 2023 competition," 10 2023, pp. 1–6.
- [7] H.-L. Nguyen, D.-Q. Nguyen, H.-T. Nguyen, T.-T. Pham, H.-D. Nguyen, T.-A. Nguyen, T.-H.-Y. Vuong, and H.-T. Nguyen, "Neco@alqac 2023: Legal domain knowledge acquisition for low-resource languages through data enrichment," 2023.
- [8] N.-M. Pham, H.-T. Nguyen, and T.-H. Do, "Multi-stage information retrieval for vietnamese legal texts," 09 2022.
- [9] O. Kurland and J. S. Culpepper, "Fusion in information retrieval: Sigir 2018 half-day tutorial," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1383–1386. [Online]. Available: <https://doi.org/10.1145/3209978.3210186>
- [10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing Management*, vol. 24, no. 5, pp. 513–523, 1988. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0306457388900210>
- [11] S. Robertson, H. Zaragoza, and M. Taylor, "Simple bm25 extension to multiple weighted fields," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, ser. CIKM '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 42–49. [Online]. Available: <https://doi.org/10.1145/1031171.1031181>
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:52967399>
- [13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:198953378>
- [14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: <https://arxiv.org/abs/1908.10084>
- [15] M. Yoshioka, Y. Aoki, and Y. Suzuki, "Bert-based ensemble methods with data augmentation for legal textual entailment in collee statute law task," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ser. ICAIL '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 278–284. [Online]. Available: <https://doi.org/10.1145/3462757.3466105>
- [16] C.-N. Chau, T.-S. Nguyen, and L.-M. Nguyen, "Vnlawbert: A vietnamese legal answer selection approach using bert language model," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, 2020, pp. 298–301.
- [17] T.-T. Tieu, C.-N. Chau, N.-M.-H. Bui, T.-S. Nguyen, and L.-M. Nguyen, "Apply bert-based models and domain knowledge for automated legal question answering tasks at alqac 2021," in *2021 13th International Conference on KSE*, 2021, pp. 1–6.
- [18] H.-L. Nguyen, T.-B. Nguyen, T.-M. Nguyen, H.-T. Nguyen, and H.-Y. T. Vuong, "Vlh team at alqac 2022: Retrieving legal document and extracting answer with bert-based model," in *2022 14th International Conference KSE*, 2022, pp. 1–6.
- [19] R. Benham and J. S. Culpepper, "Risk-reward trade-offs in rank fusion," in *Proceedings of the 22nd Australasian Document Computing Symposium*, ser. ADCS '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: <https://doi.org/10.1145/3166072.3166084>
- [20] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 758–759. [Online]. Available: <https://doi.org/10.1145/1571941.1572114>
- [21] K. M. Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *Image and Video Retrieval*, W.-K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 61–70.
- [22] D. Q. Nguyen and A. Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.92>
- [23] N. Q. T. T. H. V. Nguyen Hoang Gia Khang, Nguyen Minh Nhat, "Vietnamese legal text retrieval." [Online]. Available: <http://ds.libol.fpt.edu.vn/handle/123456789/3716>
- [24] N. Q. Duc, L. H. Son, N. D. Nhan, N. D. N. Minh, L. T. Huong, and D. V. Sang, "Towards comprehensive vietnamese retrieval-augmented generation and large language models," 2024.
- [25] L. Gao and J. Callan, "Condenser: a pre-training architecture for dense retrieval," in *Conference on Empirical Methods in Natural Language Processing*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:237581068>

SHREADS: A Shrimp Health Diagnosis and Advisory System using Agentic RAG

Thao Ngo
CICT, Can Tho University
Can Tho, Vietnam
thaob2207566@student.ctu.edu.vn

Phu-An Thai
CICT, Can Tho University
Can Tho, Vietnam
anb2207512@student.ctu.edu.vn

Tan Vo
CICT, Can Tho University
Can Tho, Vietnam
tanb2207561@student.ctu.edu.vn

The-Khanh Chau
CICT, Can Tho University
Can Tho, Vietnam
khanhb2207528@student.ctu.edu.vn

Thanh Ma
CICT, Can Tho University
Can Tho, Vietnam
mtthanh@ctu.edu.vn

Minh-Thai Truong
CICT, Can Tho University
Can Tho, Vietnam
tmthai@ctu.edu.vn

Abstract—Shrimp aquaculture, a key driver of Vietnam’s economy, faces persistent disease threats that cause severe production and financial losses. Traditional diagnostics often fail to enable timely responses. We propose SHREADS (*SHR*imp *hEALTH* Diagnosis and advisory System), an intelligent platform providing rapid, accurate disease detection and management. SHREADS conducts an Agentic Retrieval-Augmented Generation (A-RAG) framework with a multi-agent architecture for optimized query processing. Farmers submit text or image queries through a mobile application, which are processed through specialized models, i.e., a ResNet50-based CNN for image-based diagnosis and a FastText model for textual symptom analysis. Experiments demonstrate high performance, with ResNet50 achieving an F1-score of 92.85% for classifying healthy shrimp and FastText reaching 82.4%. By integrating Agentic RAG and multimodal deep learning, SHREADS enhances early intervention, supports sustainable aquaculture, and promotes technological adoption in Vietnam. Our dataset and implementation are published in Github: <https://github.com/p4wndev/shreads>.

Keywords—Shrimp Disease Detection, Agentic RAG, Advisory Systems, Deep Learning, Sustainable Aquaculture.

I. INTRODUCTION

Shrimp aquaculture represents a critical pillar of Vietnam’s rural economy, yet its long-term viability is persistently threatened by recurrent disease outbreaks that devastate production and impose severe economic burdens on farming communities. Existing disease management practices are often grounded in anecdotal experience or fragmented local knowledge, offering limited predictive value and lacking the scientific rigor necessary for swift, effective response. For instance, farmers frequently rely on visual assessments or informal peer advice, which can delay intervention and exacerbate stock mortality. This persistent vulnerability underscores a pressing need for accessible, data-driven diagnostic solutions that leverage advanced computational methods (i.e., Agent AI) to enable early detection, evidence-based decision-making, and sustainable aquaculture practices.

Although prior research has examined technology-driven solutions—such as machine learning-based predictive models by Do Thi Thanh et al. [1], Xun Ran et al. [2], and Yadav

et al. [3], alongside global efforts integrating IoT and AI for real-time aquaculture monitoring [4]—substantial obstacles persist. High capital investment, system heterogeneity, and the digital divide among smallholder farmers hinder large-scale deployment. For example, while advanced IoT platforms offer real-time water quality sensing and AI-enabled disease forecasting, their cost and complexity often exceed the financial and technical capacity of rural operators, who constitute the majority of shrimp producers in Vietnam. These limitations underscore the urgent need for a cost-effective, user-friendly, and fully integrated system specifically designed for the socio-economic realities and operational constraints of Vietnamese aquaculture, enabling equitable access to technological innovations and sustainable farming practices.

Moreover, diagnosing shrimp diseases is notoriously complex due to a confluence of various factors: (1) clinical signs for diseases like White Spot Syndrome Virus (WSSV) and Early Mortality Syndrome (EMS) can be highly variable, changing with the shrimp species, life stage, and environmental conditions; (2) the dynamic pond ecosystem, with fluctuating water quality parameters such as pH, salinity, and dissolved oxygen, can either mask or mimic disease symptoms; (3) the possibility of pathogen co-infections often presents a blended and confusing set of symptoms, confounding accurate visual diagnosis; (4) the continuous evolution of pathogen strains necessitates adaptive diagnostic systems. These challenges underscore the need for advanced tools that can move beyond simple observation.

This paper introduces SHREADS, a novel system designed to overcome these challenges. By integrating a multi-agent architecture with advanced AI, SHREADS provides an accessible and robust platform for disease diagnosis and advisory services. Our primary contribution is a system that leverages multimodal inputs (*text and images*) within an Agentic RAG framework to deliver accurate, context-aware guidance, thereby reducing economic losses and promoting sustainable aquaculture practices in Vietnam.

II. BACKGROUND

This section outlines the theoretical foundations and key techniques underpinning SHREADS, a system designed to address complex, multimodal farmer queries in shrimp aquaculture. We introduce the core concepts and rationale for our approach, with implementation details provided in Section III.

A. Leveraging Agentic RAG for System-wide Coordination

To handle multifaceted farmer queries (e.g., combining image-based diagnosis with treatment inquiries), we adopt an Agentic Retrieval-Augmented Generation (RAG) framework [5]. Unlike monolithic RAG pipelines, which lack flexibility for composite requests, Agentic RAG decomposes queries into sub-tasks, orchestrated by a Supervisor Agent guided by a Semantic Router. This approach mimics an expert's diagnostic workflow, integrating retrieval with multi-step reasoning.

Definition 1 (Agentic RAG Principle in SHREADS): The Agentic RAG process in SHREADS handles a query q by identifying its intent and orchestrating sub-tasks (e.g., classification, retrieval) via a Supervisor Agent, leveraging conversation history H for context-aware responses. The final answer A is synthesized as:

$$A = G(q, H, C(q), R(q, K)),$$

where $C(q)$ is the output of a Classifier Agent, $R(q, K)$ retrieves relevant documents D_{rel} from a knowledge base K , and G is a Generator Agent (LLM).

B. Semantic Contextual Chunking for Shrimp Disease Documents/Reports

To optimize knowledge retrieval, SHREADS employs Semantic Contextual Chunking [6]. This technique segments documents based on semantic coherence, creating self-contained, meaningful chunks that preserve information integrity, unlike fixed-size chunking. This ensures accurate retrieval for the downstream RAG pipeline.

Definition 2 (Semantic Contextual Chunking Process): Given a document D , Semantic Contextual Chunking segments it into coherent chunks based on semantic shifts, enriches them with LLM-generated summaries, and indexes them for efficient retrieval (see Fig. 1 and Section III-B).

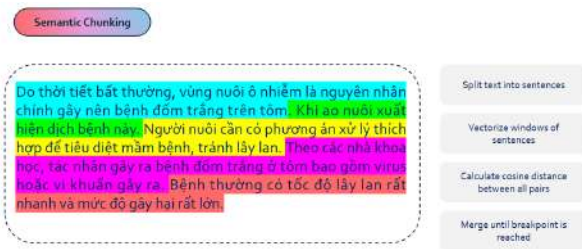


Fig. 1. Illustration of the Semantic Contextual Chunking process. Sentences are embedded, and chunk boundaries are created at points of high semantic distance, ensuring contextually coherent segments for indexing.

C. Hybrid Search Techniques for RAG

SHREADS implements a hybrid search mechanism, combining keyword-based (BM25) and semantic vector search (FAISS) to capture both lexical and conceptual relevance. Results are merged using Reciprocal Rank Fusion (RRF) to produce a highly relevant document set (see Section III-B for details).

Definition 3 (Hybrid Search): Hybrid search integrates BM25 for keyword matching and FAISS for semantic similarity, with RRF combining results for robust ranking:

$$\text{RRFScore}(d) = \sum_i \frac{1}{k + \text{rank}_i(d)},$$

where $\text{rank}_i(d)$ is the rank of document d in list i , and k (e.g., 60) mitigates outlier ranks.

D. Shrimp Disease Region Detection

To ensure diagnostic accuracy from noisy farm images, SHREADS uses a two-stage process to isolate disease regions, focusing analysis on relevant visual biomarkers (see Section III-A for implementation).

Definition 4 (Object Detection Role in SHREADS): Object detection isolates potentially diseased regions in an input image I , enabling focused classification of visual biomarkers.

III. METHODOLOGY

Building on the theoretical foundations in Section II, we developed SHREADS, an intelligent system engineered to deliver timely and accurate disease management support in shrimp aquaculture. The system's central innovation is its Agentic RAG architecture, where specialized AI agents collaborate to process complex, multimodal user queries. This modular design enhances scalability and enables multi-step reasoning for comprehensive analysis. The overall architecture, depicted in Figure 2, shows a central Supervisor Agent orchestrating the workflow between specialized agents.

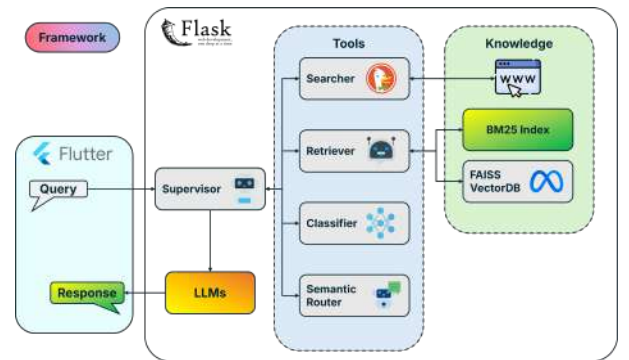


Fig. 2. The SHREADS system architecture. The Supervisor Agent orchestrates the workflow, routing queries to the Classifier Agent and Retriever Agent before generating a final response.

The workflow begins when a user submits a query q , which may consist of both text and an image. Following the Agentic RAG principle (Definition 1), the process unfolds as follows:

- 1) **Intent Recognition:** A lightweight, LLM-based Semantic Router performs initial intent classification, categorizing the query (e.g., DIAGNOSIS, INFORMATION, GREETING) to determine the appropriate downstream processing path.
- 2) **Agentic Orchestration:** The Supervisor Agent acts as the system’s central controller, dynamically constructing an execution plan based on the recognized intent and conversation history H , deciding whether to invoke the diagnostic module, the retrieval module, or both, and in what sequence.

A. Multimodal Diagnostic Module

For queries requiring a diagnosis, the Supervisor Agent dispatches the task to the Classifier Agent, which employs a multi-pathway strategy to analyze both visual and textual information comprehensively, as shown in Figure 3.

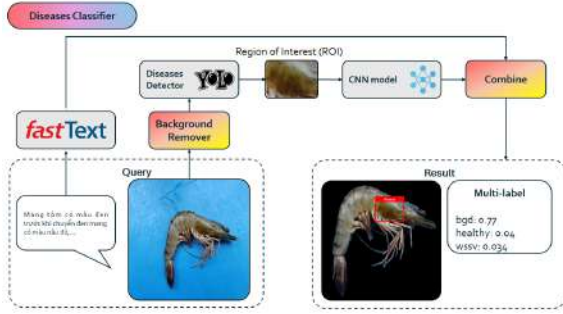


Fig. 3. Workflow of the Multimodal Diagnostic Module. An input image is processed by a YOLO and ResNet50 pipeline, while text is analyzed by FastText. The resulting predictions are fused to produce a final diagnosis.

a) *Image-based Diagnosis:* As outlined in Definition 4, the visual analysis pipeline starts with a **YOLO** model, selected for its state-of-the-art balance of speed and accuracy in object detection. This model localizes shrimp disease regions within a preprocessed image I' , yielding a set of bounding boxes $B = \{b_j\}$ with associated confidence scores $S = \{s_j\}$. A subset of boxes $B_{filt} = \{b_j \in B \mid s_j > \theta\}$ (e.g., $\theta = 0.7$) is cropped to form Regions of Interest (ROIs). These ROIs are fed to a **ResNet50-based CNN**, chosen for its ability to train deep convolutional networks through residual connections, mitigating vanishing gradient issues. Its deep stack of convolutional layers extracts complex textural features, color anomalies, and morphological changes—such as white spots and gill discoloration—indicative of shrimp diseases, producing probabilistic predictions $\mathcal{P}_{cnn}$.

b) *Text-based Diagnosis:* If the user provides a textual description of symptoms (e.g., “shrimp are lethargic and have white spots on their shells”), the text is analyzed by a **FastText** classification model, selected for its efficiency and robustness to misspellings and colloquialisms through character n-grams. This model produces a parallel set of predictions, $\mathcal{P}_{text}$.

c) *Probabilistic Fusion:* To generate a robust final diagnosis, the system fuses outputs from both modalities, mitigating errors from a single source (e.g., low-quality images or ambiguous descriptions). The final confidence score for each disease class l_k is computed by averaging probabilities:

$$\bar{p}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(l_i, p_i) \in \mathcal{S}_k} p_i,$$

where

$$\mathcal{S}_k = \{(l_i, p_i) \in \mathcal{P}_{cnn} \cup \mathcal{P}_{text} \mid l_i = l_k\}.$$

This ensemble approach leverages complementary evidence to improve diagnostic accuracy.

B. Hybrid Retrieval Module

For informational queries or post-diagnostic management protocols, the Supervisor Agent invokes the Retriever Agent, which queries a specialized knowledge base K using a hybrid retrieval strategy (Definition 3), as illustrated in Figure 4.

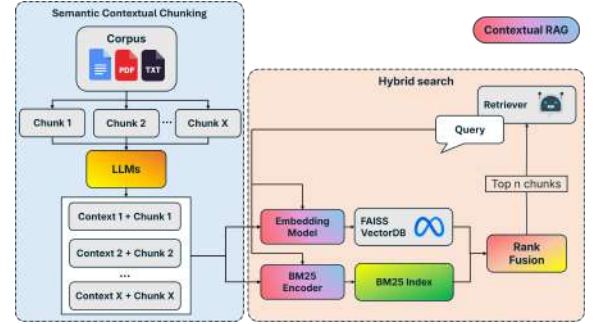


Fig. 4. The Hybrid Retrieval Module. Parallel lexical (BM25) and semantic (FAISS) searches are conducted, and results are merged via Reciprocal Rank Fusion (RRF) to retrieve the most relevant documents for the LLM.

a) *Knowledge Base Preparation:* The knowledge base K is preprocessed using Semantic Contextual Chunking (Definition 2). The document D is partitioned into sentences $\{s_1, s_2, \dots, s_m\}$, with embeddings $e_i = \text{Embed}(s_i)$. Chunk boundaries are inserted between sentences s_j and s_{j+1} if their semantic distance $d(e_j, e_{j+1}) > \tau$, forming coherent chunks c_k . Each chunk is enriched with an LLM-generated summary $ctx_k = \text{LLM-Summarize}(c_k)$, forming $c'_k = ctx_k \oplus c_k$, which is embedded and indexed:

$$\text{Index}(\text{Embed}(c'_k)).$$

This ensures self-contained, semantically meaningful units, reducing context fragmentation (see Fig. 1).

b) *Hybrid Search Execution:* The agent performs parallel searches to maximize recall and precision:

- **Lexical Search:** A sparse retriever using the BM25 algorithm identifies documents containing exact query keywords (e.g., “WSSV treatment”), scoring them as: $\text{Score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$, where $f(q_i, D)$ is the term frequency, $|D|$ is document

length, avgdl is average document length, and k_1, b are hyperparameters.

- **Semantic Search:** A dense retriever using a FAISS index maps chunks c'_k to embedding vectors $v_k = \text{Embed}(c'_k)$, indexed as $\mathcal{I} = \text{FAISSIndex}(\{v_k\})$. For a query Q embedded as $q_v = \text{Embed}(Q)$, FAISS retrieves the top- k nearest neighbors:

$$\text{NearestNeighbors}(q_v, \mathcal{I}, k) = \{(v_{j_1}, \dots, v_{j_k})\}.$$

Results are merged using RRF (Definition 3), selected for its robustness and simplicity, effectively combining both retrieval methods without complex score normalization.

IV. SHREADS ALGORITHM

The end-to-end SHREADS workflow is formalized in Algorithm 1. The algorithm accepts a multimodal query q and conversation history H . It first performs semantic routing (line 5) to determine the query type τ . Based on this, the Supervisor conditionally activates the diagnostic module (lines 9-17), the retrieval module (lines 19-23), or both. The diagnostic path involves object detection, multimodal classification, and prediction fusion. The retrieval path executes a hybrid search followed by rank fusion. Finally, all collated evidence, including classification results C_q and retrieved documents D_{rel} , is synthesized by an LLM to generate the final answer A (line 26).

A. Complexity Analysis

The computational complexity of the SHREADS workflow is governed by its most intensive components. Let N be the number of chunks in the knowledge base and d be the embedding dimension.

- **Routing & Text Classification (FastText):** Inference is linear in input length, $O(|q_{text}|)$, which is effectively constant time, $O(1)$, for typical queries.
- **Image Diagnosis (YOLO + ResNet50):** Inference time depends on image size but is constant for a fixed resolution, denoted T_{vision} .
- **Hybrid Retrieval:** BM25 search is approximately $O(|q_{text}| \log N)$. With an optimized index (e.g., IVFPQ), FAISS search is sub-linear, roughly $O(d \cdot n_{probes} + \log N)$, where $n_{probes} \ll N$. Thus, retrieval complexity is dominated by $O(\log N)$.
- **LLM Generation:** This is an auto-regressive process with complexity proportional to the generated answer length L_A , denoted $T_{LLM}(L_A)$.

The overall complexity for a query requiring both diagnosis and information retrieval is approximately $O(T_{vision} + \log N + T_{LLM}(L_A))$. This analysis confirms the system's scalability, as the most data-intensive operation (retrieval) scales logarithmically with the size of the knowledge base.

V. EXPERIMENTS AND RESULTS

This section presents the empirical evaluation of the SHREADS system. We detail the experimental setup, the datasets developed for training and validation, and the performance of the core classification models.

Algorithm 1 SHREADS End-to-End Workflow

```

1: Algorithm SHREADS_WORKFLOW( $q, H$ )
2:   Input: Query  $q = \{q_{text}, q_{image}\}$ , History  $H$ 
3:   Output: Generated answer  $A$ 
4:   // Phase 1: Semantic Routing
5:    $\tau \leftarrow \text{SEMANTICROUTER}(q_{text})$ 
6:   // Phase 2: Agentic Coordination
7:    $C_q \leftarrow \emptyset, D_{rel} \leftarrow \emptyset$ 
8:   if  $\tau = \text{DIAGNOSIS} \vee q_{image} \neq \emptyset$  then
9:      $P_{cnn} \leftarrow \emptyset, P_{text} \leftarrow \emptyset$ 
10:    if  $q_{image} \neq \emptyset$  then
11:       $R \leftarrow \text{YOLO}(q_{image})$ 
12:       $P_{cnn} \leftarrow \text{RESNET50}(R)$ 
13:    end if
14:    if  $q_{text} \neq \emptyset$  then
15:       $P_{text} \leftarrow \text{FASTTEXT}(q_{text})$ 
16:    end if
17:     $C_q \leftarrow \text{FUSEPREDICTIONS}(P_{cnn}, P_{text})$ 
18:  end if
19:  if  $\tau \in \{\text{DIAGNOSIS}, \text{INFORMATION}\}$  then
20:     $L_{bm25} \leftarrow \text{BM25SEARCH}(q_{text}, K)$ 
21:     $L_{faiss} \leftarrow \text{FAISSSEARCH}(q_{text}, K)$ 
22:     $D_{rel} \leftarrow \text{RRF}(L_{bm25}, L_{faiss})$ 
23:  end if
24:  // Phase 3: Response Generation
25:   $A \leftarrow \text{LLMGENERATE}(q, H, C_q, D_{rel})$ 
26:  return  $A$ 
27: end Algorithm

```

A. Experimental Environment

All experiments were conducted on a workstation equipped with a 12th Gen Intel Core i5-12450HX CPU, 12GB of RAM, and an NVIDIA GeForce RTX 3050 (6GB) GPU. The software stack included Python with TensorFlow (2.17.1), FastText (0.9.2), LangChain (0.3.13), and Flask (3.1) for the backend API. The client application was developed using Flutter for cross-platform deployment.

B. Dataset Curation and Preprocessing

To rigorously train and evaluate our models, we curated a multimodal dataset tailored to shrimp disease diagnosis. The dataset was constructed from three primary sources: (1) images and symptom descriptions voluntarily shared by farmers on aquaculture-focused social media platforms, focusing on Black Gill Disease and White Spot Syndrome; (2) expert-validated case data provided by aquaculture specialists; and (3) annotated data extracted from peer-reviewed publications by Can Tho University researchers [7].

Each collected image and textual description was independently annotated by three domain experts with backgrounds in aquaculture pathology. Disagreements were resolved through majority voting, and Cohen's kappa was computed to assess inter-annotator agreement, yielding an average $\kappa = 0.87$, indicative of strong consensus. Low-quality samples (e.g., blurry, poorly lit, or ambiguous) were filtered.

For the textual corpus, symptom descriptions were normalized by lowercasing, removing typographical noise, and harmonizing colloquial expressions. For the image dataset, background artifacts were cropped and brightness normalized. To mitigate sampling bias, we performed stratified sampling to ensure balanced representation across major disease categories. All datasets were partitioned into 70% training, 15% validation, and 15% testing splits, ensuring no farm-level data leakage across partitions.

The composition of the datasets is reported in Table I and Table II. Tables report the disease codes (labels), human-readable disease names, and the number of instances retained after preprocessing.

TABLE I
TEXT-BASED DATASET COMPOSITION AFTER EXPERT ANNOTATION.

Label	Disease	#Samples
wfs	White Feces Syndrome	142
lsd	Luminous Vibriosis Disease	121
wssv	White Spot Syndrome Virus	96
imn	Infectious Myonecrosis	85
bsd	Black Spot Disease	78
pd	Perkinsus Disease	73
ahpnd	Acute Hepatopancreatic Necrosis	71
fpd	Fouling Protozoan Disease	65
bgd	Black Gill Disease	43
Total		774

TABLE II
IMAGE-BASED DATASET COMPOSITION AFTER EXPERT ANNOTATION.

Label	Disease	#Samples
wssv	White Spot Syndrome Virus	339
bgd	Black Gill Disease	401
healthy	Healthy Shrimp	283
Total		1011

C. Model Performance

1) *Text-based Disease Classification*: We benchmarked multiple classical machine learning models (Naïve Bayes, k-Nearest Neighbors, Logistic Regression, Linear SVM, Random Forest) against FastText for text classification. All models were trained on the training split and hyperparameters selected via five-fold cross-validation on the validation set. Performance metrics are reported on the held-out test set. Accuracy (Acc.), Precision (Pre.), Recall (Rec.), F1-score (F1), as well as training and inference time are reported to provide a comprehensive comparison of model capacity and efficiency. Table III summarizes results. FastText outperformed all base-lines with an F1-score of 82.40%, demonstrating robustness to noisy farmer-generated text via its subword embedding mechanism.

TABLE III
PERFORMANCE COMPARISON OF TEXT CLASSIFIERS ON THE TEST SET.

Model	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)	Train Time (s)	Test Time (ms)
NB	79.49	80.11	79.49	79.05	0.02	9.13
KNN	61.54	67.28	61.54	57.36	0.02	43.43
LG	79.49	80.92	79.49	79.67	1.17	5.18
LSVC	74.36	74.25	74.36	73.37	0.06	5.04
RF	73.08	74.20	73.08	72.68	0.75	110.77
FastText	82.25	83.09	82.58	82.40	2.21	20.30

2) *Image-based Disease Classification*: We conducted two stages of image analysis: object detection and subsequent image classification.

a) *Object Detection*.: Six detectors, including recent YOLO variants and RF-DETR-base, were compared. All models were trained on the training split with standardized augmentations (random flipping, scaling). The evaluation metrics include mAP₅₀, mAP₇₅, mAP_{50–95}, precision, recall, and per-image inference latency measured on GPU. As shown in Table IV, YOLO11n provided an attractive balance between accuracy and low latency, making it suitable for real-time deployment. While RF-DETR-base reported stronger mAP₇₅ and recall, its significantly higher latency (46.7 ms) precluded its use for interactive diagnosis.

TABLE IV
OBJECT DETECTION RESULTS (TEST SET).

Model	mAP ₅₀ (%)	mAP ₇₅ (%)	mAP _{50–95} (%)	Pre. (%)	Rec. (%)	Latency (ms)
YOLO8n	87.21	54.41	51.37	92.45	77.47	8.8
YOLO9t	85.67	45.78	51.22	89.08	80.80	8.2
YOLO10n	86.41	53.43	51.65	96.46	73.76	7.1
YOLO11n	87.04	47.34	47.56	93.05	78.37	6.3
YOLO12n	81.49	55.90	47.87	85.96	71.70	9.2
RF-DETR-base	87.00	62.00	54.00	91.67	86.84	46.7

b) *Image Classification*.: We evaluated deep convolutional architectures (NASNetMobile, MobileNet, VGG16, VGG19, ResNet50, EfficientNetB0, InceptionV3, ConvNeXt-Tiny) under identical training protocols. Models were fine-tuned with Adam optimizer (1×10^{-4} learning rate), Batch Normalization, and Early Stopping with patience of 10 epochs. Accuracy, Precision, Recall, and F1-score were computed on the test set. Inference time (per image) was also measured. Results in Table V indicate that ResNet50 achieved the best tradeoff, attaining an F1-score of 92.85% with competitive inference time compared to lighter models.

TABLE V
IMAGE CLASSIFICATION PERFORMANCE (TEST SET).

Model	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)	Inf. Time (s)
NASNetMobile	80.39	80.24	80.39	80.28	0.207
MobileNet	89.22	89.25	89.22	88.98	0.027
VGG19	91.18	91.18	91.18	91.18	0.082
VGG16	91.18	91.23	91.18	91.20	0.062
ResNet50	92.93	92.97	92.93	92.85	0.095
EfficientNetB0	71.72	58.25	71.72	64.22	0.128
InceptionV3	76.77	64.08	76.77	69.12	0.119
ConvNeXtTiny	83.84	84.78	83.84	83.42	0.107

VI. APPLICATION

The practical implementation of SHREADS is delivered through an intuitive mobile application, developed using Flutter to ensure a seamless cross-platform user experience. This application serves as the primary interface for shrimp farmers, abstracting the underlying complexity of the Agentic RAG

architecture and presenting a user-friendly diagnostic tool. The workflow is designed for simplicity and efficiency: a farmer can initiate a query by either typing a description of observed symptoms in their native language (e.g., "Shrimp stop eating and have white spots on their shells.") or by uploading an image of an affected shrimp, or both.

Upon submission, the query is processed by the SHREADS backend, which leverages the full multi-modal diagnostic and hybrid retrieval pipeline described in Section III. The system provides a response directly within the app, as illustrated in Figure 5. The result includes a primary diagnosis with a confidence score (e.g., "White Spot Syndrome: 91.2%"), derived from the fused predictions of the image and text classifiers. Crucially, the application goes beyond simple identification by presenting actionable, context-aware advice retrieved from the knowledge base. This includes detailed descriptions of the disease, recommended treatment protocols, and preventative measures, thereby empowering farmers with the knowledge to make timely and informed interventions to protect their stock.

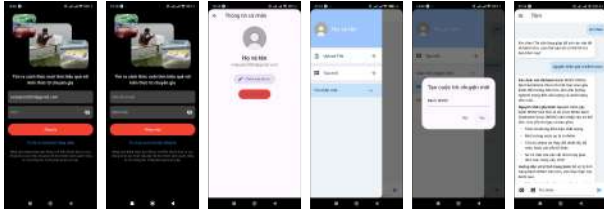


Fig. 5. The user interface of the SHREADS mobile application.

VII. CONCLUSION

This paper introduced SHREADS, a novel intelligent advisory system designed to address the significant challenge of disease management in Vietnam's shrimp aquaculture sector. By architecting the system around a collaborative team of specialized AI agents, our Agentic RAG framework successfully deconstructs and manages complex, multi-modal user queries. Our empirical evaluation demonstrates the efficacy of this approach. The multi-modal diagnostic module, which fuses a ResNet50-based computer vision model and a FastText classifier, achieves high accuracy in disease identification with F1-scores of 91.2% for images and 82.4% for text, respectively. Furthermore, our hybrid retrieval Agent, combining BM25 and FAISS with Reciprocal Rank Fusion, ensures that farmers receive relevant, context-rich information to guide their actions. The primary contribution of SHREADS is its potential to democratize access to veterinary expertise, providing small-scale farmers with a cost-effective and scalable tool for early disease detection and management.

Future work will focus on three key areas. First, we will expand the dataset to include a wider range of diseases and environmental conditions to further improve model robustness. Second, we plan to integrate real-time environmental data from IoT sensors, such as water quality parameters, to enhance the system's predictive capabilities. Finally, we will explore the use of more advanced generative models and multi-agent

conversational frameworks to provide a more interactive and dynamic advisory experience.

VIII. ACKNOWLEDGEMENTS

This study was funded by the Technical Cooperation Project of the Japan International Cooperation Agency (JICA).

REFERENCES

- [1] D. T. Thanh and H. T. K. Lien, "Machine learning technology in predicting and preventing diseases for farmed shrimp in the mekong delta," *Journal of Biotechnology*, 2022.
- [2] X. Ran, Y. Liu, H. Pan, J. Wang, and Q. Duan, "Shrimp phenotypic data extraction and growth abnormality identification method based on instance segmentation," *Computers and Electronics in Agriculture*, vol. 229, p. 109701, 2025.
- [3] V. K. Yadav, L. Paul, A. A. Sambhe, V. S. Bharti, A. Sharma, and A. Deo, "An integrated framework for shrimp detection, tracking, counting, and behavioral analysis using yolov8 and deep learning-based computer vision model," in *ICSTSA'24*, 2024, pp. 311–323.
- [4] S. L. Chen and J. S. Liu, "Development of a real-time disease monitoring system for aquaculture using iot and machine learning," *Sensors*, 2021.
- [5] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei, "Agentic retrieval-augmented generation: A survey on agentic rag," *arXiv preprint arXiv:2501.09136*, 2025.
- [6] R. Qu, R. Tu, and F. Bao, "Is semantic chunking worth the computational cost?" *arXiv preprint arXiv:2410.13070*, 2024.
- [7] Q. L. Da, P. T. Nghia, T. T. Hung, and N. C. Ngon, "Testing ai algorithms in shrimp disease detection through images," *Can Tho University Journal of Science*, vol. 57, 2021.
- [8] H. Zhou, S. H. Kim, S. C. Kim, C. W. Kim, S. W. Kang, and H. Kim, "Instance segmentation of shrimp based on contrastive learning," *Applied Sciences*, vol. 13, no. 12, 2023.
- [9] T. Chau, "Ben tre: Effectiveness of high-tech shrimp farming," *Vietnam Fisheries Magazine*, 2025.

Multi-Stage User Intent Classification for Educational AI Agents

Quang-Minh Doan¹, Quang-Chieu Ta¹, Quang-Hoan Nguyen^{1*}, Hong-Viet Tran²

¹*Faculty of Computer Science and Engineering, Thuyloi University, Hanoi, Vietnam*

minhdqtt@gmail.com, quangchieu.ta@tlu.edu.vn, quanghoanptit@gmail.com

²*University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam*

thviet@vnu.edu.vn

** The Corresponding Author*

Abstract—User intent classification is a crucial step in building AI Agent systems. In this paper, we present and compare three intent classification approaches: (1) using large language models (LLMs – specifically LLaMA), (2) approximate matching via vector embedding (vector search), and (3) a lightweight pre-trained deep learning model, SetFit. Within the scope of an Agent system supporting academic services, each method demonstrates its own advantages and limitations. An intelligent combination of these approaches yields significantly improved prediction quality and reliability compared to relying on a single method alone. Experimental results indicate that although SetFit achieves the best standalone performance, combining all three methods in a conditionally tiered framework or through multi-source voting significantly improves the overall accuracy (achievement Accuracy 92.1% and Precision 92.5%) and reliability of the system.

Key words—Intent Classification, Agentic AI, LLMs

I. INTRODUCTION

The emergence of large language models (LLMs) such as GPT, LLaMA, Claude, and Gemini has significantly advanced the state of natural language understanding and generation in artificial intelligence systems [1]–[3]. These models serve as the core enablers for developing intelligent agents capable of interpreting user context and autonomously executing tasks on the user’s behalf.

The concept of Agentic AI represents an evolutionary step beyond traditional AI Agents, emphasizing systems that function as autonomous agents capable of making decisions and executing sequences of actions to achieve specific goals [4]. A fundamental prerequisite for such autonomy is the ability to accurately interpret user intent, which serves as the initial and essential stage in the processing pipeline of natural language interaction systems.

Intent classification can be regarded as a form of pre-command parsing that plays a routing role in directing user input to the appropriate functional components within the system. In student support systems, this entails distinguishing among various types of requests such as timetable inquiries, tuition-related queries, academic regulation lookups, and other issues pertaining to student life. Inaccurate intent classifica-

tion can result in misrouted service calls, inappropriate system responses, and a diminished user experience [5].

Traditional approaches to intent recognition have largely relied on classification models built upon handcrafted features or conventional machine learning algorithms [6]. However, with the widespread adoption of deep learning—particularly large language models (LLMs) and representation learning techniques—semantic intent classification has become significantly more effective and practically feasible [7], [8].

In this paper, we propose and analyze three methods for user intent classification, implemented and evaluated in the context of building a virtual assistant system for university students:

- (1) Leveraging a large language model (LLM), specifically LLaMA, for prompt-based intent classification [3];
- (2) Employing semantic vector search to assign intent labels through approximate matching in an embedding space [7], [9];
- (3) Training a lightweight deep learning model, SetFit, to perform fast and effective classification on limited labeled data [8].

The objective of this study is to evaluate the effectiveness of each individual method as well as the potential benefits of integrating them within a multi-stage intent analysis framework. The overarching goal is to enhance the accuracy, generalizability, and scalability of AI Agents deployed in higher education environments.

II. ADOPTED APPROACHES FOR USER INTENT CLASSIFICATION

A. Large Language Models (LLMs)

Advancements in large language models such as GPT, BERT, and LLaMA have fundamentally transformed how contextual understanding and intent recognition are approached in natural language processing [1]–[3]. These models support zero-shot and few-shot intent classification by leveraging prompt-

based mechanisms, eliminating the need for task-specific fine-tuning [10], [11].

Various techniques—such as instruction tuning, in-context learning, and chain-of-thought prompting, have been explored to enhance the performance of this approach in low-resource classification settings [2], [10]. Nevertheless, these methods share inherent limitations, particularly their sensitivity to prompt formulation and the considerable computational resources required for real-world deployment.

B. Semantic Vector-Based Retrieval (Vector Search)

An alternative approach involves mapping user queries and labeled intent examples into a shared semantic vector space using encoder-based models such as Sentence-BERT [7] or MiniLM [12]. Intent classification is then performed by computing cosine similarity between the query vector and the set of pre-embedded labeled examples. This method offers notable advantages in terms of low computational cost, ease of deployment, and scalability in practical applications. As a result, it has been widely adopted in systems based on Retrieval-Augmented Generation (RAG) [9].

Despite these benefits, the effectiveness of vector search is inherently dependent on the diversity and coverage of the labeled example set. In scenarios where the data lacks representational variety, the system may struggle to handle semantically novel or syntactically varied user queries [5].

C. Pretrained Deep Classification Models (SetFit, BERT Fine-tuning)

SetFit (Sentence Transformer Fine-tuning) [13] is a lightweight and data-efficient text classification framework designed to operate effectively in few-shot settings. It leverages the semantic representation capabilities of pretrained Sentence Transformers [7], as introduced in [8], without requiring full end-to-end fine-tuning of the underlying model. Figure 1 shows the architecture of SetFit.

SetFit¹ training pipeline consists of two stages:

- (1) A contrastive learning phase [14], in which sentence pairs are generated from a small labeled dataset. Positive pairs (from the same class) and negative pairs (from different classes) are used to adjust the embedding space such that semantically similar sentences are positioned closer together.
- (2) A classification phase, where a simple linear model—typically logistic regression or a linear support vector machine (SVM)—is trained on the resulting sentence embeddings to predict intent labels [8]

¹<https://github.com/huggingface/setfit>

Input Encoding: Given an input sentence x , the encoding process maps it into a fixed-dimensional dense vector representation using a sentence encoder:

$$\mathbf{h} = f_{\text{encode}}(x) \in \mathbb{R}^d \quad (1)$$

where:

- f_{encode} denotes the sentence encoder function (e.g., Sentence-BERT [7]);
- $\mathbf{h}$ is the output embedding vector corresponding to sentence x ;
- d is the embedding dimensionality, typically 384 or 768.

Fine-tuning the Encoder: During the contrastive learning stage, the encoder is fine-tuned using sentence pairs (x_i, x_j) sampled from the labeled dataset. The objective is to adjust the embedding space such that:

- if $x_i \sim x_j$ (i.e., both belong to the same class — positive pair), then $\cos(f(x_i), f(x_j)) \uparrow$;
- if $x_i \not\sim x_j$ (i.e., from different classes — negative pair), then $\cos(f(x_i), f(x_j)) \downarrow$,

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between two embedding vectors. This encourages semantically similar samples to be closer in the vector space, while pushing dissimilar samples apart.

Contrastive Loss Function: The contrastive learning objective aims to optimize the sentence encoder so that similar sentence pairs are embedded closer together, while dissimilar pairs are pushed apart. The loss function is defined as:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k=1}^N \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)} \quad (2)$$

where:

- $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity between embedding vectors;
- τ is a temperature scaling parameter that controls the sharpness of the distribution;
- N is the number of samples in the contrastive batch.

Classification Training: In the second stage, a linear classifier is trained on the sentence embeddings produced by the fine-tuned encoder. The predicted class distribution is computed using a softmax function as follows:

$$P(y | x) = \text{softmax}(W \cdot \mathbf{h} + \mathbf{b}) \quad (3)$$

where:

- W is the weight matrix of the classifier;
- $\mathbf{b}$ is the bias vector;
- $\mathbf{h}$ is the sentence embedding of input x .

Classification Loss Function: The training objective for the classification stage is to minimize the cross-entropy loss between the predicted class distribution and the ground truth label. The loss is defined as:

$$\mathcal{L}_{\text{class}} = -\sum_{i=1}^n y_i \log P(y_i | x) \quad (4)$$

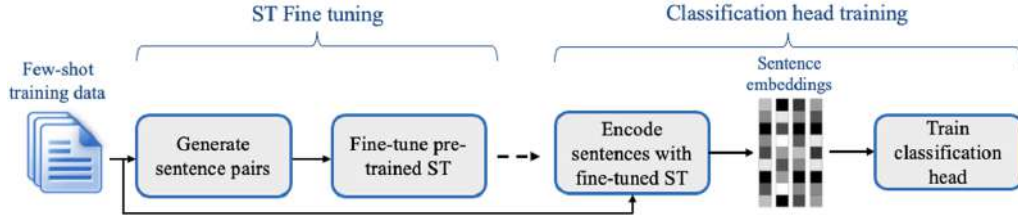


Figure 1. SETFIT's fine-tuning and training block diagram [13].

where:

- y_i is the one-hot encoded ground truth label for the input instance;
- $P(y_i | x)$ is the predicted probability of class y_i given input x .

III. PROPOSED METHOD

Based on the characteristics of each approach and the specific context of user intent analysis for university students, we propose a conditionally tiered strategy designed to optimize both computational efficiency and classification performance.

As shown in Figure 2, the system operates in three stages. First, a SetFit-based classifier is used to perform fast and high-accuracy classification on domain-specific inputs [8]. If the classification confidence is below a certain threshold, the query is passed to a semantic vector search component, which compares the encoded query with a representative set of labeled intent examples using cosine similarity [7]. In the final stage, if no confident match is found, the system falls back to a large language model (LLaMA) and uses prompt-based classification to infer the most appropriate intent [3].

This hierarchical structure allows the system to balance performance and robustness: it leverages fast classification when possible, while still maintaining the capacity to generalize and adapt to previously unseen or ambiguous queries through fallback mechanisms. Similar hybrid strategies have also been explored in prior research on retrieval-augmented and multi-stage architectures [9], [10].

A. Fast Classification Using SetFit

At this stage, the system employs a SetFit model that has been pretrained on a labeled dataset. Owing to its efficiency and accuracy in domain-specific settings, SetFit serves as the primary classifier within the pipeline.

If the model's confidence score exceeds a predefined threshold (e.g., 0.9) and the prediction is unambiguous (i.e., yields no more than one plausible label), the corresponding intent is accepted directly without invoking subsequent stages.

B. Semantic Vector Search

If the SetFit classifier produces a low-confidence or ambiguous prediction, the system proceeds to a secondary classification stage based on vector search. The process involves:

- (1) encoding the input query into a semantic vector space using Sentence-BERT (e.g., MiniLM or paraphrase-Multilingual models);
- (2) computing cosine similarity between the query vector and a curated set of representative samples for each intent class;
- (3) selecting the corresponding intent label if the highest similarity score exceeds a predefined threshold (e.g., 0.75).

This stage is designed to address cases involving novel or unseen data that may not have been captured during the SetFit training phase.

C. LLM-Based Classification

If both previous stages fail to yield a confident or consistent classification result (e.g., low confidence scores or conflicting outcomes from vector search), the system invokes an LLM-based intent classifier using prompt-based inference with the LLaMA model [3]. The model is prompted with an instruction to classify the student query into one of three predefined categories:

- **Schedule:** timetable, exam schedule, class hours, etc.
- **Regulations:** tuition, course withdrawal, academic rules, etc.
- **Other:** any other topics not covered above.

An example prompt used with the LLaMA model is shown below:

Prompt Example

You are a question classification system for university students. Classify the following question into one of three categories:

- Schedule (timetable, exam schedule, class hours)
- Regulations (tuition, course withdrawal, academic policies)
- Other (all other contents)

Question: "Hiện giờ tôi còn nợ học phí không?"

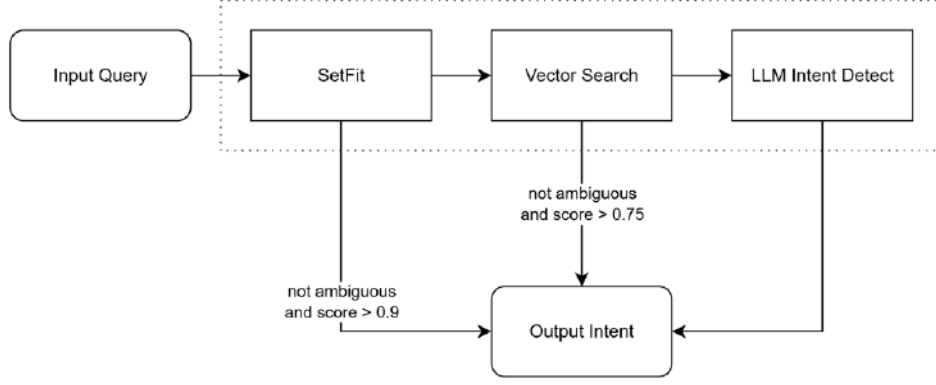


Figure 2. Intent Detection Workflow in our propose

The output returned by the LLaMA model in response to this prompt is used as the final predicted intent when preceding stages do not produce satisfactory results.

D. Conditional Multi-Stage Intent Classification Strategy (pseudocode)

To effectively balance computational efficiency and prediction robustness, we implement a conditional multi-stage intent classification pipeline combining SetFit, semantic vector search, and LLM-based classification (*Algorithm: Multi-Stage Intent Classification with Fallback*)

Algorithm 1 Multi-Stage Intent Classification with Fallback

Require: Input query x

Ensure: Predicted intent label y

```

1: Encode  $x$  to vector  $h \leftarrow f_{\text{encode}}(x)$ 
2:  $[\hat{y}_{\text{SetFit}}, p_{\text{SetFit}}] \leftarrow \text{SetFit.predict\_proba}(x)$ 
3: if  $p_{\text{SetFit}} > \tau_{\text{SetFit}}$  and  $\text{Top-2 margin} > \delta$  then
4:   return  $\hat{y}_{\text{SetFit}}$ 
5: end if
6:  $[\hat{y}_{\text{Vec}}, s_{\text{Vec}}] \leftarrow \text{VectorSearch}(h)$ 
7: if  $s_{\text{Vec}} > \tau_{\text{Vec}}$  and  $\text{Top-2 margin} > \delta$  then
8:   return  $\hat{y}_{\text{Vec}}$ 
9: end if
10: return  $\text{LLM.predict\_intent}(x)$ 

```

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

We evaluated the proposed multi-stage intent classification framework using a manually collected dataset comprising 250 real-world student queries. These queries span three major categories: academic regulations, timetabling, and other university-related topics. The data was sourced from publicly available discussions on student forums, specifically two Facebook groups affiliated with Thuyloi University:

- TLU Confessions: <https://www.facebook.com/groups/daihocthuylolvietnam/>

- TLU Student Community: <https://www.facebook.com/groups/649503585567449/>

Each query was anonymized and labeled according to its underlying intent by human annotators. The resulting dataset serves as a representative sample of the types of student inquiries encountered in practical deployments of educational AI assistants. The dataset was manually annotated into three intent classes: *schedule*, *policy*, and *other*. We split the data into 70% training, 15% validation, and 15% test sets. All experiments were conducted in Google Colab with a Tesla T4 GPU and 16GB RAM.

The SetFit model was fine-tuned on the training set using the `paraphrase-mpnet-base-v2`² encoder. The vector search used the same encoder to embed both the user query and intent examples. LLaMA 2-7B was queried by prompt-based classification using three-shot in-context examples.

The thresholds used in each stage of the hybrid pipeline were determined empirically based on validation experiments. For instance, a confidence score of 0.9 for SetFit was chosen to ensure high precision before falling back to vector search, while a similarity threshold of 0.75 provided a good balance between retrieving relevant candidates and avoiding false positives. We acknowledge that these values were not exhaustively optimized and may vary across datasets; more systematic hyperparameter tuning will be considered in future work.

B. Evaluation Metrics

We report standard metrics including Accuracy, Precision, Recall, and F1-score for each approach individually as well as for the hybrid strategy. In Table I we show the evaluation metrics intent classification performance comparison.

²<https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2>

C. Results and Analysis

Table I
INTENT CLASSIFICATION PERFORMANCE COMPARISON

Method	Accuracy	Precision	Recall	F1-score
SetFit	87.3%	88.1%	86.5%	87.3%
Vector Search	81.2%	82.5%	79.0%	80.7%
LLM (LLaMA)	85.4%	86.2%	84.0%	85.1%
Hybrid (Ours)	92.1%	92.5%	91.3%	91.9%

The hybrid approach significantly outperforms all individual methods, confirming that the conditional fallback strategy improves classification robustness. SetFit demonstrates the highest standalone accuracy, while vector search performs well on paraphrased queries. The LLM (LLaMA) is most effective in handling previously unseen queries and intent ambiguity but is computationally more expensive.

D. Ablation Study

To assess the contribution of each component, we conducted an ablation study by selectively removing modules from the pipeline:

- Removing LLM fallback reduces overall accuracy to 88.7%.
- Removing vector search causes misclassifications for novel phrasing. For example: "Em có còn nợ học phí không ?" ("Do you still owe tuition?").
- Relying solely on SetFit leads to high performance in in-domain cases, but shows fragility for edge cases.

E. Case Study

To further demonstrate the applicability of the proposed system, we present a simple real-world scenario. Suppose a student asks: "*Bao giờ thi môn Toán rồi rạc?*". The system first applies the hybrid intent detection pipeline to classify the query into the *schedule* domain. It then retrieves the corresponding timetable from the academic database and generates a natural language response via the LLM component. This process illustrates how the system supports students in accessing academic information in an efficient and user-friendly manner.

Another example concerns institutional regulations. When a student queries: "*Điều kiện xét tốt nghiệp như thế nào?*", the system identifies the *policy* domain, searches the relevant academic regulations, and provides a concise explanation. Such scenarios highlight the usefulness of the proposed agent in assisting students with both routine tasks and policy-related inquiries.

F. Limitations

Our evaluation is restricted to SetFit, vector search, and LLaMA, without comparisons to other common intent classifiers such as fine-tuned BERT or CN-N/LSTM, which limits the strength of the empirical

claims. In addition, the dataset is relatively small and drawn from a single university, making the results less generalizable. Broader benchmarks with additional methods and larger, multi-institutional datasets are left for future work.

V. CONCLUSION

This paper has presented and evaluated three user intent detection approaches in the context of deploying an AI Agent system for student support: (1) prompt-based classification using a large language model (LLM – LLaMA), (2) semantic vector search via embedding similarity, and (3) fine-tuning a lightweight deep learning model, SetFit. Each method offers distinct advantages: LLaMA demonstrates strong generalization capabilities, vector search is simple and scalable, while SetFit delivers high accuracy in domain-specific scenarios.

Experimental results indicate that although SetFit achieves the best standalone performance, combining all three methods in a conditionally tiered framework or through multi-source voting significantly improves the overall accuracy and reliability of the system. This integrated approach also helps mitigate common challenges such as training data bias, semantic undercoverage, and suboptimal prompt design—factors that frequently hinder the performance of automatic semantic dialogue systems.

The proposed strategy is highly practical for real-world applications, especially within higher education settings, where user input is linguistically diverse and intent is often implicit or ambiguous. Looking ahead, the system can be extended to support more complex tasks such as Named Entity Recognition (NER), multi-intent detection, and multi-turn dialogue management—advancing toward the development of AI Agents with increasingly robust agentic capabilities, adaptable to the evolving needs of students.

REFERENCES

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.
- [3] H. Touvron, T. Lavril, G. Izacard *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [4] J. Andreas, "Language models as agents," *arXiv preprint arXiv:2209.00788*, 2022.
- [5] C. Zhang and P. Wang, "A survey on intent detection for dialogue systems," *ACM Computing Surveys*, 2021.
- [6] G. Tur, D. Hakkani-Tur, and L. Heck, "What is left to be understood in intent classification?" in *IEEE SLT*, 2010.
- [7] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *EMNLP*, 2019.
- [8] T. Schick, H. Schütze *et al.*, "Setfit: Efficient few-shot learning without prompting," *arXiv preprint arXiv:2209.11055*, 2022.
- [9] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and V. Stoyanov, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *NeurIPS*, 2020.
- [10] V. Sanh, A. Webson, C. Raffel *et al.*, "Multitask prompted training enables zero-shot task generalization," in *ICLR*, 2022.

- [11] T. Schick and H. Schütze, “It’s not just size that matters: Small language models are also few-shot learners,” in *NAACL*, 2021.
- [12] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” 2020. [Online]. Available: <https://arxiv.org/abs/2002.10957>
- [13] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, “Efficient few-shot learning without prompts,” 2022. [Online]. Available: <https://arxiv.org/abs/2209.11055>
- [14] T. Gao, X. Yao, and D. Chen, “Simcse: Simple contrastive learning of sentence embeddings,” in *EMNLP*, 2021.

RAID: Robust Out-of-Distribution Intent Detection with Augmented and Discriminative Training

Hoang-Duy Vu, Kieu-Anh Truong Thi, Ngoc-Trang Thi Hoang, Duc-Trong Le*

VNU University of Engineering and Technology, Hanoi, Vietnam

{21021469, kieuanhtt, tranghtn, trongld}@vnu.edu.vn

Abstract—Out-of-distribution (OOD) intent detection is crucial in task-oriented dialogue systems, as models frequently encounter queries with intents beyond their training scope. This presents a significant challenge, requiring robust mechanisms to identify unseen OOD samples. To address this, we propose RAID, a method that integrates augmented representation learning and discriminative training with Feature Mixup. Specifically, RAID employs the Fast Gradient Sign Method (FGSM) together with feature mixup to generate augmented representations, improving model robustness and generalization to unseen intents. In addition, a discriminative training strategy based on negative IND–OOD pairs is introduced to refine decision boundaries and enhance semantic separation in the embedding space. We evaluate RAID on the Banking77 and CLINC150 datasets, where 50% of intents are treated as OOD. Results demonstrate that RAID significantly outperforms baselines, validating its effectiveness for OOD intent detection.

Index Terms—OOD intent detection, feature mixup, augmented representation, discriminative training, FGSM

I. INTRODUCTION

Out-of-distribution (OOD) detection has emerged as a crucial capability for improving the robustness and reliability of modern machine learning systems. In real-world applications such as conversational AI, autonomous driving, and healthcare, models frequently encounter queries or inputs that differ from their training data distribution. Misclassifying such OOD inputs as in-distribution (IND) can result in overconfident and potentially harmful predictions, eroding user trust and compromising safety [1]. Therefore, effective OOD detection is essential to prevent erroneous responses and maintain system reliability in safety-critical domains.

Recent advances in OOD detection span various modalities, including computer vision, natural language processing, and speech recognition. Established approaches include confidence-based methods such as Maximum Softmax Probability (MSP), scoring-based energy methods, and distance-based techniques like Mahalanobis distance [2]. Generative models, e.g., Variational Autoencoders and Generative Adversarial Networks, explicitly model the training distribution to detect anomalies [3]. Furthermore, large language models (LLMs) and vision-language models have recently been leveraged to integrate OOD detection into broader tasks like anomaly detection [4]. Despite these advances, distinguishing semantically similar IND and OOD samples remains challenging, particularly in intent detection [5], [6].

OOD intent detection, a specialized subproblem in NLP, aims to determine whether a given user query belongs to a known intent class or represents a novel intent not seen during training. Traditional methods rely on pre-defined thresholds to separate IND from OOD samples [7], [8], often using multiple classifiers, loss-based scoring, or clustering-based anomaly detection [9], [10]. More recent approaches focus on geometric properties of the feature space [12] or reformulating the task into a $(K+1)$ -way classification problem, where the additional class explicitly represents OOD queries [13]. Techniques such as feature mixup [16], phase distortion [15], and latent space generation [17] have been explored to generate synthetic OOD data and improve detection robustness. Probabilistic label estimation [14] further addresses uncertainty in OOD classification. While these approaches demonstrate notable progress, important gaps remain in effectively leveraging training strategies to enhance semantic discrimination between IND and OOD queries.

Building on these insights, we present **RAID**, a robust and unified framework for Out-of-Distribution Intent Detection using Augmented and Discriminative Training. It effectively mitigates semantic overlap while significantly enhancing generalization to previously unseen intents. The key contributions are summarized as follows:

- *Augmented Representation*: We propose an augmented training method that combines Fast Gradient Sign Method (FGSM) with feature mixup to generate augmented representations for IND and OOD samples. It enhances the model’s ability to generalize to unseen intents by leveraging adversarial perturbations and interpolated feature representations, improving OOD detection performance.
- *Discriminative Training Strategy*: We introduce a discriminative training approach to better distinguish between IND and OOD intents. This method optimizes the model’s decision boundary, enhancing its capability to accurately classify OOD samples while maintaining performance on in-distribution data.
- *Extensive Evaluation*: We conduct comprehensive experiments on two benchmarking intent detection datasets (Banking77 and CLINC150). Our thorough analysis demonstrates the effectiveness of the proposed methods, providing insights into their impact on OOD intent detection performance.

The remainder of this paper is organized as follows. Sec-

* Corresponding author.

tion II reviews prior research on out-of-distribution (OOD) detection and OOD intent detection, highlighting existing challenges and motivating our proposed approach. Section III presents the details of the proposed RAID framework while section IV describes datasets, experimental setup, evaluation metrics and results. Finally, Section V summarizes the key findings and outlines directions for future work.

II. RELATED WORK

OOD detection: Out-of-distribution (OOD) detection is essential for ensuring the reliability of machine learning models by identifying inputs that deviate from their training data distribution. This is particularly critical in applications like conversational AI, where misclassifying OOD queries can erode user trust, and in safety-critical domains such as autonomous driving or healthcare [1]. OOD detection mitigates overconfident errors and improves model robustness.

Common approaches include Maximum Softmax Probability (MSP) for confidence-based detection, energy-based scoring methods, and distance-based measures such as Mahalanobis distance [2]. Generative models, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), explicitly learn the data distribution to flag anomalies, while outlier exposure methods train on additional OOD samples [3]. These techniques have been widely applied across computer vision, NLP, and speech recognition. Recent work has also explored large language models and vision-language models to enhance OOD detection, integrating it with related tasks like anomaly detection [4]. Despite these advances, challenges persist in distinguishing semantically similar IND and OOD samples, where contrastive learning [5] and adversarial training [6] have shown notable promise.

OOD intent detection: OOD intent detection focuses on identifying whether a user query in a dialogue system corresponds to any predefined intent in the training set (IND) or represents an unseen intent (OOD). Traditional methods often rely on setting thresholds on model confidence scores [7], [8], building multiple classifiers for each intent, or applying loss-based strategies to separate OOD from IND queries. Other early approaches include SEG [9] and LMCL [10], which use outlier detection algorithms such as Local Outlier Factor (LOF) [11] and Gaussian clustering to detect OOD inputs.

Recent advances have moved toward modeling the geometric properties of the representation space. Zhang et al. [12] proposed identifying class centroids and radii to define OOD boundaries, improving separability in embedding space. Zhan et al. [13] redefined OOD detection as a $(K + 1)$ -way classification task with an explicit OOD class, alongside a feature mixup strategy to generate high-quality OOD samples. Lang et al. [14] further introduced probabilistic label estimation, assigning OOD samples a distribution over known labels rather than a binary OOD tag, better capturing uncertainty.

In addition, several OOD data generation strategies have been proposed to enhance robustness, including Phase Distortion [15], Feature Mixup [16], Latent Generation [17], and Open-Domain Sampling [18]. Architectural improvements,

such as dual-head classification models, have also contributed to achieving state-of-the-art performance on OOD intent detection benchmarks. Although methods like Feature Mixup [16] have demonstrated significant effectiveness in generating synthetic OOD samples for training within the $(K + 1)$ -way classification framework [13], a major limitation lies in their reliance on random selection of samples for mixing, which often results in unstable training outcomes and inconsistent performance across runs. Thus, RAID incorporates training techniques to enhance the robustness of Feature Mixup.

III. METHODOLOGY

A. Task Formulation

Given a user query $x \in \mathcal{X}$, where $\mathcal{X}$ denotes the input space of all possible queries, e.g., text strings in natural language, the objective of intent classification is to map x to an intent label y from a predefined set of known K intents $\mathcal{Y}_{\text{IND}} = \{y_1, y_2, \dots, y_K\}$. The classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ is trained on a dataset $\mathcal{D}_{\text{IND}} = \{(x_i, y_i)\}_{i=1}^N$, where each x_i is an in-distribution (IND) query with a corresponding label $y_i \in \mathcal{Y}_{\text{IND}}$.

For OOD intent classification, the task is to recognize queries in $\mathcal{X}$ that do not align with any intent in $\mathcal{Y}$. To address this, the label set $\mathcal{Y}$ is extended with an OOD label denoted as y_{OOD} , forming the extended label set $\mathcal{Y} = \mathcal{Y}_{\text{IND}} \cup \{y_{\text{OOD}}\}$. Finally, the primary task is to train a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ that satisfies two conditions namely:

- For an IND query x , the IND intent $y \in \mathcal{Y}_{\text{IND}}$ is returned.
- For an OOD query x_{OOD} , the OOD label y_{OOD} is returned.

B. RAID: Out-of-Distribution Intent Detection with Augmented and Discriminative Training

Our method is inspired by the idea of using a $(K + 1)$ -way intent classifier for OOD intent detection [13]. To generate synthetic OOD samples, we adopt the Feature Mixup strategy [16], which combines embeddings of in-distribution (IND) utterances to create diverse and informative training examples. In addition, we incorporate augmented and discriminative training into a multi-step training pipeline. This combination is designed to enhance both intent classification and OOD detection by increasing data diversity and optimizing a tailored loss function that explicitly separates IND and OOD representations. An overview of the proposed approach is illustrated in Fig. (1). Specifically, the input is first passed through a pre-trained language model (PLM) as an encoder to obtain the hidden states, also referred to as feature representations. Subsequently, these original feature representations are fed into an FGSM (Fast Gradient Sign Method) block, which augments the original data in the representation space to introduce perturbations. Following this, both the original and augmented feature representations are processed through a feature mixup module to generate synthetic training samples labeled as OOD. Finally, the entire set of data—including IND utterances, augmented versions, and mixed OOD samples—is passed through a classification head to train a $(K + 1)$ -way classifier for the K IND intents and the additional OOD intent. A detailed description of each stage is presented below:

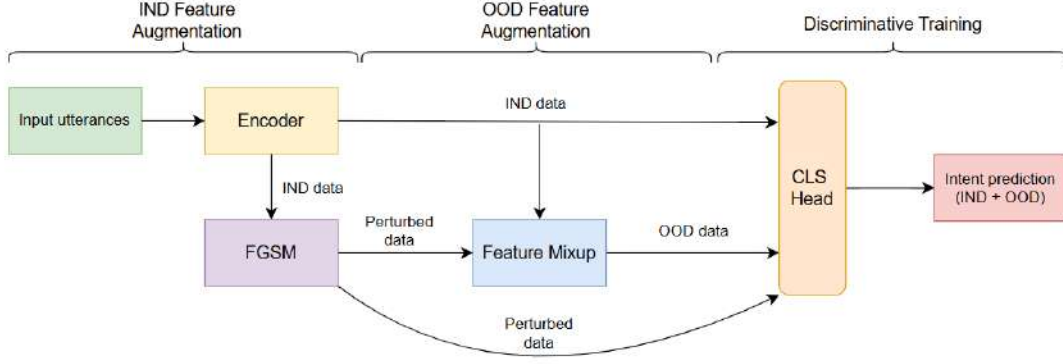


Fig. 1. Overall architecture of the RAID technique.

a) In-Distribution (IND) Feature Augmentation: For a given input query x_i , it is passed through an *Encoder*, e.g., BERT, RoBERTa, or DeBERTa, to obtain the semantic latent representation $\mathbf{e}_i \in \mathbb{R}^D$, with D is the output dimension of the Encoder. Inspired by Sengupta et al. [24], we generate an augmented presentation $\mathbf{z}_i$ by adding synthesized noise to improve the robustness of intent detectors. Here, we utilize the Fast Gradient Sign Method (FGSM) [23] to create $\mathbf{z}_i$ as the following procedure:

$$\mathbf{z}_i = \mathbf{e}_i + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x_i, y_i)) \quad (1)$$

where $J(\theta, x_i, y_i)$ is the cross-entropy loss of the intent classification task, $\nabla_x J(\theta, x_i, y_i)$ is the gradient with respect to x_i , $\epsilon \in [0, 1]$ is a small perturbation magnitude, and $\text{sign}(\cdot)$ extracts the direction of the gradient.

b) Out-of Distribution (OOD) Feature Augmentation: The augmented IND representation $\mathbf{z}_i$, together with the original IND vector $\mathbf{e}_i$, is passed through the Feature Mixup (FM) block to generate augmented OOD examples. The FM block produces examples that lie outside the IND distribution by interpolating the feature representations of two randomly selected samples with different IND intents within the same batch. This interpolation is performed in the latent space, ensuring that the resulting samples are semantically distinct from known intents. By introducing these diverse and informative synthetic OOD examples into the training set, the model is better equipped to recognize and isolate out-of-distribution queries during inference.

Feature Mixup Block. For each training batch, we randomly select a pair of input x_i, x_j with their respective latent representations $(\mathbf{e}_i, \mathbf{z}_i), (\mathbf{e}_j, \mathbf{z}_j)$. Subsequently, two OOD sample representations are generated via the following operations:

$$\mathbf{z}_{\text{OOD}}^{(1)} = \lambda \mathbf{e}_i + (1 - \lambda) \mathbf{e}_j \quad (2)$$

$$\mathbf{z}_{\text{OOD}}^{(2)} = \lambda \mathbf{z}_i + (1 - \lambda) \mathbf{z}_j \quad (3)$$

It is noted that an optimal number of OOD samples would be generated for training. The number is tuned to maximize the OOD detection performance without compromising the IND classification accuracy. This process ensures that the

OOD samples are diverse yet representative of the out-of-scope queries the model needs to detect.

c) Discriminative Training: Both IND and OOD examples, along with their respective representations, are fed into a classification (CLS) head. The CLS head is composed of a Linear Layer that reduces the input dimensionality by half, followed by a LeakyReLU activation, a Dropout Layer, and a final Linear Layer. The output of this last Linear Layer corresponds to the class logits, which are subsequently passed through a softmax function to yield the class probabilities over $K + 1$, i.e., K in-distribution (IND) intents and one OOD class. Actually, the model RAID is trained on three data types:

- *Original IND representation* $\mathbf{e}_i$ of a given input query x_i with its known intent label $y_i \in \mathcal{Y}_{\text{IND}}$;
- *Augmented IND representation* $\mathbf{z}_i$ of a given input query x_i with its known intent label $y_i \in \mathcal{Y}_{\text{IND}}$;
- *Augmented OOD representation* $\mathbf{z}_{\text{OOD}}$ generated from input queries with OOD label via Feature Mixup Block.

This training strategy enables the ability to learn from both standard and adversarial conditions, enhancing the model's robustness and OOD detection capabilities.

C. Optimization

The training process optimizes a composite loss function consisting of two main components namely:

a) Loss for original IND $\mathbf{e}_i$ and augmented OOD representation $\mathbf{z}_{\text{OOD}}$:

$$\mathcal{L}_{\text{IND}} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{DT} \quad (4)$$

where α is a trade-off hyperparameter, and $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss for intent classification:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \log p(y_i | x_i) \quad (5)$$

with N original IND samples, $p(y_i | x_i)$ is the predicted classification probability for a given IND training sample x_i using the embedding $\mathbf{e}_i$ and the label y_i . Equally important, $\mathcal{L}_{DT}$ is the discriminative loss computed as:

$$\mathcal{L}_{DT} = \frac{1}{M} \sum \max(0, \cos(\mathbf{e}_i, \mathbf{z}_{\text{OOD}}) - \tau) \quad (6)$$

TABLE I
THE STATISTICS OF EXPERIMENTAL DATASETS

Dataset	Training	Validation	Test IND	Test OOD
Banking77	4,446	494	1,520	1,560
CLINC150	7,500	1,500	2,250	3,450

where M is the number of augmented OOD samples, and τ is a cut-off threshold.

b) Loss for augmented IND $\mathbf{z}_i$ and augmented OOD representation $\mathbf{z}_{OOD}$: With the same way, it is not difficult to compute the augmented loss $\mathcal{L}_{AT}$ for augmented IND $\mathbf{z}_i$ and augmented OOD samples $\mathbf{z}_{OOD}$.

c) Combined loss: The overall loss is the average of the two losses as:

$$\mathcal{L}_{\text{overall}} = \frac{\mathcal{L}_{\text{IND}} + \mathcal{L}_{AT}}{2} \quad (7)$$

This structure ensures balanced optimization across clean and augmented data, enhancing both classification performance and OOD detection.

IV. EXPERIMENTS

A. Datasets

We conduct experiments on two widely adopted benchmarks for intent classification and OOD detection, including Banking77 and CLINC150 [22]. Banking77 contains 13,083 real banking queries annotated with 77 intents, such as “card-payment-not-recognised” and “transfer-fee-charged,” split into training (10,000), validation (1,000), and test (2,083) sets. It evaluates domain-specific performance within a single domain. CLINC150 covers 150 intents across 10 domains (15 intents per domain) with 22,500 IND samples and 1,200 explicit out-of-scope (OOS) queries, enabling multi-domain generalization assessment. Together, these datasets offer complementary challenges: Banking77 tests depth in a single domain, while CLINC150 evaluates breadth across diverse domains.

Inspired by Chen et al. [15], we employ a 50% setting, in which we randomly take 50% classes as OOD ones. Table I presents the statistics of the experimental datasets.

B. Evaluation Strategy

Referred to [14], we evaluate RAID using three metrics: Macro F1-All (F1-All), Macro F1-OOD (F1-OOD), and Macro F1-IND (F1-IND). F1-All provides a balanced measure of precision and recall across all classes, including IND intents and OOD, using macro-averaging to handle class imbalances.

To ensure reliability and reproducibility, we perform experiments with 10 random seeds, accounting for variability in model initialization and data shuffling. For each configuration, the model is trained and evaluated under all 10 seeds. We report the average values of three performance metrics - Macro F1-All (F1-All), Macro F1-OOD (F1-OOD), and Macro F1-IND (F1-IND)—across these runs. This strategy mitigates random fluctuations and provides stable, fair comparisons between different settings.

C. Implementation Setups

All models are trained on a Kaggle GPU T4×2 (16 GB) environment. Feature Mixup is applied during training to generate synthetic OOD samples on the fly. The pre-trained backbone is optimized with AdamW and learning rate 2×10^{-5} while the classification head (CLS Head) with Adam and learning rate 2×10^{-4} . Each mini-batch of size 32 includes 43 Feature Mixup OOD samples to ensure the fairest comparison with the baseline DCL method, as their paper identifies 43 OOD samples per batch as the most effective value for optimal performance. For In-distribution feature augmentation, the perturbation magnitude is configured as 0.3. For the $\mathcal{L}_{\text{IND}}$ loss, we empirically set the trade-off parameter $\alpha = 0.1$ and the margin threshold $\tau = 0.5$.

D. Baselines

Our first baseline model, namely DCL, is the original framework that employs pretrained language models (e.g., BERT, RoBERTa, DeBERTa) as the Encoder, and leverages Feature Mixup to train the model as a $K + 1$ -way classifier [13]. In addition, we compare our results against two established baseline approaches: Maximum Softmax Probability (MSP) and Adaptive Decision Boundary (ADB). MSP estimates the maximum softmax probability as a confidence score to distinguish IND from OOD samples, providing a simple yet effective heuristic for detection tasks [21]. ADB, proposed by Zhang et al. (2021) [12], introduces an adaptive decision boundary learned via a novel loss function that dynamically adjusts class boundaries to improve open-intent classification, particularly enhancing the detection of OOD samples in intent recognition tasks.

E. Main Result

Table II presents a comprehensive comparison of the proposed RAID model against baselines including MSP, ADB, and DCL. On Banking77, RAID achieves the best performance across all metrics, with an F1-All of 81.75, F1-OOD of 80.63, and F1-IND of 81.78. These values clearly surpass the baselines, indicating RAID’s robustness in handling both in-distribution and out-of-distribution samples. Particularly, the improvement in F1-OOD demonstrates RAID’s superior ability to detect unseen intents, which is critical for open intent classification tasks. Likewise, on CLINC150, RAID again outperforms all competing methods. It reaches an F1-All of 85.09, F1-OOD of 88.84, and F1-IND of 85.02, setting the highest scores across all evaluation dimensions. Compared to the strong baseline ADB (F1-OOD = 88.65), RAID still provides a slight but consistent improvement, showing that the model generalizes effectively even against advanced adaptive methods. Overall, the results confirm that RAID not only maintains strong in-distribution classification accuracy but also achieves substantial gains in OOD detection, establishing its effectiveness as a comprehensive solution for intent recognition tasks.

TABLE II
PERFORMANCE COMPARISON OF RAID VERSUS BASELINES ON BANKING77, CLINC150

Data	Model	F1-All	F1-OOD	F1-IND
BANKING77	MSP	71.18	41.19	71.97
	ADB	80.90	78.44	80.96
	DCL	80.07	78.66	80.44
	RAID	81.75	80.63	81.78
CLINC150	MSP	70.41	57.92	70.58
	ADB	85.05	88.65	85.00
	DCL	84.03	86.07	84.00
	RAID	85.09	88.84	85.02

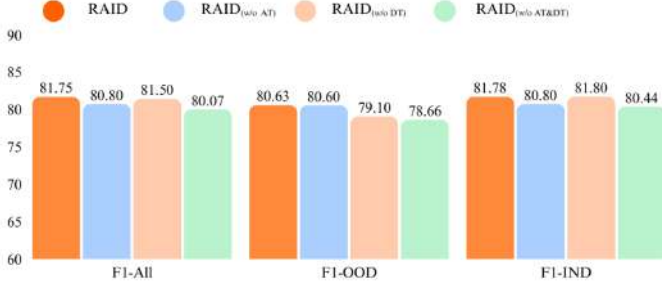


Fig. 2. Impact of Augmented & Discriminative Training On Model Performance for the Banking77 dataset. AT means ‘‘Augmented Training’’ while DT stands for ‘‘Discriminative Training’’.

F. Ablation Study

a) *Impact of data splitting strategy*: Since Banking77 does not provide explicit OOD samples, we partition its 77 intents (approximately 38 IND and 39 OOD) into three 50% split strategies: *best-split* (maximizing semantic separation), *worst-split* (minimizing semantic separation), and *random-split* (serving as a baseline). Although random splitting yields a challenging evaluation scenario, it can introduce ambiguity when semantically similar intents (e.g., card-payment-not-recognised and card-payment-issue) are assigned to different groups. As shown in Table III, using DCL with discriminative training, we observe a substantial performance gap across splits: the best-split achieves the highest F1-OOD (86.72%), whereas the worst-split drops to 75.78% due to semantic overlap. These findings highlight a key limitation of Banking77 for OOD evaluation and underscore the need for datasets explicitly designed for this purpose. In contrast, this issue is less pronounced in more diverse datasets such as CLINC150.

b) *Impact of Augmented and Discriminative Training*: To further examine the impact of discriminative training (DT), we evaluate the RAID architecture both with and without DT across two datasets. As shown in Fig. 2, incorporating DT consistently improves F1-OOD scores across all datasets. This highlights the effectiveness of DT in enhancing the model’s OOD detection capability while preserving its IND classification performance.

Augmented training enhances model robustness by introducing perturbed IND samples during training, thereby improving

TABLE III
DCL WITH DISCRIMINATIVE TRAINING PERFORMANCE ACROSS SPLITTING STRATEGIES ON BANKING77 (50%)

Splitting Strategy	F1-All	F1-OOD	F1-IND
Best-split	83.62	86.72	83.54
Worst-split	79.19	75.78	79.30
Random	80.76	80.62	80.77

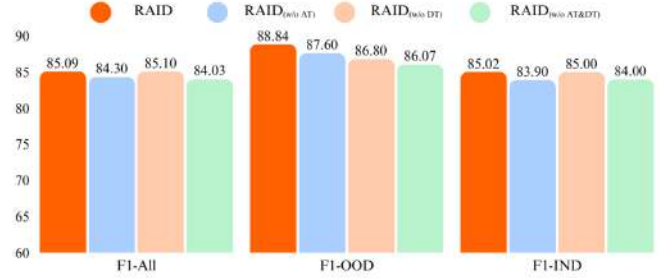


Fig. 3. Impact of Augmented & Discriminative Training On Model Performance for the CLINC150 dataset. AT means ‘‘Augmented Training’’ while DT stands for ‘‘Discriminative Training’’.

both IND classification and OOD detection. To assess its effectiveness, we compare models trained with and without augmented training (Fig.2, Fig.3). The results show consistent F1-IND improvements, demonstrating better generalization to IND variations. Furthermore, adversarial examples passed through the Feature Mixup block serve as effective OOD augmentation, confirming the benefits of augmented training for improving both IND and OOD performance.

c) *Impact of Backbone Model Selection*: To evaluate the impact of various backbone models on out-of-distribution (OOD) intent detection, we conducted experiments using three typical transformer-based models: BERT, RoBERTa [19], and DeBERTa [20]. The results are summarized in Fig 4.

Overall, BERT consistently outperforms RoBERTa and DeBERTa on both datasets, particularly in OOD detection, as reflected by its highest F1-OOD among the three models. This advantage may stem from BERT’s bidirectional training strategy, which more effectively leverages Feature Mixup-generated OOD samples. DeBERTa achieves better performance than RoBERTa, likely benefiting from its enhanced attention mechanisms, though it still falls short of BERT, possibly due to sensitivity to mixup parameters. RoBERTa’s lower performance, especially on Banking77, indicates that its optimization for large-scale pretraining does not translate as effectively to OOD intent detection without further adaptation. These findings underscore the superiority of the BERT model over its variants in feature-space operations.

V. CONCLUSION

In this work, we introduced **RAID**, a novel framework for out-of-distribution (OOD) intent detection that integrates Feature Mixup, augmented and discriminative training within an architecture. Through comprehensive experiments on the

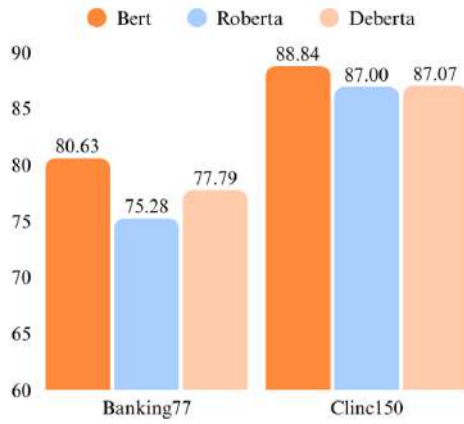


Fig. 4. F1-OD across backbone models on Banking77 and Clinc150.

Banking77 and Clinc150 datasets, our method consistently outperformed strong baselines, demonstrating its effectiveness in addressing the semantic overlap between in-distribution and OOD queries. Ablation studies further validated the contribution of each component to the overall performance. Future work will explore extending RAID to multi-turn dialogue settings, incorporating cross-lingual transfer, and leveraging unlabeled OOD data to further enhance robustness.

REFERENCES

- [1] Y. Yang, C. Zhang, Y. Lei, J. Lv, and K. Zhang, “Generalized Out-of-Distribution Detection: A Survey,” *International Journal of Computer Vision* 132.12, 5635–5662, 2023.
- [2] H. Lang, Y. Zheng, Y. Li, J. Sun, F. Huang, and Y. Li, “A Survey on Out-of-Distribution Detection in NLP,” In *TMLR*, 2023.
- [3] D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep Anomaly Detection With Outlier Exposure,” In *Proceedings of ICLR*, 2019.
- [4] A. Miyai *et al.*, “Generalized Out-of-Distribution Detection and Beyond in Vision Language Model Era: A Survey,” *arXiv preprint arXiv:2407.21794*, 2024.
- [5] Y. Wu *et al.*, “Revisit Overconfidence for OOD Detection: Reassigned Contrastive Learning With Adaptive Class-Dependent Threshold,” in *Proceedings of NAACL*, 2022, pp. 4165–4179.
- [6] Y. Wang *et al.*, “Improved OOD Generalization Via Adversarial Training and Pre-Training,” *arXiv preprint arXiv:2106.06925*, 2021.
- [7] L. Shu, H. Xu, and B. Liu, “DOC: Deep Open Classification of Text Documents,” *arXiv preprint arXiv:1709.08716*, 2017.
- [8] S. Ryu, S. Kim, J. Choi, H. Yu, and G. G. Lee, “Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems,” *Pattern Recognition Letters*, vol. 88, pp. 26–32, Mar. 2017.
- [9] L. Fan, G. Yan, Q. Li, H. Liu, X. Zhang, A. Y. S. Lam, and X.-M. Wu, “Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification,” in *Proceedings of ACL’20*, 2020, pp. 1050–1060.
- [10] T.-E. Lin and H. Xu, “Deep unknown intent detection with margin loss,” In *Proceedings of ACL*, 2019, pp. 5491–5496.
- [11] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: identifying density-based local outliers,” *SIGMOD Record*, vol. 29, no. 2, pp. 93–104, May 2000.
- [12] H. Zhang, H. Xu, and T.-E. Lin, “Deep open intent classification with adaptive decision boundary,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14374–14382, May 2021.
- [13] L.-M. Zhan, H. Liang, B. Liu, L. Fan, A. Y. S. Lam, and X.-M. Wu, “Out-of-scope intent detection with self-supervision and discriminative training,” in *Proceedings of ACL-IJCNLP*, 2021, pp. 3521–3532.
- [14] H. Lang, Y. Zheng, J. Sun, F. Huang, L. Si, and Y. Li, “Estimating soft labels for out-of-domain intent detection,” *arXiv preprint arXiv:2211.05561*, 2022.
- [15] D. Chen and Z. Yu, “Improving out-of-scope detection in dialogues using data augmentation,” In *EMNLP*, 2021.
- [16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” In *ICLR*, 2018.
- [17] P. Marek, V. I. Naik, V. Auvray, and A. Goyal, “OodGAN: Generative adversarial network for out-of-domain data generation,” *arXiv preprint arXiv:2104.02484*, 2021.
- [18] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, “Pretrained transformers improve out-of-distribution robustness,” *arXiv preprint arXiv:2004.06100*, 2020.
- [19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [20] P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-enhanced BERT with disentangled attention,” *arXiv preprint arXiv:2006.03654*, 2021.
- [21] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *arXiv preprint arXiv:1610.02136*, 2016.
- [22] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, “An evaluation dataset for intent classification and out-of-scope prediction,” in *Proceedings of EMNLP-IJCNLP*, China, Nov. 2019, pp. 1311–1316.
- [23] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” *arXiv preprint arXiv:1412.6572*, 2015.
- [24] Sengupta, S., Krone, J., and Mansour, S., “On the Robustness of Intent Classification and Slot Labeling in Goal-oriented Dialog Systems to Real-world Noise,” In *Proceedings of NLP4ConvAI*, 2021, p69-79.

EffWhiS: A Proposed Efficient Approach for Speech-to-Text Streaming Whisper

Mac Thi Quynh Nhu^{1*}, Nguyen The Hung¹, Vu Hoang Phuc², Nguyen Trung Dung¹,
Le Xuan Duc¹, Pham Thi Bich Van³

¹National Laboratory for Securing Information (NLSI), Vietnam Ministry of Defence, Hanoi, Vietnam

²School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam

³Institute of Information and Communication Technology, Le Quy Don Technical University, Hanoi, Vietnam
machthiquynhnhu.mta@gmail.com*, hungnguyenthe1211@gmail.com, phuc.vh214923@sis.hust.edu.vn,
ntdtoanud2011@gmail.com, lexuanducvn@gmail.com, vanptb@lqdtu.edu.vn

Abstract—Whisper Streaming is an emerging approach for real-time automatic speech recognition (ASR), extending OpenAI’s Whisper model, and using local agreement to commit transcriptions. However, its reliance on fixed chunk sizes and local agreement often causes unnecessary delays or premature outputs. To address this, we propose an efficient Whisper Streaming (EffWhiS), a refined streaming algorithm that enhances Whisper’s responsiveness and accuracy through two key strategies: (1) majority-vote commit, which commits a word if it appears in at least two of the last three decoding passes; and (2) segment consistency constraint, which allows commits only when consecutive segment predictions agree. The benchmark ESIC corpus and two hours of recordings of actual Vietnamese meetings are used to evaluate our proposed EffWhiS. The findings show that, in comparison to the original Whisper Streaming, EffWhiS achieves a lower Word Error Rate and a significantly lower Average Commit Latency. Notably, EffWhiS maintains high accuracy even at a short chunk size of 0.5 seconds, confirming its robustness for practical streaming ASR deployments. These improvements are consistent across various chunk sizes in both English and Vietnamese, highlighting the generalizability of our proposed approach. With its balance between accuracy and latency, EffWhiS offers a potential path for effective and deployable streaming ASR systems.

Index Terms—Speech to Text, Automatic Speech Recognition, Whisper streaming, Efficient Streaming

I. INTRODUCTION

Whisper [1], a cutting-edge automated speech recognition (ASR) model, is a solid basis for a variety of speech-related applications since it can recognize speech in multiple languages and translate. Faster-Whisper [2], an optimized variant, offers significant speed improvements through inference using CTranslate2, a quick inference engine for Transformer models, and a smaller memory footprint, making it well-suited for deployment on edge devices or resource-constrained environments. However, both models are primarily designed for offline inference, assuming full audio availability before processing and lacking native support for low-latency or streaming scenarios, which limits their applicability in real-time or interactive speech recognition tasks.

Unlike offline ASR, streaming ASR is essential for applications like live captioning, voice assistants, and interactive

systems since it concentrates on transcribing speech in real-time or with no latency [3]. Some recent research utilizes streaming inference over Whisper models, addressing the model’s issues in low-latency scenarios while leveraging its strengths in offline ASR conditions. The streaming methods used in these studies were designed to enable real-time transcription. Whispy [4] is a production-ready streaming wrapper for Whisper that combines an ffmpeg-based input pipeline, voice activity detection (VAD), contextual retranscription, hallucination filtering, and suggestion generation using edit distance to enable real-time transcription. Additionally, Whisper-T [5] is a streaming framework that enhances Whisper for low-latency deployment by incorporating CPU/GPU pipelining for efficient resource allocation on edge devices, beam trimming for faster alignment, and learnable “hush words” to mitigate hallucinations. The authors of the study [6] propose an additional technique, termed Simul-Whisper, which builds upon Whisper Streaming [7]. By utilizing cross-attention alignment for autoregressive decoding and introducing a truncation detection model to mitigate boundary errors, this method achieves strong performance without the need for model fine-tuning, thereby enabling chunk-based streaming ASR.

Nevertheless, similar to Whisper Streaming, the streaming ASR problem of striking a balance between speed and accuracy is a key challenge for the aforementioned study. While boosting context improves accuracy at the expense of responsiveness, reducing latency restricts contextual information and may degrade recognition quality. This is significant for real-time applications such as conversational agents or live captioning. Although it provides a workable solution for low-latency transcription, the original Whisper Streaming still relies on set chunk sizes and heuristics, such as local agreement, to assess output stability. Even though these techniques work well in most situations, they may cause unnecessary delays or premature outputs under challenging circumstances, such as overlapping speech, loud input, or rapidly changing conditions. These drawbacks highlight the need for a more flexible and contextually aware method of real-time speech detection.

In this paper, we propose several enhancements to the

Whisper Streaming framework to improve transcription quality and responsiveness in real-time meeting scenarios:

- Majority-vote commit: Commit a word if it appears in at least two of three decoding passes, relaxing the LocalAgreement-2 rule to boost responsiveness with minimal accuracy loss.
- Segment consistency constraint: Commit only when at least two consecutive segments agree, reducing unstable outputs and improving accuracy.

In the remaining paper, the introduction is given in Section I, the overview of related work is given in Section II, the suggested approach is described in section III, the experimental design is outlined in Section IV, and the findings and discussion are presented in Section V.

II. RELATED WORK

Whisper is a sequence-to-sequence Transformer model that was first presented by OpenAI in 2022 [1]. It was trained using 680,000 hours of supervised audio-text pairs that were gathered from the Internet. Whisper’s training corpus is noisy and extremely diverse, which helps explain its resilience across accents, contexts, and languages, in contrast to earlier ASR systems that frequently rely on curated datasets. It is one of the most flexible ASR models available, supporting 96 languages for English translation and 97 languages for transcription. Whisper uses a typical encoder-decoder Transformer architecture in theory. Interestingly, it integrates several speech-related tasks into a single multitask model, including language identification, voice-to-text, and speech translation. The CTranslate2 inference engine is used to reimplement the Whisper model in Faster-Whisper [8], which reduces latency and memory footprint by enabling hardware acceleration, quantization, and efficient batching. This makes Faster-Whisper more deployable in real-world scenarios, especially on edge or low-resource devices. However, because its slow inference speed renders it less suitable for real-time applications like voice interaction or live captioning, neither Whisper nor Faster-Whisper is viable for real-time speech recognition in streaming ASR settings.

Integrating streaming ASR into chatbot systems [9] or virtual conference secretaries [10] is essential for achieving low-latency, real-time speech transcription, enabling more natural and efficient voice-based interactions. Whisper Streaming uses a Whisper-like API to transform into a real-time streaming transcription system. It employs the LocalAgreement policy [11] (specifically, the LocalAgreement-2 method) to guarantee segment stability before finalization. This policy continually analyzes an audio buffer and commits the longest common prefix of consecutive outputs. Whisper Streaming, when tested on the ESIC dev dataset [12], maintains transcription quality that is comparable to offline performance while achieving an average latency of roughly 3.12 seconds on A40 with VAD enabled. This framework’s status as a reliable foundation for real-time ASR systems has been cemented by its resilience and usability in live multilingual conferences. Despite achieving constant output and low latency, the approach has several drawbacks. It utilizes the basic Silero VAD [13], which may

be less successful in noisy or multi-speaker settings, has a high computational cost due to recurrent inference, and lacks token-level confidence control for early commitment. Low-latency transcription is made possible by Whisper Streaming, but it still depends on set chunk sizes and local agreement techniques, which can lead to needless delays or premature output commitment. Whisper Streaming’s computational efficiency in the real world has not yet been evaluated because it has not been verified on inexpensive hardware. Furthermore, as mentioned in its original study [7], latency and quality results derived from the ESIC corpus could not translate well to other languages or dialects. Simul-Whisper outperforms the LocalAgreement strategy on the LibriSpeech test-clean set [14], achieving significantly lower word error rates under similar streaming constraints [6]. However, it does not provide explicit latency measurements or perform system-level optimizations, making its real-time responsiveness uncertain in practice.

III. EFFWHIS: A PROPOSED EFFICIENT WHISPER STREAMING

Figure 1 below shows the general workflow of our streaming ASR system, showing the data flow from the unstructured speech input to the structured output. Our technology can interpret audio from several speakers and provide structured transcripts that include speaker attribution. The figure illustrates how various models and processing stages combine to create a polished transcript. From initial transcription to diarization, speech identification, and final text formatting, every element of the pipeline contributes to increasing accuracy. Utilizing cutting-edge speech processing technology, this system is built to manage intricate multi-speaker settings effectively.

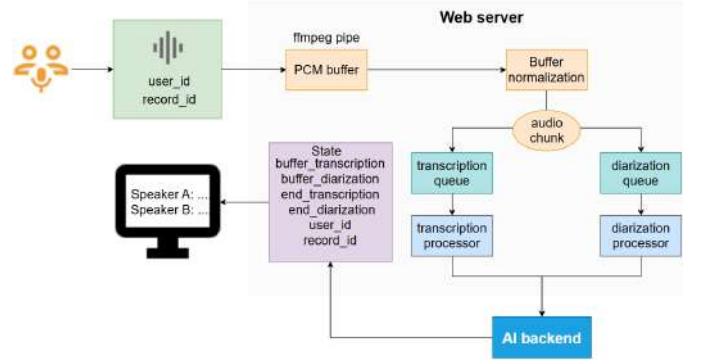


Fig. 1. The overall architecture of our streaming ASR system

In this paper, we present EffWhiS, an efficient Whisper Streaming to improve transcription quality and responsiveness in real-time meeting scenarios. There are two noteworthy improvements: First, the majority-vote commit relaxes the LocalAgreement-2 rule to increase responsiveness with minimal accuracy loss by allowing a word to be chosen if it appears in at least two of three decoding passes; second, the segment consistency constraint allows for committing only when at

least two consecutive segments agree, which reduces unstable outputs and improves accuracy.

A. Majority-vote commit

To accelerate word commitment without significantly compromising transcription quality, we propose the Majority-Vote Commit strategy, which adopts a more relaxed condition. Specifically, it compares the three most recent ASR transcripts and commits a word if it appears in at least two out of the three segments. This approach generalizes the original two-segment agreement strategy, while also being more tolerant of minor inconsistencies between successive predictions. By allowing flexible matching across segments, this method improves system responsiveness, making it particularly suitable for real-time ASR applications. It maintains a good balance between speed and accuracy, enabling faster feedback without introducing substantial errors. Algorithm 1 provides a detailed illustration of our suggested Majority-Vote Commit approach.

Algorithm 1 Flush Buffer with Majority-vote commit

Require: N : buffer of incoming uncommitted words,
 B : buffer of previous uncommitted words,
 F : buffer of first uncommitted words,
Ensure: C : list of committed words

```

1:  $C \leftarrow \emptyset$ 
2: while  $?N \neq \emptyset$  do
3:    $(a, b, t, p) \leftarrow N.\text{head}()$ 
4:   if  $B = \emptyset$  and  $N = \emptyset$  then
5:     break
6:   else if  $t = B.\text{head}().\text{text}$  or  $t = F.\text{head}().\text{text}$  then
7:      $C \leftarrow C \cup \{(a, b, t, p)\}$ 
8:     Remove head from  $F, B, N$ 
9:     Update last committed word and timestamp
10:  else
11:    break
12:  end if
13: end while
14:  $F \leftarrow B$ 
15:  $B \leftarrow N$ 
16: return  $C$ 

```

Figure 2 shows the main running steps in our proposed Majority-Vote Commit approach. Tokens with the same color box across updates are aligned and compared. If a token appears in at least two out of three updates, it is considered committed and shown in a green box. In the final line, these stable tokens are committed and displayed in green text without boxes.

B. Segment consistency constraint

Unlike Whisper Streaming’s default LocalAgreement-2 policy, which commits words by finding the longest common prefix between two consecutive decoding passes, we propose a stricter constraint that delays committing until the current output contains at least two segments.

Update N-1: To they will conduct

Update N: Today I will contact us for

Update N+1: Today I will conduct our meeting. First, my eye

Update N+1: Today I will conduct our meeting. First, I invite

Fig. 2. Illustration of Majority-Vote Commit.

Our motivation stems from the observation that each segment returned by faster-whisper often corresponds to a semantically coherent unit, such as a phrase or sentence. By waiting until at least two segments have been generated, we ensure that the model has accumulated enough context before applying the agreement check. This reduces the likelihood of premature or unstable outputs, particularly for longer or more complex utterances. Once this segment-length condition is satisfied, we apply the standard LocalAgreement-2 across two consecutive decoding passes to determine the committed portion. Algorithm 2 provides a detailed illustration of our suggested Segment consistency constraint approach.

Algorithm 2 Flush Buffer with Segment consistency constraint

Require: S : incoming ASR segments,
 N : buffer of incoming uncommitted words,
 B : buffer of previous uncommitted words
Ensure: C : list of committed words

```

1:  $C \leftarrow \emptyset$ 
2: if  $\text{len}(S) < 2$  then
3:   return  $C$ 
4: end if
5: while  $?N \neq \emptyset$  do
6:    $(a, b, t, p) \leftarrow N.\text{head}()$ 
7:   if  $B = \emptyset$  then
8:     break
9:   else if  $t = B.\text{head}().\text{text}$  then
10:     $C \leftarrow C \cup \{(a, b, t, p)\}$ 
11:    Remove head from  $B$  and  $N$ 
12:    Update last committed word and timestamp
13:  else
14:    break
15:  end if
16: end while
17:  $B \leftarrow N$ 
18: return  $C$ 

```

Figure 3 shows the main running steps in our proposed Segment consistency constraint approach. Words with the same underline color belong to the same segment. In Update N, only one segment is present, so the LocalAgreement-2 (LA-2) policy is not applied, and no output is committed. From Update N+1, two segments appear. The policy then compares the unconfirmed right part across two updates and finds the longest common prefix (highlighted with a green background).

Tokens in this prefix are committed and displayed in green text in Update N+2.

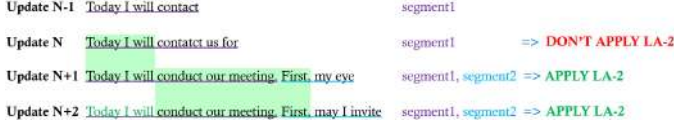


Fig. 3. Illustration of segment consistency constraint.

IV. EXPERIMENTAL DESIGN

This section describes the experimental configuration intended to address a research question: Does EffWhiS increase the latency and accuracy of Whisper streaming? This section provides the dataset, implementation specifics, and evaluation measures that were used to gauge the method’s effectiveness.

A. Datasets

- Real meeting recordings: a collection of 30 Vietnamese meeting audio sessions totaling approximately 2 hours. These recordings were collected in real meeting rooms under varying acoustic conditions.

B. Experimental setup

The models are run on a supercomputer running Ubuntu 20.04 and equipped with an Intel Core i7 CPU, 32GB of RAM, and an RTX 4090 24GB GPU. As indicated in Table I, we set up five system variants to assess the efficacy of the suggested approaches. To examine their respective effects, we first test the Segment Consistency Constraint (EffWhiS-cons) and Majority-Vote Commit (EffWhiS-maj) independently using the baseline streaming model (Base-WhisStream). EffWhiS-comb examines the complementary impacts of both tactics by combining them. Lastly, Offline-Whisper acts as a non-streaming reference for upper-bound performance and reflects the offline ASR.

All streaming variants are evaluated on the ESIC and Vietnamese administrative datasets, using chunk sizes of 0.5s, 1s, and 2s, with VAD enabled. This setup enables us to compare methods under various latency conditions and evaluate their contributions to recognition stability and accuracy. We use a streaming simulation without real-time delays: each chunk is processed immediately after the previous one, with no waiting time, so system latency equals the chunk size.

C. Evaluation method

Word Error Rate (WER) [15]: Measures the accuracy of transcription by comparing the system output with a ground truth reference. Lower WER indicates better transcription performance and reflects how well the system handles real-world audio.

$$\text{WER} = \frac{S + D + I}{N} \quad (1)$$

where:

- S is the number of substitutions (words that were incorrectly recognized),

TABLE I
EVALUATED VARIANTS OF STREAMING AND NON-STREAMING SYSTEMS

ID	System Variant	Description
Base-WhisStream	Baseline Whisper Streaming [7]	Baseline streaming inference using Faster Whisper.
EffWhiS-maj	Majority-Vote Commit	Commit a token if it appears in at least two out of three decoding passes.
EffWhiS-cons	Segment Consistency Constraint + Whisper Streaming	Commit only when two consecutive segments produce the same output.
EffWhiS-comb	Majority-Vote Commit + Segment Consistency Constraint	Combines both constraints for more stable and reliable output.
Offline-Whisper	Faster-Whisper [2] (non-streaming)	Offline decoding as reference upper-bound performance.

- D is the number of deletions (words that were omitted or missing in the hypothesis),
- I is the number of insertions (extra words added in the hypothesis), and
- N is the total number of words in the reference (ground truth).

Commit Latency: Average delay between the time a token’s corresponding audio is first streamed into the system and the time that token is committed in the transcription. For each token, the system records its initial input timestamp and compares it with its commit point. This metric reflects the system’s responsiveness and is used to compare different configurations. Since the system simulates streaming without real-time delays, the computed commit latency directly depends on the chunk size. Therefore, comparisons between configurations should be made using the same chunk size to ensure fairness.

Let:

- T_i^{input} : Timestamp when token i ’s corresponding audio is first received by the system.
- T_i^{commit} : Timestamp when token i is committed (finalized) in the transcription.
- n : Total number of tokens.

Then, the Average Commit Latency (ACL) is calculated as:

$$\text{ACL} = \frac{1}{n} \sum_{i=1}^n (T_i^{\text{commit}} - T_i^{\text{input}}) \quad (2)$$

V. RESULT AND DISCUSSION

A. Evaluation on Benchmark ESIC Dataset

The evaluation of our proposed approach on the ESIC development set is shown in Table II. With a maximum WER of 9.30% at 1 second (s), the results demonstrate that the baseline setup Base-WhisStream performs moderately across all chunk sizes, but neither latency nor accuracy exhibits any notable strengths. While its WER is still a little higher, EffWhiS-maj, which prioritizes responsiveness, achieves the lowest latency of 1.00s at 2s chunk size, making it perfect for

TABLE II
PERFORMANCE OF STREAMING ASR CONFIGURATIONS ON ESIC DEV DATASET

ID	0.5s		1s		2s	
	WER	ACL	WER	ACL	WER	ACL
Base-WhisStream	11.29	1.06	9.30	1.20	9.30	1.20
EffWhiS-maj	11.33	1.05	9.33	1.17	9.38	1.00
EffWhiS-cons	10.57	2.06	9.40	1.79	8.91	2.12
EffWhiS-comb	10.67	2.05	9.38	1.77	9.23	1.44
Offline-Whisper	WER average = 9.52					

real-time applications. In contrast, EffWhiS-cons prioritizes precision, achieving the highest WERs for all chunk sizes, including 10.57% at 0.5s and 8.91% at 2s, indicating its efficacy in steady transcribing. With a WER of 9.23% at 2s and better latency than EffWhiS-cons, EffWhiS-comb provides a solid compromise and is a well-balanced solution. According to Offline-Whisper, the three approaches all perform better than the baseline on average in either accuracy, latency, or both. The average WER for all three approaches is 9.45%.

B. Evaluation on real meeting Vietnamese recordings

The evaluation of our suggested approach using Vietnamese recordings of actual meetings is shown in Table III. The baseline Base-WhisStream, constrained by the stringent LocalAgreement-2 constraint, produces a WER of 10.01% with a latency of 2.0s at a 0.5s chunk size. By permitting earlier commits by majority voting, EffWhiS-maj greatly improves responsiveness (latency: 1.82s), although WER marginally rises to 10.32%. Conversely, EffWhiS-cons reduces WER to 8.86% by enforcing agreement across segment boundaries, but at the expense of a greater latency of 6.66s because of more stringent commit conditions.

TABLE III
PERFORMANCE OF STREAMING ASR CONFIGURATIONS ON THE VIETNAMESE DATASET

ID	0.5s		1s		2s	
	WER	ACL	WER	ACL	WER	ACL
Base-WhisStream	10.01	2.00	8.12	1.08	8.31	3.27
EffWhiS-maj	10.32	1.82	8.55	1.96	8.75	2.49
EffWhiS-cons	8.86	6.66	8.12	3.08	8.04	5.54
EffWhiS-comb	8.75	6.10	8.05	4.81	8.82	5.22
Offline-Whisper	WER average = 9.12					

By combining the greatest features of EffWhiS-maj and EffWhiS-cons, our integrated EffWhiS-comb achieves the lowest WER at 0.5s (8.75%) and a lower latency of 6.1s, making it accurate and responsive enough for streaming. EffWhiS-comb maintains its good performance with a WER of 8.05% at a chunk size of 1.0s. Crucially, the accuracy of all three streaming options is higher than that of non-streaming. Additionally, our approach shows that even with small chunk sizes, high recognition quality can be attained. Its potential for real-time transcribing workloads without compromising

performance is demonstrated by the fact that 0.5s already achieves optimal accuracy.

Figures 4 and 5 provide a visual summary of WER and ACL trends across different chunk sizes. As shown in Figure 4, our proposed methods consistently achieve lower WER compared to baselines, particularly at shorter chunk sizes (0.5s), demonstrating strong performance even under low-latency constraints. Figure 5 complements this by showing

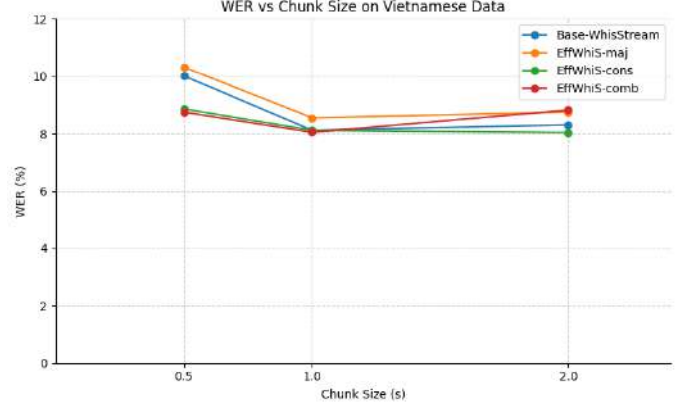


Fig. 4. WER vs chunk sizes on Vietnamese data

that while EffWhiS-cons sacrifices latency for accuracy, the combined approach (EffWhiS-comb) effectively balances both, maintaining competitive WER with reduced ACL, thereby offering a more practical solution for real-time applications.

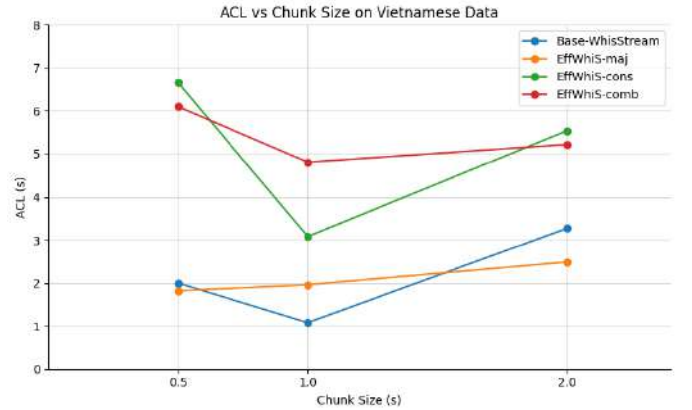


Fig. 5. ACL vs chunk sizes on Vietnamese data

Across both the ESIC and Vietnamese datasets, our proposed streaming strategies demonstrate consistent improvements over the baseline in either recognition accuracy, responsiveness, or both. The majority-vote commit (EffWhiS-maj) significantly enhances latency across chunk sizes, making it suitable for real-time applications, while the segment consistency constraint (EffWhiS-cons) achieves superior accuracy, particularly evident in lower WERs across both corpora. Notably, the combined strategy (EffWhiS-comb) effectively balances these strengths, consistently ranking among the top

performers in both WER and latency. This cross-lingual robustness affirms the generalizability and practicality of our approach, confirming its effectiveness for efficient and accurate streaming ASR beyond a single language or dataset.

VI. CONCLUSION

In this work, we proposed EffWhiS, an efficient approach for real-time transcription based on Whisper Streaming. By introducing two key strategies—majority-vote commit and segment consistency constraint—we address key limitations of the original system, such as delayed output and instability under noisy or complex conditions. Experiments on two real-world datasets (ESIC and Vietnamese meeting recordings) demonstrate that our method achieves better accuracy-latency trade-offs across various chunk sizes, especially maintaining low Word Error Rate even at short segment lengths. Compared to the baseline Whisper Streaming, EffWhiS-comb reduces WER from 10.01% to 8.75% at 0.5s chunk size on Vietnamese data, and from 11.29% to 10.67% on ESIC English data. Moreover, it maintains competitive latency (ACL = 6.1s and 2.05s, respectively) while still preserving high transcription quality. These results suggest that EffWhiS offers a practical and more adaptive alternative for real-time ASR in challenging environments. Notably, EffWhiS proves robust across both English and low-resource languages without requiring model fine-tuning. This highlights its effectiveness as a lightweight, deployable enhancement to Whisper’s inference pipeline. For future work, we aim to integrate dynamic chunk sizing based on signal characteristics, which may further improve responsiveness and stability.

REFERENCES

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [2] G. Klein, “Faster-Whisper: a reimplementation of openai’s whisper using ctranslate2,” <https://github.com/SYSTRAN/faster-whisper>, 2025, accessed: 2025-07-28.
- [3] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri, and M. Almojel, “Automatic speech recognition: Systematic literature review,” *Ieee Access*, vol. 9, pp. 131 858–131 876, 2021.
- [4] A. Bevilacqua, P. Saviano, A. Amirante, and S. P. Romano, “Whispy: Adapting stt whisper models to real-time environments,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.03484>
- [5] R. Wang, Z. Xu, and F. Xiaozhu Lin, “WhisperFlow: speech foundation models in real time,” *arXiv e-prints*, p. arXiv:2412.11272, Dec. 2024.
- [6] H. Wang, G. Hu, G. Lin, W.-Q. Zhang, and J. Li, “Simul-whisper: Attention-guided streaming whisper with truncation detection,” in *Interspeech 2024*, ser. Interspeech. ISCA, Sep. 2024, pp. 4483–4487.
- [7] D. Macháček, R. Dabre, and O. Bojar, “Turning whisper into real-time transcription system,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.14743>
- [8] G. Klein, F. Hernandez, V. Nguyen, and J. Senellart, “The opennmt neural machine translation toolkit: 2020 edition,” in *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 2020, pp. 102–109.
- [9] N. T. Hung, D. V. Khanh, L. Q. Minh, L. X. Duc, N. T. Dung, and P. T. B. Van, “A proposed hybrid chatbot for retrieving information in the domain of digital transformation,” in *Proceedings of the 17th National Conference on Fundamental and Applied Information Technology Research (FAIR)*. Hanoi, Vietnam: Vietnam Academy of Science and Technology, Aug. 2024.
- [10] M. T. Q. Nhu, V. H. Phuc, N. D. Son, N. T. Hung, D. V. Khanh, N. T. Dung, and P. T. B. Van, “A proposed architecture of ai application for virtual conference secretary,” in *Proceedings of the 18th National Conference on Fundamental and Applied Information Technology Research (FAIR)*. Hanoi, Vietnam: Vietnam Academy of Science and Technology, Aug. 2025.
- [11] D. Liu, G. Spanakis, and J. Niehues, “Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection,” in *Interspeech 2020*, 2020, pp. 3620–3624.
- [12] D. Macháček, M. Žilínek, and O. Bojar, “Lost in interpreting: Speech translation from source or interpreter?” in *Interspeech 2021*, 2021, pp. 2376–2380.
- [13] S. Team, “Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier,” 2024.
- [14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [15] A. Ali and S. Renals, “Word error rate estimation for speech recognition: e-wer,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics (ACL), 2018, pp. 20–24.

LAURA: A Legal Agent Using Reasoning and Acting for Vietnamese Law

Van-Khanh Tran* and Van-Hieu Nguyen

Institute of Applied Science and Technology - IAST

University of Information and Communication Technology - ICTU

Thai Nguyen City, Vietnam

{tvkhanh, dtc205480201clc0046}@ictu.edu.vn

Abstract—Accessing legal information in Vietnam is fraught with challenges, primarily due to the sheer volume of legal documents and the complexity of legal language. While Large Language Models (LLMs) present a potential solution, they are often hindered by issues of “hallucination” and outdated knowledge. To overcome these obstacles, this paper introduces LAURA (A Legal Agent Using Reasoning and Acting), a novel agent-based question-answering system tailored for Vietnamese law. LAURA operates on an agentic workflow, leveraging an LLM as its core reasoning engine. The system is built on the ReAct paradigm, which synergizes reasoning with the use of retrieval tools to process legal queries effectively. To enhance accuracy and minimize errors, a custom retrieval module was developed, combining Reciprocal Rank Fusion (RRF) with the BM25 sparse retriever and the multilingual-e5 dense retriever. Furthermore, LAURA integrates Tavily Search to access the latest legal information from external sources in real-time. This integrated approach demonstrates the significant potential of combining agentic reasoning with multi-source retrieval to create an accurate and interpretable legal question-answering system for the Vietnamese language.

Index Terms—Legal question answering, document retrieval, Legal Agent Using Reasoning and Acting

I. INTRODUCTION

In the current era of digitalization and rapid artificial intelligence development, accessing legal information has become more convenient than ever. However, the Vietnamese legal system [1] is characterized by a vast number of documents, including decrees and circulars, written in complex legal language. This complexity creates significant barriers for individuals and businesses without legal expertise, making it difficult for them to search, interpret, and comply with regulations to protect their legitimate rights. Traditionally, obtaining legal information has required consulting with lawyers and legal advisors or conducting manual searches across multiple sources, processes that are both time-consuming and costly. This gap has created an urgent demand for an intelligent legal support system capable of providing accurate and clear answers to legal questions.

The emergence of Large Language Models (LLMs) like LLaMA 3 [2] has opened new opportunities in legal support, given their ability to process vast amounts of text. However, their standalone application is limited by several critical issues, including outdated knowledge, a lack of source

traceability, and a tendency toward “hallucination,” where the model generates incorrect content. To mitigate these problems, approaches like Retrieval-Augmented Generation (RAG) [3] were introduced to enhance accuracy by combining external knowledge sources with the LLM’s internal knowledge.

Nevertheless, even standard RAG systems encounter challenges in the legal domain, such as difficulty in integrating complex legal context and a lack of multi-step reasoning needed for iterative analysis. To address these remaining gaps, this paper introduces LAURA (A Legal Agent Using Reasoning and Acting), an intelligent, agentic question-answering system built upon the ReAct paradigm [4]. Coordinated by the LLaMA 3.3-70B model, LAURA employs an agentic workflow that combines step-by-step reasoning with dynamic tool use and iterative knowledge retrieval. By automating key stages of legal information retrieval, LAURA aims to enhance the transparency and reliability of legal answers, making complex legal knowledge more accessible to individuals and businesses alike.

II. RELATED WORKS

Answering legal questions is a highly challenging task due to the complexity of legal text processing, which involves specialized terminology, complex sentence structures, and strict logical relationships. Large Language Models (LLMs) such as LLaMA 3 [2] and GPT-4 [5] have made significant advancements in Natural Language Processing (NLP), improving not only question answering but also the ability to understand and respond accurately to user intent. However, LLMs still face several limitations, such as outdated knowledge (due to static training data), lack of source traceability, and especially the issue of “hallucination,” where the model generates incorrect or inconsistent content. To address these problems, the RAG approach [3] was introduced, enhancing answer accuracy and reliability by combining external knowledge sources with the LLM’s internal knowledge. To further refine this process, advanced methods like Corrective Retrieval-Augmented Generation (CRAG) [6] were developed. CRAG introduces a self-correction layer that assesses the relevance of retrieved documents. If the information is deemed insufficient or irrelevant, the system can trigger corrective actions, such as rewriting the query or searching external sources to augment the knowledge base. This proactive validation helps ensure

* Corresponding author: tvkhanh@ictu.edu.vn

the context provided to the LLM is of higher quality, directly addressing the issue of grounding.

Nevertheless, even with these advancements, RAG-based systems still encounter three major challenges in legal question answering:

- **Difficulty in integrating legal context:** While RAG can retrieve relevant legal documents, it struggles to seamlessly incorporate them into responses, often producing fragmented, generic, or inapplicable answers.
- **Lack of multi-step reasoning:** Complex legal queries typically require iterative retrieval, analysis, and synthesis from multiple sources. Traditional RAG systems cannot adjust retrieval strategies based on intermediate information, leading to incomplete or inaccurate answers.
- **Scalability and latency issues:** As legal data—such as laws, case law, regulations, and guidelines—continues to grow, querying and ranking relevant documents becomes computationally expensive, slowing down response times. This limits the use of RAG systems in real-time applications like online legal consultations or courtroom assistance.

These challenges drive the integration of autonomous agents into legal QA systems. Such agents enable dynamic decision-making and multi-step reasoning, allowing the system to adapt its strategy to each specific query. This approach improves answer accuracy, reduces latency, and enhances scalability when addressing complex legal questions.

III. DATASETS

A. ALQAC Dataset

The primary dataset used in this study is derived from the Automated Legal Question Answering Competition (ALQAC 2023) [1], which focuses on two core tasks related to Vietnamese legal documents: legal information retrieval and legal question answering. This makes it a suitable benchmark for evaluating our agent-based system.

The dataset consists of 17 laws and a total of 2,249 articles. Table I summarizes the legal content, while Table II provides statistics for the question sets used in the competition.

TABLE I
VIETNAMESE LAWS IN THE ALQAC DATASET

Item	Quantity
Number of Laws	17
Number of Articles	2,249

Table II presents detailed statistics of the question sets used in different evaluation splits of the ALQAC dataset, including the development test set, the public test set, and the private test set. Each split contains a comparable number of questions, with variations in question length measured by token count.

Specifically, the number of questions ranges from 100 to 110, while the maximum number of tokens per question varies from 60 (Dev test) to 90 (Public test). The average token length also gradually increases from 24.91 to 27.91, indicating that questions in the Private test set tend to be slightly longer

TABLE II
STATISTICS OF ALQAC QUESTION SETS

Split	Dev test	Public Test	Private Test
Number of Questions	100	100	110
Max Tokens	60	90	87
Min Tokens	6	4	10
Avg. Tokens	24.91	25.91	27.91

and potentially more complex. These variations are important for assessing the robustness and generalization ability of the question-answering system.

B. Datasets for Fine-tuning

- **ALQAC 2022:** Contains legal content from 4 laws, totaling 1,337 articles and 520 question samples.
- **Zalo AI Challenge:** A large-scale dataset with 3,271 laws, 61,425 articles, and 3,196 QA samples.

TABLE III
STATISTICS OF FINE-TUNING DATASETS

Dataset	ALQAC 2022	Zalo
Number of Laws	4	3,271
Number of Articles	1,337	61,425
Number of Questions	520	3,196

Data Augmentation: The first step in the data processing pipeline is to apply the BM25 [7] algorithm, a retrieval method based on a vector embedding model, to extract documents relevant to each question. After retrieval, each document is labeled according to the following rules:

- **Negative label (label = 0):** The document is irrelevant or not suitable for the question.
- **Positive label (label = 1):** The document is relevant to the question.

From this process, a total of 4,879 valid samples were collected.

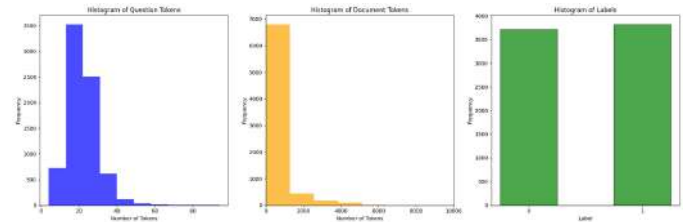


Fig. 1. Visualization of the fine-tuning dataset

Figure 1 illustrates the distribution of token lengths in questions, legal documents, and label balance:

- **Questions (blue, left):** Average: 21.63 tokens; Min: 5; Max: 94. Most questions range from 10–30 tokens, peaking around 20. The distribution is left-skewed.
- **Documents (orange, center):** Average: 236.79 tokens; Min: 17; Max: 500. Token counts are more spread out, mostly between 100–300 tokens, with a right-skewed distribution.

- **Labels (green, right):** Positive and negative labels are balanced (2,400 each), which helps prevent model bias during training.

IV. METHODS

A. The LAURA Agentic Framework

The proposed system, LAURA (A Legal Agent Using Reasoning and Acting), is an agentic framework designed to navigate the complexities of the Vietnamese legal system. At its core, an LLM—specifically LLaMA 3.3-70B—serves as the central reasoning engine, coordinating a suite of advanced tools to deliver responses that are accurate, reliable, and interpretable. The framework is built upon the ReAct paradigm [4], which allows the system to simulate the iterative workflow of a legal expert by tightly integrating step-by-step reasoning with concrete actions. The overall workflow of the LAURA agent is illustrated in Figure 2.

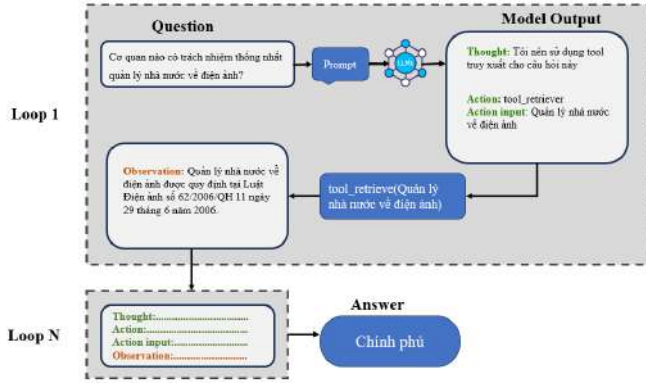


Fig. 2. The LAURA Agentic Framework for the legal QnA system.

The LAURA workflow operates on a loop of three key states: **Thought**, **Action**, and **Observation**. At each step t , the process is defined as follows:

- 1) **Thought (Reasoning):** The LLM model analyzes the current state $\mathcal{S}_{t-1}$ to generate an internal thought process and formulate a plan for the next action:

$$(\text{Thought}_t, \text{ActionPlan}_t) = \text{LLM}_{\text{reason}}(\mathcal{S}_{t-1}) \quad (1)$$

- 2) **Action:** The planned action is executed, typically by invoking an external tool such as a legal retriever or web search.
- 3) **Observation:** The outcome of the action is recorded and integrated into the agent's state:

$$\text{Observation}_t = \text{Execute}(\text{Action}_t) \quad (2)$$

- 4) **State Update:** The system's state is updated with the new information gathered from the cycle:

$$\mathcal{S}_t = \mathcal{S}_{t-1} \cup \{(\text{Thought}_t, \text{Action}_t, \text{Observation}_t)\} \quad (3)$$

This *Thought-Action-Observation* loop repeats until the agent determines that sufficient evidence has been collected, at which point a final answer is generated based on the complete reasoning history. This iterative process enables the

LAURA agent to maintain a coherent chain of legal reasoning and produce well-supported responses that are aligned with authoritative legal documents.

B. The Agent's Toolkit

The **Action** state in the LAURA workflow is executed by invoking a suite of specialized tools designed to gather and process legal information. A better retrieval tool is necessary to overcome the inherent limitations of standard search methods when dealing with complex legal texts, while fine-tuning a dense vector embedding teaches the model to understand relevance in this specialized domain better.

1) *Web Search Tool:* The system integrates Tavily Search¹, a real-time search engine, to augment its capabilities. This tool allows the agent to retrieve up-to-date information from reliable external sources like government portals and academic research, ensuring the context used for answers is both relevant and current.

2) *Hybrid Retrieval Tool:* To overcome the limitations of standard search methods in complex legal texts, a custom retrieval engine was developed. Its key components are:

- *Fine-tuned Dense Embeddings:* To improve the model's ability to discern relevance, the dense vector embedding model is fine-tuned using Contrastive Loss. This process adjusts the Euclidean distance between vector representations of question-document pairs, minimizing the distance for relevant pairs (label=1) and maximizing it for irrelevant pairs (label=0). This tailors the model to the specific nuances of the legal domain, ensuring more meaningful semantic representations
- *Hybrid Architecture:* The engine employs a two-pronged retrieval strategy:
 - **Vector Transformation:** Legal documents and queries are converted into both dense vectors using models like BERT [8] for semantic understanding and sparse vectors using lexical models like BM25, which effectively ranks documents by balancing term frequency and document length.
 - **Result Fusion:** Results from both the dense (semantic) and sparse (lexical) retrievers are combined using the Reciprocal Rank Fusion (RRF) [9] method. RRF assigns a score to each document based on its rank across the different result lists, prioritizing documents that rank highly in at least one list to improve overall accuracy.

3) *RAG Tool:* The LAURA framework incorporates the functionality of Retrieval-Augmented Generation (RAG), an approach designed to enhance answer accuracy and reliability by combining external knowledge sources with an LLM's internal knowledge. Within the agentic workflow, RAG is not treated as a standalone system but as a foundational tool that the LAURA agent can execute during its **Action** state. When the LAURA agent calls its Hybrid Retrieval Tool, it is effectively performing a RAG-like action: retrieving relevant

¹<https://www.tavily.com/>

TABLE IV
RETRIEVAL PERFORMANCE ON THE DEVELOPMENT SET

	top 1			top 2			top 3			top 4			top 5		
	P	R	F2	P	R	F2	P	R	F2	P	R	F2	P	R	F2
TF-IDF	0.86	0.86	0.86	0.65	0.92	0.83	0.47	0.94	0.76	0.3	0.95	0.69	0.3	0.96	0.64
BM25	0.9	0.9	0.9	0.68	0.95	0.86	0.5	0.95	0.78	0.38	0.95	0.71	0.3	0.95	0.65
vietnamese-sbert	0.74	0.74	0.74	0.58	0.83	0.75	0.44	0.88	0.7	0.33	0.89	0.65	0.27	0.91	0.61
vietnamese-sbert-hybrid	0.76	0.76	0.76	0.66	0.95	0.85	0.5	0.96	0.78	0.36	0.97	0.7	0.3	0.98	0.65
vietnamese-sbert-ft	0.61	0.61	0.61	0.49	0.73	0.65	0.34	0.75	0.59	0.27	0.76	0.54	0.22	0.82	0.52
vietnamese-sbert-ft-hybrid	0.6	0.6	0.6	0.58	0.92	0.8	0.47	0.95	0.76	0.36	0.96	0.7	0.38	0.97	0.63
multilingual-e5-large	0.93	0.93	0.93	0.7	0.95	0.8	0.54	0.97	0.8	0.42	0.97	0.74	0.34	0.97	0.68
multilingual-e5-large-hybrid	0.93	0.93	0.93	0.72	0.96	0.88	0.55	0.98	0.81	0.41	0.99	0.75	0.32	0.99	0.68
multilingual-e5-large-ft	0.49	0.49	0.49	0.4	0.63	0.56	0.32	0.68	0.54	0.27	0.72	0.51	0.21	0.73	0.48
multilingual-e5-large-ft-hybrid	0.49	0.49	0.49	0.59	0.95	0.83	0.45	0.96	0.75	0.37	0.97	0.7	0.3	0.97	0.64

context from its knowledge base to ground the final generated answer.

4) *Corrective Retrieval (CRAG) Tool*: To enhance the robustness of information retrieval, the LAURA framework can incorporate the principles of Corrective Retrieval-Augmented Generation (CRAG). Unlike standard RAG, which directly uses retrieved documents, CRAG introduces a self-correction mechanism to evaluate and improve the quality of the retrieved information before it is used for answer generation. This makes it a powerful and specialized tool that the LAURA agent can invoke for complex or ambiguous queries.

The CRAG workflow is built around a lightweight retrieval evaluator that assesses the relevance of retrieved documents for a given query. Based on the confidence score from the evaluator, one of three actions is triggered:

- **Correct**: If the retrieved documents are highly relevant, they are passed on to the knowledge refinement step.
- **Incorrect**: If the documents are irrelevant, they are discarded. CRAG then performs a corrective action by triggering a web search to find more accurate information.
- **Ambiguous**: If the relevance is uncertain, a web search is also initiated to supplement the retrieved information.

Within the LAURA framework, this entire CRAG process functions as a single, powerful **Action** that the agent can choose to execute. For instance, if the agent’s initial **Thought** determines that a query is high-stakes or likely to yield irrelevant results from the static database, it can invoke the CRAG tool instead of a simple retriever. The **Observation** returned to the LAURA agent would then be a high-quality, pre-vetted set of information, either from the internal database or corrected via a web search. By integrating CRAG as a tool, the LAURA agent can dynamically decide between a fast, standard retrieval and a more robust, self-correcting retrieval, directly addressing the limitations of traditional RAG systems, such as their inability to adjust strategies based on intermediate information.

V. EXPERIMENTAL SETUPS

Our methods are implemented using the Python programming language, leveraging open-source and widely-used libraries in deep learning and natural language processing, such as Sentence Transformer and Huggingface. Moreover, we integrate the Together API to efficiently and flexibly access the capabilities of large language models (LLMs), specifically the

LLaMA 3.3–70B model. The Tavily API is also employed to support querying and aggregating information from various sources across the Internet. Furthermore, we utilize LangGraph to construct interactive, multi-step language processing workflows with enhanced control and traceability, and LangChain to build modular chains that integrate LLMs with external tools and data sources, enhancing context-aware reasoning and response generation. Our experiments are conducted on Kaggle and Google Colab platforms. Specifically, Kaggle provides an NVIDIA Tesla P100 GPU with 16GB of memory for up to 30 hours of free usage, along with an Intel Xeon CPU running at 2.00 GHz.

To evaluate and compare the performance of the legal retrieval task of traditional lexical retrieval models against modern dense retrieval models on the Vietnamese legal document dataset, the experiment used two lexical models, TF-IDF [10] and BM25, and two dense models, vietnamese-sbert² and multilingual-e5-large³. Performance was measured by retrieving the top 1 to 5 most relevant documents and calculating precision, recall, and F2 scores on the development test set and the private test set. Results are shown in Table IV and V.

To evaluate the end-to-end question-answering performance of different LAURA configurations and compare them against non-agentic baselines, we conducted an ablation study of the LAURA Agent’s tools. The experiment tested four LAURA configurations on the dev test set and the public test set: LAURA using only RAG tool (LAURA-RT), LAURA using CRAG tool (LAURA-CT), LAURA using only its internal hybrid retrieval tool (LAURA-IT), and LAURA using both the hybrid retrieval tool and the web search tool (LAURA-2T). Performance was measured by the total number of correct answers across True/False, Multiple-Choice, and Free Text questions.

VI. RESULTS

A. Legal Retrieval Task

This section of the study undertakes a comprehensive evaluation to identify the most effective information retrieval method to serve as a core tool within the LAURA framework. The experiments systematically compare the performance of

²<https://huggingface.co/keepitreal/vietnamese-sbert>

³<https://huggingface.co/intfloat/multilingual-e5-large>

TABLE V
RETRIEVAL PERFORMANCE ON THE PRIVATE TEST SET

	top 1			top 2			top 3			top 4			top 5		
	P	R	F2	P	R	F2	P	R	F2	P	R	F2	P	R	F2
TF-IDF	0.7	0.68	0.68	0.52	0.76	0.67	0.4	0.82	0.66	0.32	0.87	0.63	0.37	0.88	0.58
BM25	0.7	0.66	0.67	0.55	0.76	0.68	0.41	0.81	0.66	0.34	0.87	0.64	0.28	0.87	0.64
vietnamese-sbert	0.6	0.58	0.58	0.45	0.67	0.59	0.34	0.71	0.56	0.26	0.72	0.51	0.21	0.72	0.47
vietnamese-sbert-hybrid	0.56	0.54	0.55	0.5	0.79	0.69	0.39	0.83	0.65	0.32	0.85	0.61	0.26	0.87	0.57
vietnamese-sbert-ft	0.46	0.44	0.45	0.33	0.51	0.45	0.25	0.55	0.43	0.2	0.58	0.4	0.16	0.59	0.37
vietnamese-sbert-ft-hybrid	0.41	0.39	0.39	0.48	0.78	0.67	0.36	0.81	0.63	0.29	0.83	0.58	0.24	0.84	0.54
multilingual-e5-large	0.82	0.78	0.79	0.6	0.83	0.75	0.48	0.87	0.72	0.37	0.88	0.66	0.3	0.89	0.61
multilingual-e5-large-hybrid	0.82	0.78	0.79	0.63	0.88	0.8	0.47	0.9	0.74	0.37	0.9	0.67	0.3	0.91	0.62
multilingual-e5-large-ft	0.5	0.49	0.49	0.38	0.57	0.5	0.31	0.61	0.49	0.25	0.66	0.47	0.2	0.67	0.44
multilingual-e5-large-ft-hybrid	0.5	0.49	0.49	0.5	0.8	0.71	0.4	0.85	0.67	0.32	0.87	0.63	0.27	0.87	0.57

traditional lexical models, modern dense models, and a hybrid approach across both a development set and a private test set from the ALQAC 2023 competition.

Table IV shows the evaluation of various document retrieval models using top-1 to top-5 document extraction. *Lexical Models (TF-IDF vs. BM25)*: When comparing the two traditional, keyword-based models, BM25 consistently demonstrates a performance edge over TF-IDF, achieving a higher precision of 0.90 on the development set. This suggests that BM25’s probabilistic model, which better accounts for term frequency and document length, is more effective for ranking legal documents than the simpler TF-IDF weighting scheme. *Dense Models (Semantic Retrieval)*: The experiments highlight a significant performance gap between the two dense models tested. The multilingual-e5-large model is shown to be superior to both the lexical models and the vietnamese-sbert model. The primary reason for this success is the model’s larger input size. With a capacity of 512 tokens, multilingual-e5-large can process more of the complex context found in legal documents compared to vietnamese-sbert (256 tokens), an advantage that holds despite the latter’s specific design for the Vietnamese language.

The most critical finding from the retrieval task is the validated efficacy of the hybrid approach. This approach combines the best-performing lexical model (BM25) with the best-performing dense model (multilingual-e5-large) and fuses their ranked results using the Reciprocal Rank Fusion (RRF) method. This hybrid configuration consistently achieved the best results across both datasets. On the development set, it yielded the highest F2 and precision scores (both 0.93) and the highest recall (0.99). On the more challenging private test set, it again attained the top performance with an F2 score of 0.80 and a recall of 0.91. The success of the hybrid model demonstrates a powerful synergy. It leverages the strength of BM25 in matching specific legal keywords and terminology while simultaneously capitalizing on the deep contextual understanding of the multilingual-e5-large model. This fusion creates a retrieval system that is both lexically precise and semantically robust.

When benchmarked against other systems in the ALQAC 2023 [1] competition, the proposed hybrid retriever ranked fourth overall with an F2 score of 0.80, as shown in Table VI. Despite its fourth-place overall ranking, the retriever demonstrated a crucial advantage by achieving the highest recall

TABLE VI
COMPARATIVE RANKINGS OF ALL PARTICIPATING TEAMS IN TASK 1 ON THE PRIVATE TEST SET.

Team	Precision	Recall	F2
Neco	0.9	0.86	0.86
NOWJI	0.86	0.86	0.83
Sonic	0.83	0.83	0.81
LAURA’s Retriever	0.63	0.88	0.8
AI EPU	0.7	0.68	0.68
ST	0.24	0.72	0.52

score (0.88) among all competitors, notably outperforming the first-place team in this specific metric. This finding is highly significant for a legal domain application. High recall ensures that the system is less likely to miss a potentially relevant legal document, which is critical for providing comprehensive and reliable answers. This makes the hybrid retriever the optimal choice for the LAURA framework, as it prioritizes completeness in its evidence-gathering stage.

B. Legal Question Answering Task

TABLE VII
STATISTICS OF THE NUMBER OF CORRECT ANSWERS FOR EACH METHOD ON THE DEV TEST SET.

Question Type	LAURA-RT	LAURA-CT	LAURA-IT	LAURA-2T
True/False	46	45	45	45
Multi-Choice	37	37	37	38
Free Text	9	10	7	10
Total	92	92	89	93

This section evaluates the end-to-end question-answering performance of the system by comparing four distinct LAURA configurations. The objective is to assess how different tool integrations impact accuracy across various question types and to benchmark the framework against other systems. The study tested the following four configurations of the LAURA framework:

- LAURA-RT: The agent using only a standard RAG tool.
- LAURA-CT: The agent using a Corrective RAG (CRAG) tool.
- LAURA-IT: The agent using only its internal, custom Hybrid Retrieval Tool.
- LAURA-2T: The agent using both the Hybrid Retrieval Tool and the Web Search Tool.

Table VII breaks down the performance of each configuration on the dev test set, revealing nuanced strengths and weaknesses.

- *Overall Performance:* The LAURA-2T configuration, which leverages both internal and external knowledge sources, achieved the highest overall score, with 93 out of 100 correct answers. This outperforms all other configurations, including the LAURA-RT and LAURA-CT models, which both scored 92.
- *Performance by Question Type:* For True/False questions, the LAURA-RT configuration performed best, correctly answering 46 out of 50 questions. For more complex Multiple-Choice and Free Text questions, the LAURA-2T configuration demonstrated superior performance.

The strong performance of LAURA-2T on complex questions suggests that its ability to dynamically integrate real-time web search with its internal knowledge base is a key advantage. While a direct retrieval (LAURA-RT) is highly effective for fact-based verification (True/False), answering multi-faceted or open-ended questions benefits significantly from the broader context provided by the web search tool. The LAURA-1T configuration, relying solely on its internal retriever, had the lowest score (89), underscoring the performance gains from more advanced RAG techniques and external tool integration.

TABLE VIII
COMPARISON OF RESULTS WITH OTHER TEAMS ON THE PUBLIC TEST SET
FOR QUESTION ANSWERING

Team / Method	Accuracy
Sonic	29%
NOWJ1	67%
NeCo	74%
LAURA-1T	84%
LAURA-2T	87%
LAURA-RT	92%
LAURA-CT	92%

Table VIII provides a comparison of the LAURA configurations against other teams from the ALQAC 2023 competition on the public test set. In this context, both LAURA-RT and LAURA-CT achieved the highest overall accuracy at 92%, significantly outperforming competing teams like NeCo (74%), NOWJ1 (67%), and Sonic (29%). The LAURA-2T configuration scored 87%, and LAURA-1T scored 84%. There is a notable discrepancy in the performance ranking between the two test sets. While LAURA-2T was the top performer on the dev test set (Table VII), it was surpassed by the LAURA-RT and LAURA-CT configurations on the public test set (Table VIII). This suggests that the performance of each toolchain is data-dependent. The public test set may contain questions that are better aligned with the more direct retrieval-and-correction pipelines of RAG and CRAG. Conversely, the dev test set appears to have more questions where the flexible, multi-tool reasoning of LAURA-2T provides a distinct advantage. In conclusion, the results validate the overall efficacy of the LAURA agentic framework, as all its configurations outperformed external competitors. The analysis also highlights that the optimal tool configuration within the framework can vary depending on the specific nature of the legal queries being addressed.

CONCLUSION

This study demonstrates the effectiveness of the LAURA agent framework for Vietnamese legal question answering, successfully bridging traditional information retrieval methods with modern AI techniques to develop a practical and intelligent legal support system. By integrating a hybrid retrieval engine—combining a BM25 sparse retriever with a dense encoder trained using contrastive learning—with an agentic workflow that leverages step-by-step reasoning, action planning, and flexible interaction with external knowledge sources, the system significantly enhances both interpretability and retrieval performance.

Despite these promising results, the system faces several limitations that guide future research directions. These include slower processing speeds when handling long legal documents, occasional errors in complex reasoning tasks, and limited adaptability to newly issued legal regulations. To address these issues, future work will focus on several key areas: (i) optimizing retrieval for long and structurally complex legal texts; (ii) integrating larger-scale legal databases and developing more effective knowledge representation methods; (iii) leveraging domain-specific Vietnamese large language models to improve contextual understanding; deploying the system on real-world platforms such as online legal consultation portals; and (v) improving user experience through intelligent legal assistant interfaces. In summary, this research not only advances the state of the art in Vietnamese legal question answering but also contributes to the broader goals of legal transparency and greater public access to justice.

REFERENCES

- [1] C. Nguyen and e. a. Luu, Son T., “A summary of the alqac 2023 competition,” in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6.
- [2] A. Grattafiori and A. D. et al., “The llama 3 herd of models,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [3] P. Lewis and e. a. Ethan Perez, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” 2021. [Online]. Available: <https://arxiv.org/abs/2005.11401>
- [4] S. Yao and e. a. Jeffrey Zhao, “React: Synergizing reasoning and acting in language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2210.03629>
- [5] OpenAI, J. Achiam, and e. a. Steven Adler, “Gpt-4 technical report,” 2024. [Online]. Available: <https://arxiv.org/abs/2303.08774>
- [6] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, “Corrective retrieval augmented generation,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.15884>
- [7] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: Bm25 and beyond,” *Foundations and Trends in Information Retrieval*, vol. 3, pp. 333–389, 01 2009.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [9] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, “Reciprocal rank fusion outperforms condorcet and individual rank learning methods,” in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’09. New York, NY, USA: Association for Computing Machinery, 2009, p. 758–759. [Online]. Available: <https://doi.org/10.1145/1571941.1572114>
- [10] L. Cheng, Y. Yang, K. Zhao, and Z. Gao, “Research and improvement of tf-idf algorithm based on information theory,” in *The 8th International Conference on Computer Engineering and Networks (CENet2018)*, Q. Liu, M. Misir, X. Wang, and W. Liu, Eds. Cham: Springer International Publishing, 2020, pp. 608–616.

Toward Effective Domain Adaptive Continual Pre-training for Vietnamese Large Language Models

Tien-Khoi Nguyen¹, Xuan-Son Quan¹, Hong-Viet Tran^{1*}, Van-Vinh Nguyen¹, Phuong-Thai Nguyen¹

¹University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

22022658@vnu.edu.vn, quanxuanson2004@gmail.com, thviet@vnu.edu.vn, vinhnv@vnu.edu.vn, thainpv@vnu.edu.vn

* The Corresponding Author

Abstract—Large language models (LLMs) have achieved several outstanding results on natural language related tasks. By continuing training the LLMs on specific domains can eventually boost the performance on some specific tasks. However, continual training the large language model still needs huge amount of data and there are some domains that is limited in high quality text to training or hard to get. Our approach can address this by augmenting the original domain text corpus with reasoning model to increase the amount of data for continual training, furthermore, we propose a training method to keep the model’s ability while gaining new knowledge. The result shows that after training our method improves the result by 2% on question answering test set and 4 to 6% on domain specific MCQ test set of VMLU [1]

Key words—Continual Learning, Continual Pretraining, Pretraining Data, LoRA, Synthetic Data, Continual LoRA

I. INTRODUCTION

While Large Language Models (LLMs) excel at general tasks through large-scale pre-training [2], their effectiveness in specialized domains such as medicine or law remains limited due to insufficient domain knowledge [3]. This limitation is more severe for low-resource languages like Vietnamese, where the lack of high-quality domain text hinders continual pre-training—a key strategy for incremental adaptation. Moreover, direct fine-tuning often causes catastrophic forgetting [4], erasing general knowledge.

To overcome these challenges, we propose a framework for domain-adaptive continual pre-training of Vietnamese LLMs. Our method addresses data scarcity through intelligent data augmentation and preserves prior knowledge using advanced continual learning. The main components of our approach are as follows:

- A two-phase data pipeline that generates synthetic domain data using a smaller reasoning model to enhance limited corpora.
- The adoption of C-LoRA to enable efficient domain adaptation while mitigating catastrophic forgetting through orthogonality constraints.

II. RELATED WORKS

Our research builds on recent advancements in continual learning, domain adaptation, and synthetic data generation, applying them to the unique challenges of

developing Large Language Models (LLMs) for low-resource Vietnamese.

A. Continual Learning for Large Language Models

Continual learning (CL) mitigates catastrophic forgetting, where an LLM loses performance on prior tasks after fine-tuning on new data. Existing strategies include data replay [5], regularization [6], and parameter isolation. Among these, Low-Rank Adaptation (LoRA) [7] is efficient, introducing small trainable low-rank matrices while freezing most parameters. We adopt C-LoRA [4], which extends LoRA by enforcing orthogonality constraints, ensuring new knowledge is learned in non-interfering subspaces and effectively preserving prior knowledge.

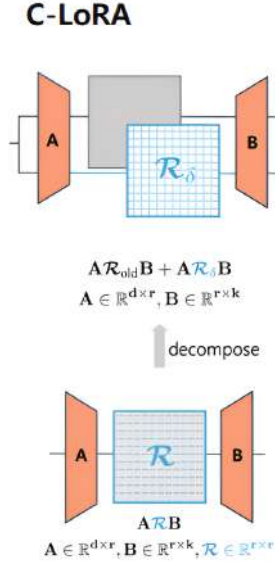


Fig. 1. Overview of C-LoRA’s orthogonality mechanism.

B. Domain Adaptation and Synthetic Data

Adapting LLMs to specialized domains is critical but computationally expensive via full fine-tuning. Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA provide a more viable alternative. We align with recent work demonstrating that continual pre-training is more effective when raw text is converted into structured, instruction-style formats [8].

To overcome data scarcity in specialized domains, synthetic data generation has emerged as a promising solution. This typically involves using a capable model to generate new training examples from a limited seed set, a technique demonstrated by methods like self-instruction [9]. Our framework leverages a smaller reasoning model to enrich a sparse domain corpus with high-quality, structured synthetic data for effective continual pre-training.

C. Challenges for Vietnamese LLMs

Developing LLMs for Vietnamese is challenging due to a smaller volume of high-quality text corpora and a scarcity of domain-specific benchmarks compared to high-resource languages. Consequently, existing Vietnamese LLMs often struggle with specialized tasks [10], [11]. Our work proposes a practical framework to build specialized Vietnamese LLMs that are robust and knowledgeable, even when high-quality domain data is limited.

III. METHODOLOGY

A. Synthetic-data augmentation for pretraining

Large Language Models (LLMs) have demonstrated remarkable capabilities in addressing open-domain tasks. However, their deployment in specific domains remains challenging due to a lack of domain-specific knowledge, limited capacity to utilize such knowledge effectively, and inadequate adaptation to domain-specific data formats. To overcome these limitations, a growing body of research has explored *continual pre-training* as a strategy to adapt LLMs incrementally to new domains.

The research like [8] draws inspiration from human learning behavior by transforming raw corpora into reading comprehension datasets. Their study reveals that while training on raw domain-specific data improves factual knowledge, it can significantly degrade the model’s performance on question-answering tasks, suggesting a need for carefully curated training inputs during continual adaptation.

Our Approach: To address these challenges, we propose a two-phase data processing pipeline that leverages both machine learning-based and heuristic-based methods for data synthesis and filtering. Our approach involves generating high-quality synthetic data from raw domain-specific corpora using the reasoning ability of small-scale LLMs, followed by perplexity-based filtering using the target LLM to ensure training efficiency and stability [3].

Phase 1: Synthetic Data Generation and Filtering

In the first phase, we begin with a pre-processing step using Data-Juicer [12] to perform text normalization (e.g., UTF-8 encoding), removes personal information, filters null samples, and eliminates duplicates. This ensures a clean and consistent dataset for downstream processing.

Next, we utilize recent advancements in small language models, which have shown strong reasoning capabilities comparable to larger models, which capable of generating richer content [13]. The model is asked to analyze each input document to:

- Identify its genre and infer the likely target audience [13];
- Extract key entities relevant to the domain and audience [14];
- Rewrite the document to better suit the identified genre and audience, emphasizing these entities.

Phase 2: Perplexity-based Filtering

In the second phase, we augment our dataset with additional educational-domain documents sourced from CommonCrawl. To improve training efficiency and prevent the model from struggling with low-quality or overly complex inputs, we evaluate each sample using the perplexity score computed by the target LLM. Samples with excessively high perplexity are filtered out, as they are less likely to contribute effectively to model learning and may degrade overall performance.

TABLE I
SAMPLE DISTRIBUTION FROM VNU PAGES AND WEB SOURCES.

Sources	Tokens (Million)	Samples
UET	6.8	6,847
HUS	7.3	8,131
UEB	24	11,821
IS	2.3	1,781
UEd	3.5	5,274
NXB	0.38	975
VNU	9	5,000
ITI	1.4	1700
Edu docs	50.4	52,148
Total	95.1	99,094

TABLE II
ABBREVIATIONS

Abbreviation	Full Name
UET	University of Engineering and Technology
HUS	University of Science
UEB	University of Economics and Business
IS	International School
UEd	University of Education
NXB	Publishing House
VNU	Vietnam National University (General)
ITI	Institute of Information Technology

B. Continual Pretraining with C-LoRA

We adopt the continual learning method from C-LoRA [4] to address catastrophic forgetting, a critical challenge where models degrade performance on previously learned tasks when acquiring new knowledge. C-LoRA achieves this by enforcing orthogonality between the current learning direction and the accumulated knowledge from past tasks, all within a single shared LoRA adapter. This approach enables efficient continual pretraining on new domains without compromising previously acquired abilities. In our

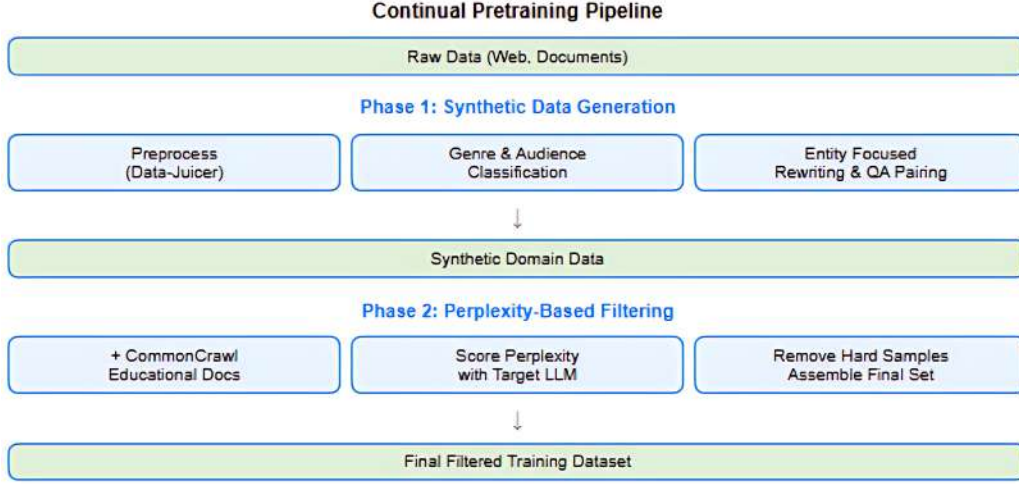


Fig. 2. Our pipeline in our proposed method

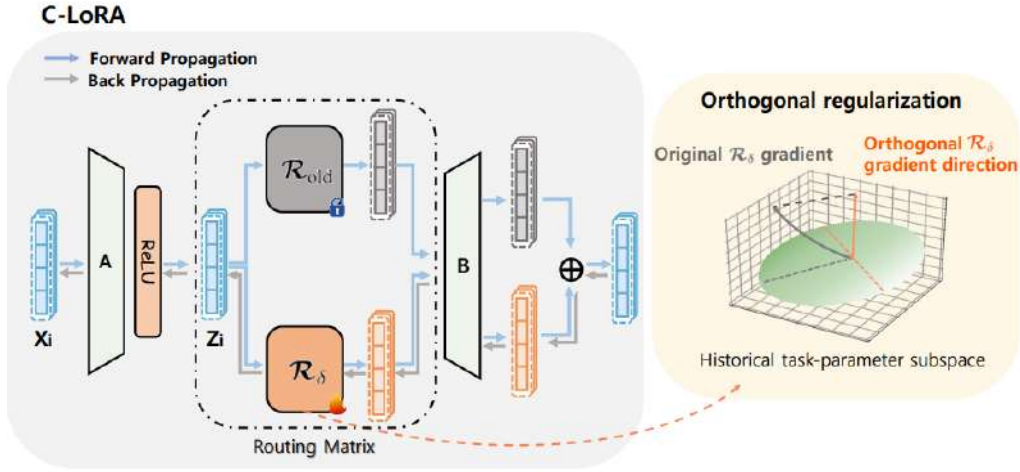


Fig. 3. The C-LoRA architecture mitigates catastrophic forgetting by decoupling R to constrain updates within the parameter space of previous tasks and enforcing orthogonality between R updates and the low-rank parameter space of past tasks.

implementation of C-LoRA, we modify the standard LoRA layer. For each linear layer W_0 (e.g., query, key, value projections in attention), we introduce trainable low-rank matrices $A \in \mathbb{R}^{d_{in} \times r}$ and $B \in \mathbb{R}^{r \times d_{out}}$, where r is the LoRA rank. Additionally, we introduce an internal rotation matrix, $R_{\delta} \in \mathbb{R}^{r \times r}$, and maintain an accumulated rotation matrix, $R_{old} \in \mathbb{R}^{r \times r}$, initialized as an identity matrix. The effective weight update for the forward pass becomes:

$$W_{eff} = W_0 + \Delta W = W_0 + \frac{\alpha}{r} B(R_{old} + R_{\delta})A \quad (1)$$

where α is the LoRA scaling factor. Crucially, R_{old} is detached during training, only accumulating updates between tasks. A and B are treated as task-specific parameters. For each new learning task, these matrices are re-initialized and trained from scratch. The continual knowledge is not stored within A and B across tasks, but is instead exclusively consolidated within the accumulated rotation matrix, R_{old} .

1) Training Procedure and Orthogonality Constraint: Our training process for each new domain involves learning A , B , and R_{δ} while enforcing an orthogonality constraint. After each task $t - 1$ is completed, the learned R_{δ} is added to R_{old} for the next task, and R_{δ} is reset to zeros for task t . The core mechanism of C-LoRA is to ensure that the newly learned updates are "orthogonal" to past knowledge, preventing interference.

The overall training objective for the current task t is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{LM} + \lambda_{ortho} \cdot \mathcal{L}_{ortho} \quad (2)$$

Here, $\mathcal{L}_{LM}$ is the standard causal language modeling loss (e.g., cross-entropy), and λ_{ortho} is a hyperparameter controlling the strength of the orthogonality regularization.

The orthogonality loss, $\mathcal{L}_{ortho}$, is specifically designed to enforce a relationship between the current task's rotation update (R_{δ}) and the input-side LoRA

matrix from the previous task (A_{prev}). This is formulated as the Frobenius norm of their product:

$$\mathcal{L}_{\text{ortho}} = \|R_{\text{delta}} \cdot A_{\text{prev}}^T\|_F^2 \quad (3)$$

where A_{prev} represents the A matrix learned from the immediately preceding task. Minimizing this norm encourages the columns of A_{prev} to be in the null space of R_{delta} . This means the new rotation R_{delta} has minimal effect on the directions captured by the previous adapter, thus preserving the knowledge encoded in the A_{prev} to B_{prev} pathway.

IV. EXPERIMENTAL RESULT

A. Data processing details and result

1) *Phase 1:* For synthetic data generation, we use Qwen3-4B as the generator model. The model is inferred through vLLM with reasoning mode enabled to enhance data diversity. We evaluated multiple Qwen3 variants for this task. Smaller models (0.6B and 1.7B) produced format-consistent but instruction-weak outputs, often limited to paraphrasing the original texts. In contrast, the larger 8B model generated more diverse content but frequently failed to adhere to the required format, complicating subsequent data processing. Therefore, we select Qwen3-4B as the optimal balance between instruction following, format correctness, and content diversity. After we complete phase one of pipeline, the distribution is shown:

TABLE III
NEW DISTRIBUTION FROM VNU PAGES AND WEB SOURCES
AFTER PHASE 1.

Sources	Tokens (Million)	Samples
VNU ¹	44.7	29,964
VNU (Gen.) ²	5.1	13,230
Edu docs ³	36.2	52,148
Total	100.2	95,342

2) *Phase 2:* To improve data quality and training stability, we implemented a two-stage filtering pipeline.

First, we established data-driven difficulty thresholds. We profiled the perplexity distribution of a high-quality reference corpus (Vietnamese Wikipedia [15]) using our target model. This analysis revealed that 96% of the content had a perplexity score below 25 in which the range from 0 to 10 account for 46% and the range from 10 to 25 is 47%. Based on this, we defined three difficulty categories: 'easy' (perplexity < 10), 'medium' (10 to 25), and 'hard' (perplexity > 25).

Second, we trained a lightweight classifier to apply this filter at scale. A classifier trained directly on the Wikipedia distribution would suffer from severe class imbalance, as only 4% of samples are 'hard'. To address this, we constructed a balanced training set for a FastText classifier. We sourced 'easy' and

'medium' samples from the labeled Wikipedia corpus and supplemented them with 'hard' samples sourced from a subset of the lower-quality Fineweb dataset [16]. All samples were labeled using the aforementioned perplexity thresholds computed by the target LLM. After training on this balanced dataset, the classifier was used to predict difficulty labels for our entire corpus, and all samples classified as 'hard' were subsequently removed. The final distribution is shown:

TABLE IV
NEW DISTRIBUTION FROM VNU PAGES AND WEB SOURCES
AFTER PHASE 2.

Sources	Tokens (Million)	Samples
VNU ¹	44.6	29,811
VNU (Gen.) ²	5.1	13,230
Edu docs ³	36.2	36,297
Total	85.9	79,338

As we can see, nearly third of those data taken from education documents on the internet are filtered, those data from internet contains more outliers, while documents from VNU universities often more well structure so only a few documents are filtered, also because we left out a lot of VNU document using heuristic rules.

B. Training details

1) *Hyperparameters and Base Model:* For our experiments, we use the Qwen/Qwen3-0.6B model as the base LLM. The C-LoRA hyperparameters are set as follows: LoRA rank $r = 8$, LoRA scaling factor $\alpha = 16$, and orthogonality regularization strength $\lambda_{\text{ortho}} = 10$. We observe that, when applying C-LoRA to LLMs, the rank can be much smaller than in the original C-LoRA experiments on image classification models. Consequently, the orthogonal loss is relatively weaker compared to the language modeling loss, requiring a stronger regularization term. The model is trained using 4-bit quantization with compute type set to `torch.bfloat16`. These settings balance performance and computational efficiency during continual pre-training. For fair comparison, the standard LoRA configuration is aligned with the same hyperparameters to highlight the advantages of C-LoRA.

2) *Result of C-LoRA training on medical datasets and literature datasets:* We tested our method of C-LoRA training on 3 different domain focus dataset: medical [17], literature [18] and our own data. The medical and literature data is collected by VTSNLP [19]. We measure the ability of model after training by using VMLU [1] benchmark. The VMLU benchmark tests model on wide range of 58 domains, we report model's result on overall of 58 domains and specific domain that related to dataset we trained the

¹<https://huggingface.co/datasets/group2sealion/vnu-hard-clean>

²https://huggingface.co/datasets/zerostratos/augment_edu

³<https://huggingface.co/datasets/zerostratos/syn-edu>

model. For training process on medical dataset, we show model result on clinical pharmacology and basic medicine.

TABLE V
PERFORMANCE OF MODEL TRAINED ON MEDICAL DATASET
COMPARES TO OTHER MODELS.

Model	Pharmacology	Medicine	Overall
Qwen3-0.6B(baseline)	23.89	36.84	34.26
Phi4-Mini-reasoning	24.44	17.54	27.94
Llama3.2-3B	32.78	27.49	34.09
Our med-clora	31.11	40.94	34.59

We show model’s score on Vietnamese language, Middle school exercises which is dominant in the training dataset for training process on literature dataset:

TABLE VI
PERFORMANCE OF MODEL TRAINED ON LITERATURE DATASET
COMPARES TO OTHER MODELS.

Model	Language	Middle school	Overall
Qwen3-0.6B(baseline)	27.01	39.66	34.26
Phi4-Mini-reasoning	29.31	33.33	28.33
Llama3.2-3B	32.18	34.44	34.09
Our litr-clora	29.31	45.4	36.58

Our continual pretraining process was structured into three sequential tasks. Task 1 was an initialization step, consisting of a single training iteration. This was necessary to populate the previous task’s parameters A_{prev} required for the orthogonality loss calculation in the subsequent task, addressing a procedural requirement of C-LoRA. Task 2 involved training the model on the curated real-world datasets from VNU and the filtered educational documents. Finally, Task 3 focused on training exclusively with the newly generated synthetic data to instill stylistic and domain reasoning improvements. The result of model training on our domain data is shown in Table VII.

TABLE VII
PERFORMANCE OF DIFFERENT MODELS ON EDUCATION DOMAIN
DATASETS.

Model	VMLU	Educational QA
Qwen3-0.6B (baseline)	34.26	33.18
Qwen3-0.6B-LoRA-no-synthetic	34.16	32.27
Our Method	36.39	35.0

By constraining model not to forget, we can see that the ability on VMLU benchmark slightly improve, showing that model can acquire new knowledge while most of it’s own ability. We test the model’s answer on education question answering test⁴, which is created from our own dataset including 220 questions. An answer will be categorized into correct, incorrect. The correct answer must satisfy all three conditions: factual accuracy, completeness, and adherence.

The result shows that after training, model of 0.6B can answer 35% of test questions strictly correct. Our

method improve the result by 2%. Our reported result also show that by using LoRA with no synthetic data, the model perform slightly worse than the baseline on VMLU score, showing sign of model forget what it has learned, the score of question answering also worsen.

V. DISCUSSION

As shown in the performance tables, we select the pre-trained model as the baseline. Our results demonstrate that the proposed approach effectively preserves general knowledge while learning new domain abilities. Training performance is further improved by incorporating synthetic data compared to standard LoRA trained only on crawled data. The original C-LoRA paper [4] has already demonstrated its advantage over LoRA in multi-task and continual learning, while recent work [13] similarly confirms that combining synthetic and crawled data enhances training efficiency.

In question answering evaluations, we observe that reasoning models trained on domain corpora also generate more natural Vietnamese text with reduced code-mixing. The models not only provide more accurate answers but also produce clearer reasoning steps. In writing tasks, continual pre-training improves fluency and semantic understanding.

By integrating well-written synthetic and domain-relevant data, we expand both knowledge and stylistic understanding. However, the reliability of synthetic data evaluation—judged by Gemini and ChatGPT—remains limited, and tuning the new loss term introduces hyperparameter sensitivity. Moreover, while the orthogonal loss shows promise, its effectiveness on large-scale continual pre-training remains to be validated.

VI. CONCLUSION

In this work, we presented an effective framework for domain-adaptive continual pretraining of Vietnamese Large Language Models, specifically addressing the critical challenge of limited high-quality domain-specific data. Our method leverages a novel two-phase data pipeline that augments sparse domain corpora with high-quality synthetic data generated by a smaller reasoning model. Combined with the C-LoRA continual learning strategy, which effectively prevents catastrophic forgetting through orthogonality constraints, our framework enables robust knowledge acquisition in new domains.

Our framework successfully demonstrated its efficacy on single-domain Vietnamese datasets, proving significant improvements in in-domain question-answering capabilities. Beyond quantitative performance gains, the models exhibited enhanced generative fluency in Vietnamese, a notable reduction in mix-language outputs. These findings highlight the potential of our approach to not only expand a model’s

⁴<https://huggingface.co/datasets/zerostratos/Education-QA>

factual knowledge but also refine its stylistic understanding and reasoning coherence within specialized domains.

While our method shows promising results for single-domain adaptation, the challenges of continual pretraining across diverse and imbalanced multi-domain datasets, along with the need for more objective evaluation of qualitative aspects, indicate clear avenues for future research. Our work paves the way for developing more specialized and robust Vietnamese LLMs from constrained data resources, contributing significantly to the practical application of AI in low-resource language settings [20].

Acknowledgments. This research has been done under the research project QG.23.73 of Vietnam National University, Hanoi.

REFERENCES

- [1] ZaloAI-Jaist, “Vmlu: Vietnamese multitask language understanding benchmark,” <https://github.com/ZaloAI-Jaist/VMLU>, 2023, dataset and leaderboard with 10,880 multiple-choice questions over 58 subjects.
- [2] T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, and G. Haffari, “Continual learning for large language models: A survey,” *arXiv preprint arXiv:2402.01364*, 2024.
- [3] A. Jadon, A. Patil, and S. Kumar, “Enhancing domain-specific retrieval-augmented generation: Synthetic data generation and evaluation using reasoning models,” *arXiv preprint arXiv:2502.15854*, 2025.
- [4] X. Zhang, L. Bai, X. Yang, and J. Liang, “C-LoRA: Continual low-rank adaptation for pre-trained models,” *arXiv preprint arXiv:2502.17920*, 2025.
- [5] D. Lopez-Paz and M. Ranzato, “Gradient episodic memory for continual learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6467–6476. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/7f9f1f8ee95144ac98b00c4c185c2f6d-Abstract.html>
- [6] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017. [Online]. Available: <https://www.pnas.org/doi/10.1073/pnas.1611835114>
- [7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022, presented January 2022; published at ICLR 2022. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [8] D. Cheng, S. Huang, and F. Wei, “Adapting large language models via reading comprehension,” *CoRR*, vol. abs/2309.09530, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2309.09530>
- [9] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, “Self-instruct: Aligning language models with self-generated instructions,” in *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
- [10] D. Q. Nguyen, L. T. Nguyen, C. Tran, D. N. Nguyen, D. Phung, and H. Bui, “Phogpt: Generative pre-training for vietnamese,” *arXiv preprint arXiv:2311.02945*, 2023.
- [11] C. D. Tran, N. H. Pham, A. T. Nguyen, T. S. Hy, and T. Vu, “Videberta: A powerful pre-trained language model for vietnamese,” in *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, 2023, pp. 1071–1078.
- [12] D. Chen, Y. Huang, X. Pan, N. Jiang, H. Wang, C. Ge, Y. Chen, W. Zhang, Z. Ma, Y. Zhang, J. Huang, W. Lin, Y. Li, B. Ding, and J. Zhou, “Data-juicer 2.0: Cloud-scale adaptive data processing for foundation models,” *arXiv preprint arXiv:2501.14755*, 2024.
- [13] X. Hao, R. Zhu, G. Zhang, K. Shen, and C. Li, “Reformulation for pretraining data augmentation,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.04235>
- [14] Z. Yang, N. Band, S. Li, E. Candès, and T. Hashimoto, “Synthetic continued pretraining,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.07431>
- [15] Wikimedia, “Wikipedia,” <https://huggingface.co/datasets/wikimedia/wikipedia>, 2023, accessed: Jul. 15, 2025.
- [16] HuggingFaceFW, “Fineweb-2,” <https://huggingface.co/datasets/HuggingFaceFW/fineweb-2>, 2024, accessed: Jul. 15, 2025.
- [17] VTSNLP, “Vtsnlp medical datasets,” https://huggingface.co/VTSNLP/base_yTe, 2025, accessed: Jul. 15, 2025.
- [18] —, “Vtsnlp literature datasets,” https://huggingface.co/VTSNLP/base_vanhoc_general, 2025, accessed: Jul. 15, 2025.
- [19] —, “Vtsnlp models and datasets,” <https://huggingface.co/VTSNLP>, 2025, accessed: Jul. 15, 2025.
- [20] Z. Ke, Y. Shao, H. Lin, T. Konishi, G. Kim, and B. Liu, “Continual pre-training of language models,” in *Proc. of International Conference on Learning Representations (ICLR)*, 2023.

Improving Legal Reasoning in LLMs through Structured Prompting

Van-Khanh Tran* and Ke-Luong Nguyen

Institute of Applied Science and Technology - IAST

University of Information and Communication Technology, ICTU

Thai Nguyen City, Vietnam

tvkhanh@ictu.edu.vn, nguyenuongd3k@gmail.com

Abstract—Legal Textual Entailment Recognition (LTER) requires sophisticated reasoning to navigate the syntactic complexity and logical structures inherent in legal language. This study introduces a novel structured prompting approach that leverages Large Language Models (LLMs) without resource-intensive fine-tuning. We design specialized prompts with intricate logical structures that guide model inference to align with the patterns of legal reasoning. Through systematic evaluation on COLIEE 2024 and VLSP 2023 benchmarks, our approach achieves 97% accuracy on Vietnamese legal texts and outperforms GPT-4 and strong previous research on multiple English legal reasoning tasks. These results demonstrate that carefully designed prompting strategies can unlock complex legal logic in LLMs while maintaining computational efficiency.

Index Terms—LTER, Structured Prompting, Legal Document Analysis.

I. INTRODUCTION

The digital transformation of legal systems has created an urgent need for intelligent document analysis tools capable of complex legal reasoning [1]. Legal Textual Entailment Recognition (LTER) - determining whether a legal statement logically follows from given legal text - represents a critical challenge in this domain [2]. Traditional approaches rely on fine-tuning pre-trained models like Legal-BERT [3], requiring substantial computational resources and large labeled datasets. In contrast, this study explores whether structured prompting can achieve comparable performance without parameter updates, making legal AI more accessible and cost-effective [4].

In this approach, instead of modifying the weights of the LLM models, we design and implement specialized prompts with intricate logical structures. These prompts are meticulously optimized to reflect the unique linguistic characteristics and reasoning patterns of legal discourse, a critical step for adapting general models to specialized domains [5]. By doing so, they guide the model’s inference process to align with domain-specific legal logic, a technique that can be enhanced by structuring the prompt

to elicit step-by-step reasoning [6]. This enables the model to quickly adapt to deep semantic requirements, complex reasoning patterns, and specialized terminology without the need for costly model retraining [7].

Our contributions are fourfold:

- Novel modular prompt architecture for legal reasoning
- Comprehensive evaluation across 7 prompting strategies and multiple model architectures
- Demonstration that specialized, smaller models outperform larger general-purpose ones
- Analysis of cross-linguistic performance patterns in legal reasoning

II. RELATED WORKS

Early legal NLP relied on traditional machine learning with Support Vector Machines [8] and lexical matching [9], often enhanced with legal ontologies [10]. These approaches struggled with the deep semantic relationships in legal text. The introduction of transformer-based models like BERT [5], [11] and Legal-BERT [3] significantly improved performance through pre-training on legal corpora. However, full fine-tuning remains computationally expensive and requires large domain-specific datasets [7].

Recent works have explored LLMs like GPT-3.5 and Llama for legal tasks using zero-shot and few-shot learning [12], [13]. At the same time, prompt-tuning—a technique that enables task adaptation without modifying model weights—remains underutilized in this domain [14]. While promising, most studies have not fully leveraged LLMs’ reasoning capabilities for rigorous logical tasks, such as LTER. General-purpose prompting often fails to unlock the nuanced, domain-specific inference required. Advanced adaptation techniques that do not modify model weights, such as prompt-tuning, remain underutilized in this domain. This reveals a critical gap: a method is needed to specifically guide an LLM’s reasoning process to align with formal legal logic without costly retraining.

Chain-of-Thought (CoT) [6] prompting has shown success in mathematical and logical reasoning tasks. Tree-of-Thought [15] and self-consistency [16] methods

* Corresponding author: tvkhanh@ictu.edu.vn

further enhance reasoning by exploring multiple solution paths. However, their application to domain-specific legal reasoning has not been systematically studied.

In this study, we address gaps by proposing a **structured prompting** approach for LLMs that effectively leverages legal linguistic knowledge. By designing task-specific prompts focused on argument structure and legal terminology, our method unlocks the sophisticated reasoning capabilities of LLMs, achieving significant advances in recognizing entailment relations between legal text segments.

III. METHODS

A. Structured Prompt Design Framework

Unlike conventional reasoning, legal texts exhibit distinct characteristics such as complex structures, implicit logic, and specialized terminology. Therefore, in this study, we develop a prompt design strategy that adapts to the model architecture as well as common reasoning errors in legal inference identified from error samples of large language models (LLMs). Legal texts exhibit distinct characteristics such as complex structures, implicit logic, and specialized terminology that differ from conventional reasoning. To unlock the logic embedded within these texts, this study develops a **structured prompting** strategy. This strategy is meticulously designed to adapt to specific model architectures and to mitigate common legal inference errors that were identified from analyzing the outputs of LLMs.

1) Design Principles:

a) *Principle 1: Modular Composition:* We adopt a compositional approach where complex prompts are built from atomic functional blocks. This enables systematic experimentation while maintaining consistency across variants.

b) *Principle 2: Legal Domain Adaptation:* Each component explicitly addresses legal reasoning challenges identified through error analysis of baseline LLM outputs on legal tasks, including:

- Implicit legal knowledge requirements
- Complex conditional reasoning chains
- Careful negation handling
- Preservation of legal meaning through interpretation

c) *Principle 3: Reasoning Strategy Specialization:* Different legal reasoning tasks may benefit from different cognitive strategies (sequential, parallel, meta-cognitive), requiring distinct prompt architectures.

2) *Functional Block Taxonomy:* We define eight atomic blocks with specific legal reasoning functions:

$$\mathcal{B} = \{B_{\text{basic}}, B_{\text{guide}}, B_{\text{1shot}}, B_{\text{few}}, B_{\text{cot}}, B_{\text{step}}, B_{\text{self}}, B_{\text{tot}}\} \quad (1)$$

where:

- B_{basic} (Basic Block): Role definition, task objective, entailment definition.

- B_{guide} (Guidelines Block): Encodes domain-specific reasoning constraints: Inference scope limitation, negation and conditional statement handling.
- B_{1shot} (One-shot Block): Contains one carefully selected example demonstrating: Clear positive entailment case, explanation of legal reasoning process, common error avoidance.
- B_{few} (Few-shot Block): Includes 3-5 examples covering: Positive and negative entailment cases, edge cases with subtle logical relationships, examples of common reasoning pitfalls, and diverse legal contexts and terminology.
- B_{cot} (Chain-of-Thought Block): Implements structured analytical framework.
- B_{step} (Step-back Block): High-level principle analysis incorporating meta-cognitive reflection.
- B_{self} (Self-consistency Block): Employs multiple reasoning approaches: Direct textual analysis, Legal principle application, Analogical reasoning.
- B_{tot} (Tree-of-Thought Block): Implements parallel reasoning with voting.

B. Prompt Generation System

1) *Mathematical Formalization:* Each prompt variant P_i is generated through selective block composition:

$$P_i = B_{\text{basic}} \oplus \bigcup_{j \in S_i} B_j \quad (2)$$

where $S_i \subseteq \{\text{guide}, \text{1shot}, \text{few}, \text{cot}, \text{step}, \text{self}, \text{tot}\}$ represents the set of additional blocks activated for prompt variant i , and $\oplus$ denotes the structured concatenation operator.

2) *Configuration Constraints:* Our framework generates seven distinct prompt variants through the following configurations:

Prompt Type	Configuration Formula	Block Activation
Zero-shot	$P_{\text{zero}} = B_{\text{basic}} \oplus B_{\text{guide}}$	$S_{\text{zero}} = \{\text{guide}\}$
One-shot	$P_{\text{one}} = B_{\text{basic}} \oplus B_{\text{guide}} \oplus B_{\text{1shot}}$	$S_{\text{one}} = \{\text{guide}, \text{1shot}\}$
Few-shot	$P_{\text{few}} = B_{\text{basic}} \oplus B_{\text{guide}} \oplus B_{\text{few}}$	$S_{\text{few}} = \{\text{guide}, \text{few}\}$
Chain-of-Thought	$P_{\text{cot}} = B_{\text{basic}} \oplus B_{\text{cot}}$	$S_{\text{cot}} = \{\text{cot}\}$
Step-back	$P_{\text{step}} = B_{\text{basic}} \oplus B_{\text{guide}} \oplus B_{\text{step}}$	$S_{\text{step}} = \{\text{guide}, \text{step}\}$
Self-consistency	$P_{\text{self}} = B_{\text{basic}} \oplus B_{\text{self}}$	$S_{\text{self}} = \{\text{self}\}$
Tree-of-Thought	$P_{\text{tot}} = B_{\text{basic}} \oplus B_{\text{tot}}$	$S_{\text{tot}} = \{\text{tot}\}$

TABLE I: Configuration Matrix

Certain blocks have logical dependencies that constrain valid configurations:

1) Mutual Exclusivity Constraints:

- $B_{\text{1shot}} \cap B_{\text{few}} = \emptyset$: Cannot activate both simultaneously.
- $|\{\text{cot}, \text{step}, \text{self}, \text{tot}\} \cap S_i| \leq 1$: Single advanced reasoning strategy.

2) Dependency Requirements:

- $B_{\text{basic}} \in P_i$ for all i : Basic block always required.
- B_{guide} recommended for complex reasoning strategies
- Advanced blocks (*cot*, *step*, *self*, *tot*) require careful model selection.

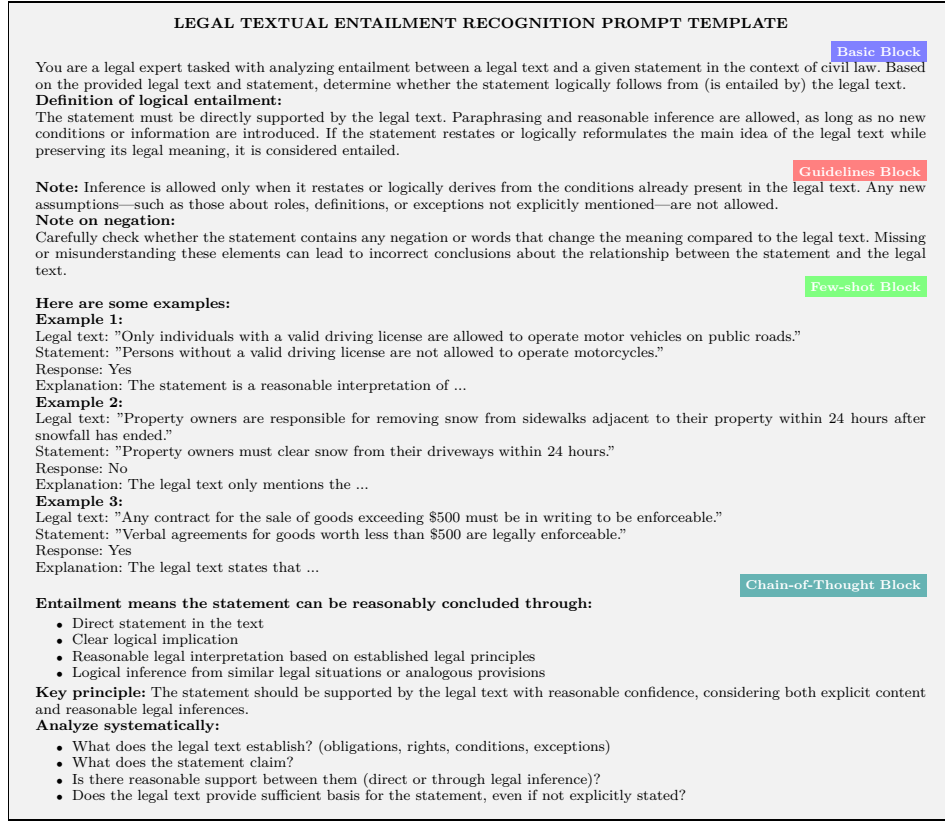


Fig. 1: LTER Structured Prompt Template - Part 1: Basic, Guidelines, One-shot, Few-shot, and Chain-of-Thought Blocks

3) *Component Formalization*: Each prompt component is formally defined as follows:

Expert Task (ET): Establishes the model’s role as a legal expert specializing in civil law textual entailment.

Objective (O): Defines the core task of determining logical entailment between legal text and statements.

Definition (D): Specifies entailment criteria through four pathways (details in Sec. III-C1).

Key Principle (KP): Establishes that statements must be supported with reasonable confidence, considering both explicit content and reasonable legal inferences.

Analysis Framework (AF): Defines the systematic four-step analytical process for controlled reasoning workflow.

Examples Block (B_{examples}): Contains demonstration cases where $E_i = \langle T_i, S_i, L_i, X_i \rangle$ where T_i is the legal text, ...

where T_i is the legal text, S_i is the statement, $L_i \in \{\text{Yes}, \text{No}\}$ is the label, and X_i is the explanation.

Reasoning Block ($B_{\text{reasoning}}$): Implements specific reasoning methodologies: *CoT* (Chain-of-thought reasoning), *SB* (Step-back prompting), *SC* (Self-consistency), *ToT* (Tree-of-thought).

C. Legal Domain Specification

1) *Entailment Definition Framework*: We operationalize legal textual entailment through four validated pathways:

- **Type 1: Direct Explicit Statement**: The statement directly restates information present in the legal text without modification. Example: Text mentions "property owners must..." → Statement "property owners are required to..."
- **Type 2: Clear Logical Implication** The statement follows necessarily from explicit conditions in the text through formal logical inference. Example: Text "Only licensed drivers may operate vehicles" → Statement "Unlicensed drivers cannot operate vehicles"
- **Type 3: Established Legal Principle Application** The statement represents reasonable interpretation based on widely accepted legal principles and precedents. Example: Contract law principles applied to specific contractual terms
- **Type 4: Analogical Legal Reasoning** The statement extends from similar legal situations or analogous provisions within the same legal framework. Example: Extending liability principles from one context to similar situations

2) Legal Reasoning Constraints:

- **Constraint 1: No Novel Information Introduction** Statements cannot introduce new facts, conditions, or assumptions not present or reasonably derivable from the source text.
- **Constraint 2: Negation Preservation** Logical negations must be carefully preserved and cannot be

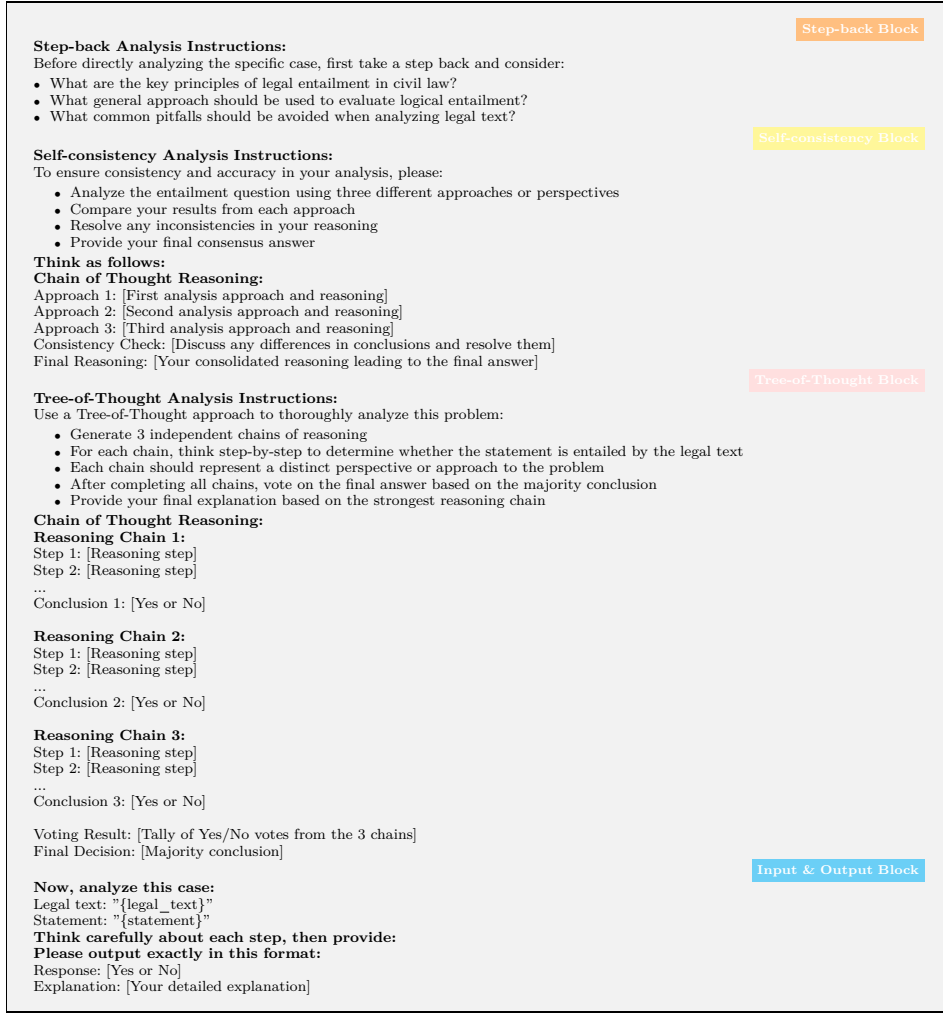


Fig. 2: LTER Structured Prompt Template - Part 2

altered through inference processes.

- **Constraint 3: Conditional Structure Maintenance** If-then relationships and conditional dependencies must be maintained in any derived statements.
- **Constraint 4: Legal Context Sensitivity** Interpretations must remain within the established legal context and jurisdiction specified in the source text.

D. Model-Specific Optimization

1) *Architecture-Dependent Configuration: Reasoning-Specialized Models (Qwen3-32B, QwQ-32B)*: Prioritize CoT and ToT strategies, leverage enhanced logical reasoning capabilities, use complex conditional reasoning blocks; *General-Purpose Models*: Emphasize few-shot learning with rich examples, provide explicit reasoning templates, include detailed constraint specifications; *Smaller Models (<10B parameters)*: Focus on few-shot learning strategies, provide comprehensive examples, avoid overly complex reasoning chains.

2) *Configuration Selection Framework*: The optimal configuration selection follows a systematic approach

based on task requirements and reasoning complexity:

$$S^* = \arg \max_{S \in \mathcal{S}_{\text{valid}}} \text{Effectiveness}(P_{\text{basic}} \oplus \bigcup_{j \in S} B_j) \quad (3)$$

where $\mathcal{S}_{\text{valid}}$ represents the set of all valid configurations satisfying the dependency rules, and Effectiveness measures the prompt’s performance on legal entailment tasks.

3) *Template Instantiation Process*: The complete prompt generation process can be formalized as:

$$\text{Generate}(\text{strategy}, \text{config}) = \bigcup_i B_i \cdot \mathcal{I}_{\text{config}}(B_i)$$

where $\mathcal{I}_{\text{config}}(B_i)$ is the indicator function that includes block B_i if enabled in the configuration matrix for the specified strategy. This modular architecture ensures consistency across prompt variants while enabling systematic experimentation with different reasoning approaches for legal entailment tasks, providing both theoretical rigor and practical flexibility in prompt design.

IV. EXPERIMENTAL SETUP

A. Datasets

Our experiments were conducted on two specialized legal textual entailment datasets. The first is **VLSP 2023** [17], a Vietnamese-language dataset focused on the domain of civil law. It is composed of premise-hypothesis pairs that require sophisticated logical reasoning to identify legal relationships. The second dataset is **COLIEE 2024** [18], [19], from the Competition on Legal Information Extraction/Entailment. This dataset is presented in English but features complex legal reasoning tasks derived from the Japanese civil code, demanding a nuanced understanding of statutory logic.

B. Model Selection

We evaluated a diverse suite of LLMs (1.5B to 72B parameters), including reasoning-specialized models (Qwen3-32B, QwQ-32B), general-purpose architectures (LLaMA variants, SaoLa series), and established baselines such as GPT-3.5, GPT-4, and models from competition participants.

C. Prompting Strategies

We systematically compared seven prompting strategies, ranging from basic methods like *Zero-shot* (task definition only), *One-shot* (a single example), and *Few-shot* (multiple examples) to more advanced reasoning techniques. These included *Chain-of-Thought*, which elicits step-by-step logic; *Step-back*, which encourages meta-cognitive reflection; *Self-consistency*, which uses a voting mechanism across multiple reasoning outputs; and *Tree-of-Thought*, which explores parallel reasoning chains.

V. RESULTS

A. Overall Performance Summary

Our structured prompting approach achieved state-of-the-art performance on both evaluation datasets. On VLSP 2023 (Vietnamese legal texts), the best configuration reached **97.00%** accuracy using Qwen3-32B with few-shot prompting. On COLIEE 2024 (English legal reasoning), the optimal setup achieved **90.83%** accuracy using QwQ-32B with chain-of-thought prompting.

B. Model Architecture Analysis

1) *Parameter Efficiency vs. Performance*: Reasoning-specialized models significantly outperformed larger general-purpose architectures. Qwen3-32B (32B parameters) achieved 95.6% average accuracy on VLSP 2023, surpassing SaoLa-Llama-3.3-70B-Instruct (93.7%) despite having 54% fewer parameters. This pattern was more pronounced on COLIEE 2024, where QwQ-32B (32B) averaged 82.1% compared to 77.3% for the 70B model.

2) *Top-Performing Models*: Table II presents the top-5 models ranked by average performance across all prompting strategies on both datasets.

TABLE II: Performance Comparison of LLMs on Legal Benchmarks

Rank	Model	VLSP 2023 (%)	COLIEE 2024 (%)	Parameters
1	Qwen3-32B	95.6	78.9	32B
2	QwQ-32B	95.4	82.1	32B
3	SaoLa-Llama-3.3-70B	93.7	77.3	70B
4	Qwen2.7B-Instruct	87.1	74.2	7B
5	Llama-3.1-8B-Instruct	79.0	68.1	8B

C. Benchmark Comparisons

1) **VLSP 2023 Competition**: Our approach substantially outperformed official competition participants, achieving 97.0% accuracy compared to the winning team’s 76.26% (+20.74 percentage points). Table III summarizes the complete comparison.

TABLE III: VLSP 2023 Competition Benchmark

Team/Model	Source	Accuracy (%)
Structured Prompting	This Study	97.00
NOWJ	VLSP 2023	76.26
CAN NOT STOP	VLSP 2023	71.98
A3N1	VLSP 2023	71.94
Angels	VLSP 2023	54.68
HNO3	VLSP 2023	53.24

2) **COLIEE 2024 Competition**: Our structured prompting approach demonstrates substantial improvements over both baseline commercial models and competition participants.

TABLE IV: Comparison of Top-3 Structured Prompting Strategies with Top Results on COLIEE 2024

Team / Method	#Correct	R05	R02	R01	H30
Top-1 Structured Prompting	N/A	N/A	0.8765	0.7385	0.8857
Top-2 Structured Prompting	N/A	N/A	0.8642	0.8154	0.8286
Top-3 Structured Prompting	N/A	N/A	0.8519	0.7846	0.7857
CAPTAIN2	90	0.8257	0.7901	0.7568	0.7543
JNLP1*	89	0.7901	0.6937	0.7429	0.6937
UA_slack	87	0.7982	0.7401	0.7165	0.7429
UA_encoder_decoder	87	0.7982	0.7401	0.7165	0.7429
CAPTAIN1	86	0.8148	0.7703	0.7407	0.7286
JNLP2*	86	0.7890	0.8395	0.7207	0.7286
UA_gpt	85	0.7798	0.7401	0.7165	0.7143
AMHR.ensembleA50	84	0.7901	0.8148	0.7387	0.7286
AMHR.single	84	0.7798	0.7284	0.7165	0.7143
H11	83	0.7651	0.7284	0.6937	0.6857
NOWJ.pandap46*	82	0.7523	N/A	N/A	N/A
AMHR.ensembleA0	80	0.7778	0.7423	0.6423	0.7000
NOWJ.llanT5-panda*	80	0.7401	N/A	N/A	N/A
OVGU1+	77	0.7059	0.7631	0.6374	0.6286
KIS2+	76	0.6937	0.6790	0.6036	0.6000
OVGU2+	76	0.6972	0.6543	0.5664	0.6000
KIS1	70	0.6147	0.6173	0.5679	0.6000
H13	66	0.5826	0.6049	0.5936	0.5714
H12	63	0.5678	0.5926	0.6306	0.6429
KIS3	62	0.5688	0.5926	0.6306	0.6429

Table IV provides a comprehensive overview of the official COLIEE 2024 competition results [18]. Our structured prompting approach demonstrated exceptional performance on the COLIEE 2024 benchmark, achieving substantial improvements over established competition methods and commercial models. The top-performing configuration reached 90.83% accuracy on group R03, while consistently outperforming all competition teams across multiple test groups, particularly excelling on R02 (87.65% vs. best competitor’s 79.01%) and H30 (88.57% vs. 74.29%). Notably, even our third-best prompting variant surpassed the strongest competition entry (CAPTAIN2 with 90 correct answers), demonstrating the robustness of our methodology.

Table V compared against commercial baselines [20]. The highest accuracy of 90.83% (R03) and 90.00% (H25)

TABLE V: Performance Comparison: GPT Models vs. Top Structured Prompting Methods (English Only)

Year	GPT-3.5 (EN)	GPT-4 (EN)	Top-1 Structured	Top-2 Structured	Top-3 Structured
H18	0.7222	0.6944	0.8333	0.8333	0.7222
H19	0.6486	0.7838	0.8919	0.8919	0.7568
H20	0.8049	0.8293	0.7805	0.8049	0.8537
H21	0.7778	0.7222	0.7963	0.7963	0.8519
H22	0.6596	0.6383	0.7660	0.7447	0.7660
H23	0.7561	0.7561	0.8537	0.8293	0.8293
H24	0.6456	0.6709	0.7089	0.6962	0.7215
H25	0.8167	0.8333	0.9000	0.8833	0.8000
H26	0.6486	0.7162	0.8243	0.8514	0.7297
H27	0.7755	0.7755	0.7959	0.7551	0.6735
H28	0.8571	0.8980	0.8776	0.8571	0.9184
H29	0.6207	0.7414	0.7931	0.8448	0.7586
H30	0.5857	0.7429	0.8857	0.7857	0.8286
R01	0.6486	0.8108	0.7385	0.7846	0.8154
R02	0.6914	0.8148	0.8765	0.8519	0.8642
R03	0.7156	0.8440	0.9083	0.9083	0.8073

significantly outperforms GPT-4’s comparable results, where GPT-4 achieved 84.40% on R03 and 83.33% on H25. The results show consistent improvements across most test groups compared to both GPT-3.5 and GPT-4 as documented in the black-box analysis study. For instance, on group H19, our approach achieved 89.19% compared to GPT-4’s 78.38% from the referenced study (+13.8%), and on H30, our result of 88.57% substantially exceeds GPT-4’s 69.44% (+27.5%).

These results validate that carefully designed structured prompting can unlock sophisticated legal reasoning capabilities in language models, providing a computationally efficient alternative to resource-intensive fine-tuning approaches while achieving state-of-the-art performance in complex legal textual entailment tasks.

VI. CONCLUSIONS

This study demonstrates that a structured prompting framework can unlock sophisticated legal reasoning in LLMs without costly fine-tuning. Our method achieved state-of-the-art results, reaching 97% accuracy on the VLSP 2023 dataset and significantly outperforming GPT-4 on the complex COLIEE 2024 benchmark. The central finding is that smaller, reasoning-specialized models, when guided by domain-specific prompts, are more effective than larger general-purpose models, suggesting that model architecture is more critical than raw parameter count for legal logic. While limitations in integrating external legal knowledge persist, future work will focus on hybrid fine-tuning techniques for more complex tasks. This research validates our approach as a powerful and efficient pathway toward developing more intelligent and accessible legal AI tools.

REFERENCES

- [1] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, “How does nlp benefit legal system: A summary of legal artificial intelligence,” in *Proceedings of ACL*, 2020.
- [2] C. Steging and L. V. Leeuwen, “A hybrid approach to legal textual entailment,” in *Proceedings of the Eighteenth International Workshop on Juris-Informatics (JURISIN 2024)*, 2024.
- [3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “Legal-bert: The muppets straight out of law school,” 2020. [Online]. Available: <https://arxiv.org/abs/2010.02559>
- [4] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” 2021. [Online]. Available: <https://arxiv.org/abs/2107.13586>
- [5] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The muppets straight out of law school,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2898–2904. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.261/>
- [6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2201.11903>
- [7] J. Dodge, T. Prewitt, R. T. D. Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan, “Measuring the carbon intensity of ai in cloud instances,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.05229>
- [8] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.
- [9] D. S. Carvalho, V. D. Tran, K. Van Tran, V. D. Lai, and M.-L. Nguyen, “Lexical to discourse-level corpus modeling for legal question answering,”
- [10] J. Rabelo, M. Y. Kim, H. Babiker, R. Goebel, and N. Farruque, “Information extraction and entailment for statute law and case law,” in *JURISIN 2018*, 2018.
- [11] S. Wehnert, S. Dureja, L. Kutty *et al.*, “Applying bert embeddings to predict legal textual entailment,” *Revista de Socienetwork Strategica*, vol. 16, pp. 197–219, 2022.
- [12] H.-T. Nguyen, R. Goebel, F. Toni, K. Stathis, and K. Satoh, “Black-box analysis: Gpts across time in legal textual entailment task,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.05501>
- [13] A. Al Zubaer, M. Granitzer, and J. Mitrović, “Performance analysis of large language models in the domain of legal argument mining,” *Frontiers in Artificial Intelligence*, vol. Volume 6 - 2023, 2023. [Online]. Available: <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1278796>
- [14] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” 2021. [Online]. Available: <https://arxiv.org/abs/2104.08691>
- [15] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, “Tree of thoughts: Deliberate problem solving with large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.10601>
- [16] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-consistency improves chain of thought reasoning in language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2203.11171>
- [17] V. Tran, H.-T. Nguyen, T. Vo, S. T. Luu, H.-A. Dang, N.-C. Le, T.-T. Le, M.-T. Nguyen, T.-S. Nguyen, and L.-M. Nguyen, “Vlsp 2023 – lter: A summary of the challenge on legal textual entailment recognition,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.03435>
- [18] R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka, “Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024,” in *New Frontiers in Artificial Intelligence*, T. Suzumura and M. Bono, Eds. Singapore: Springer Nature Singapore, 2024, pp. 109–124.
- [19] A. Nigohjkar, K. Jiang, L. Fields, O. Bilgin, S. Steinle, Y. Sadybekov, and J. Licato, “Amhr coliee 2024 entry: Legal entailment and retrieval,” in *JSAT International Symposium on Artificial Intelligence*. Singapore: Springer Nature Singapore, 2024, pp. 200–211.
- [20] H. Nguyen, R. Goebel, F. Toni, K. Stathis, and K. Satoh, “Black-box analysis: Gpts across time in legal textual entailment task,” *arXiv preprint arXiv:2309.05501*, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2309.05501>

Optimizing Image Dehazing with Balanced Loss and Feature Map Enhancement

Trung-Nam Bui-Huynh

*Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
s2310440@jaist.ac.jp*

Atsuo Yoshitaka

*Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
ayoshi@jaist.ac.jp*

Abstract—Image dehazing, a low-level task in computer vision, aims to estimate haze-free images from their corresponding degraded counterparts and has demonstrated substantial performance improvements through deep neural networks. However, several limitations persist, including an insufficient loss function, underutilization of channel relationship information, and a suboptimal integration of convolution operations, all of which may decrease its performance. To address these challenges of image dehazing, we introduce a novel approach, DEA-Net-CR-LapSE, which enhances the state-of-the-art Detail-Enhanced Attention Network (DEA-Net) through several contributions. First, we integrate an additional loss, named Laplacian Pyramid Loss (LPL), into the total loss function to enhance the efficacy of Difference Convolutions (DCs) in capturing high-frequency information. Second, we adjust the channel attention within the DEA-Net to mitigate the influence of spatial attention outputs on the channel attention outputs. Third, we propose an alternative fusion strategy that combines four DCs and one Vanilla Convolution (VC) to improve the stability of feature maps. Comprehensive experiments conducted on various benchmark datasets demonstrate the superior performance of DEA-Net-CR-LapSE, surpassing the original model, through PSNR, SSIM, and LPIPS metrics.

Index Terms—image dehazing, Laplacian Pyramid Loss, data-based method

I. INTRODUCTION

Image dehazing is the process of enhancing images degraded by haze or other atmospheric conditions, which often result in blur, noise, or color distortion. The primary goal of the image dehazing task is to produce clear, sharp, and high-resolution images. Since the performance of many vision applications is heavily influenced by the input data quality, image dehazing becomes a crucial preprocessing step for enhancing image clarity and visibility. Image dehazing is particularly beneficial for applications such as autonomous driving, video surveillance, and aerial imaging, where it aids in reliable scene understanding and object detection in challenging environments. Therefore, researching and developing low-level tasks such as image dehazing is not only necessary but also important, as it provides a foundation for and supports the advancement of high-level tasks.

Recent deep learning advancements have shifted image dehazing from traditional physical models, such as Atmospheric Scattering Model (ASM) [1], to data-driven approaches, exemplified by DEA-Net [2]. This DEA-Net model integrates Difference Convolution (DC) [3] to extract high-frequency information and create a Spatial Importance Map (SIM) for each input channel, therefore improving its performance in

single image dehazing. However, several limitations exist: (1) Despite DEA-Net’s integration of DCs and VC to enhance feature extraction, the L1 loss function’s limited sensitivity to high-frequency information restricts the model’s learning ability; (2) Applying channel attention outputs to spatial attention outputs — rather than directly to input features — may reduce the effectiveness of channel attention mechanism; (3) The unbalanced use of 4 DCs over 1 Vanilla Convolution (VC) creates an imbalance that undermines the complementary strengths of both convolution types.

This study aims to address the single image dehazing problem, with a particular focus on the DEA-Net architecture. The main contribution includes:

- Integrating the Laplacian Pyramid Loss [4], [5] to preserve texture details and edge sharpness, enhancing DCs’ ability to capture high-frequency information.
- Refining the Squeeze-and-Excitation (SE) mechanism [6] applied to channel attention, aiming to minimize the interference of unrelated information on channel relationships, thereby enhancing the quality of feature maps.
- Revising the integration of DCs and VC to achieve a more balanced information fusion, improving feature representation quality.
- Integrating all the aforementioned modifications to develop an enhanced version of the DEA-Net model, referred to as DEA-Net-CR-LapSE.
- Experiments are conducted on various benchmark datasets with several metrics used for comprehensive model comparison. Learned Perceptual Image Patch Similarity (LPIPS) [7] metric is also used to assess image similarity, offering a more human-aligned evaluation than traditional pixel-based metrics.

II. RELATED WORK

Early image dehazing approaches, often termed prior-based methods, primarily relied on statistical priors or hand-crafted assumptions about the physical characteristics of the natural environment. A typical representative is Atmospheric Scattering Model (ASM), which can generate haze-free image by making many assumptions about the transmission map and atmospheric light. However, these approaches show unstable results that are highly sensitive to variations in surrounding factors. Subsequent methods based on the ASM were developed in succession, including DCP [8], which assumes at least one color channel has low intensity values in clear outdoor

images, or CLP [9], which indicates that small patches in clear images can be formed by the linear distribution of colors.

Later, the field of image dehazing has experienced significant advancements because of the rapid development of machine learning and deep learning. This has led to the emergence of data-driven methods, such as DCPDN [10], FFA-Net [11], and DehazeFormer [12], which utilize paired hazy and clear image datasets to train neural networks for end-to-end haze removal. Recently, the Detail-Enhanced Attention Network (DEA-Net) [2], a typical representative of data-driven methods, was introduced by integrating DCs for feature extraction and applying a SIM to each input channel. This model achieved high performance while maintaining a significantly lower parameter count compared to other approaches. Despite its strengths, the DEA-Net model has limitations, which this study aims to address.

III. METHODOLOGY

A. Incorporating Laplacian Pyramid Loss

The DEA-Net model effectively combines Difference Convolutions (DC) and Vanilla Convolution (VC) to capture both gradient-level and intensity-level information for image dehazing. However, while L1 Loss preserves low-frequency components, it struggles with high-frequency details, such as edges and sharp transitions, which can degrade learning performance. Enhancing the total loss function to better capture high-frequency variations will improve DC capability and overall model performance.

In this study, we introduce an additional loss function, Laplacian Pyramid Loss (LPL) [4], [5], because it effectively emphasizes high-frequency components by capturing the differences between the image and its smoothed version. This characteristic of LPL plays a pivotal role in improving edge sharpness and structural integrity, making it particularly beneficial for models that employ DCs such as DEA-Net. In detail, the Laplacian Pyramid will be calculated as follows:

$$Pyramid^l(J) = J_D - J_U \quad (1)$$

where $Pyramid^l(J)$ denotes the difference at the l -th level of the pyramid, and this pyramid uses model output J as the input. Specifically, J_D refers to the downsampled image with input from the previous pyramid level ($l - 1$), while J_U represents the upsampled image with input from the next level ($l + 1$) of the pyramid. Note that, J_D is the input J at the first level ($l = 0$). Additionally, the Gaussian kernel will be applied before the downsampling process. The LPL calculates the difference between the Laplacian pyramids of the model output J and the ground truth GT at each pyramid level l as follows:

$$\mathcal{L}_{Lap} = \sum_{l=0}^{L-1} \mathcal{L}_1(Pyramid^l(J), Pyramid^l(GT)) \quad (2)$$

where $\mathcal{L}_{Lap}$ is LPL, L is the total level of the pyramid, and $\mathcal{L}_1$ is L1 loss. By leveraging the multi-scale structure of LPL, the model's ability to preserve high-frequency is progressively

refined, leading to more effective and accurate outputs. The final loss will be a combination of 3 loss functions as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{CR} + \lambda_3 \mathcal{L}_{Lap} \quad (3)$$

where $\mathcal{L}_{CR}$ denotes Contrastive Regularization [13], which is used in DEA-Net-CR (a version of DEA-Net). The coefficients λ_1 , λ_2 , and λ_3 serve as weighting factors to balance the relative contributions of the individual loss components.

B. Optimization of Channel Attention Mechanism

The Content-Guided Attention (CGA) in DEA-Net [2] mitigates uneven haze distribution by generating a Spatial Importance Map (SIM) for each input feature channel. It also enables interaction between spatial and channel attention, improving feature map quality. Although the integration of spatial and channel attention offers several advantages, this mechanism has certain limitations. The channel attention mechanism in CGA is inspired by the Squeeze-and-Excitation (SE) method [6], which proves more effective in generating channel relationship information. However, in contrast to the original SE, CGA integrates the channel attention outputs with the spatial attention outputs, rather than directly combining them with the input. Consequently, the spatial attention information can interfere with the channel relationship information, potentially leading to being dominated or causing negative alterations in the channel attention.

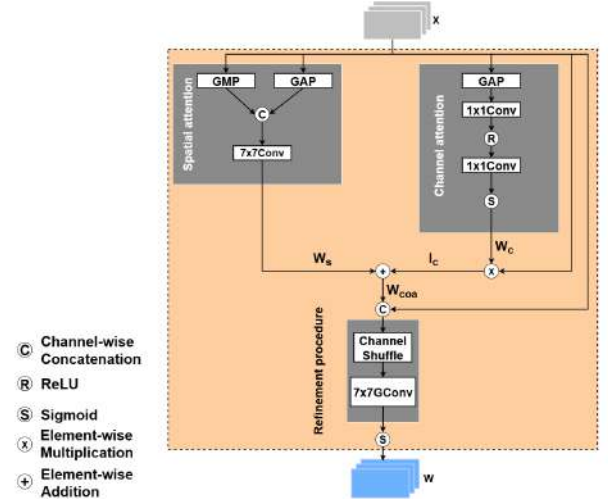


Fig. 1: Architecture of Content-Guided Attention (CGA) mechanism of DEA-Net [2] with refinement in channel attention.

To address this issue, we refine the channel attention mechanism in CGA by directly scaling the channel attention output with the CGA input before combining with the spatial attention output. As shown in Fig. 1, we modify the channel attention to fully implement the SE mechanism, ensuring that channel attention directly influences the CGA input. We hypothesize that this full SE implementation offers a more substantial contribution to the feature maps, thereby improving the quality

of outputs of CGA. In detail, channel attention in CGA will be modified as follows:

$$W_c = \sigma(C_{1 \times 1}(\max(0, C_{1 \times 1}(X_{GAP}^c)))) \quad (4)$$

where $C_{n \times n}$ denotes convolution operation with kernel size $n \times n$, and X_{GAP}^c is the features obtained by global average pooling over the spatial dimensions. $\max(0, x)$ and σ are the ReLU and Sigmoid activation function, respectively. To mitigate the influence of spatial attention information, we combine the input features X with the weight matrix W_c using element-wise multiplication $[\cdot]$ to get the final result of channel attention I_c as below:

$$I_c = X \cdot W_c \quad (5)$$

This scale step occurs immediately after the creation of W_c , effectively reducing the impact of spatial attention information on the channel relationship data. However, to preserve the influence of spatial attention information, we continue to combine the enhanced channel quality features I_c with the spatial attention information W_s to create W_{coa} as follows:

$$W_{coa} = W_s + I_c \quad (6)$$

Furthermore, all other processes will remain consistent with those in the original CGA of the DEA-Net model. As a result, the CGA can still facilitate the generation of the SIM for each channel of the input features, thus addressing the issue of uneven haze distribution at the feature level.

C. Strategy for Integrating DCs and VC

DEA-Net combines DC [3] and VC to capture both gradient-level and intensity-level information, enhancing feature map quality. However, the DEA-Net uses four DCs and a single VC with equal contributions to feature maps. As discussed in [3], there exists a relationship between Central Difference Convolution (CDC) and VC. Specifically, VC represents a special situation of CDC when CDC does not contribute to the result because of the generalized characteristic of CDC. This highlights the need to establish a balanced contribution ratio when concurrently employing both CDC (or other types of DC) and VC in the feature extraction process.

To balance the contributions of DCs and VC, we adapt the strategy of Detail-Enhanced Convolution (DEConv), particularly by modifying the re-parameterization technique. This adjustment redistributes the information extracted from DCs (gradient-level features) and VC (intensity-level features) while maintaining the total number of model parameters. As shown in Fig. 2, the feature extraction of DEConv is split into two flows: the DC process and the VC process, allowing precise control over each convolution type's contribution. The DEA-Net indicated that multiple 2D kernels of the same size can be replaced by a single equivalent kernel if their outputs are summed (in case of identical padding and stride). Using this property, we aggregate the DC kernels, including central difference (CDC), horizontal difference (HDC), vertical difference (VDC), and angular difference (ADC), into a kernel named AGG, while the VC kernel remains unchanged. Both

the AGG and VC kernels are then applied to the input to generate feature maps separately: the AGG kernel captures high-frequency and gradient-level information, and the VC kernel captures low-frequency and intensity-level features. The process is detailed in the following equations:

$$F_{dc} = \sum_{i=1}^4 F_{in} * K_i^{dc} = F_{in} * \left(\sum_{i=1}^4 K_i^{dc} \right) = F_{in} * K_{AGG} \quad (7)$$

$$F_{vc} = F_{in} * K_{VC} \quad (8)$$

where F_{in} represents the input features, K_i^{dc} denotes the i -th DC kernel, K_{AGG} is the aggregated kernel formed by summing the four DC kernels. Besides, K_{VC} is the VC kernel and $*$ is convolution operation. Note that the bias for K_{AGG} is generated through the same process used to create K_{AGG} . After that, only feature maps derived from the DC process

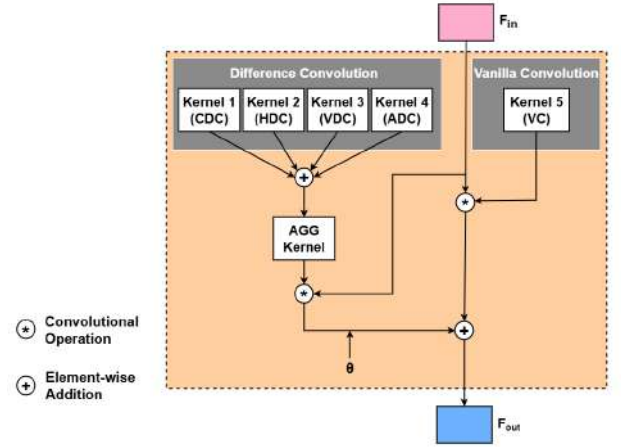


Fig. 2: Application of a new strategy to Detail-Enhanced Convolution (DEConv).

are applied balancing coefficient to reduce its influence on the final feature, as we assume that the information captured by the four DC kernels is significantly greater than that obtained from the single VC kernel. In detail, the balancing procedure is implemented as follows:

$$F_{out} = F_{vc} - \theta F_{dc} \quad (9)$$

where θ denotes the balancing coefficient, which is utilized to adjust the relative contributions of the information derived from the DC process and VC process. By balancing between F_{dc} and F_{vc} , we ensure that the feature maps maintain a balance between gradient-level and intensity-level information. After training is completed, the learned kernel weights of the trained model are summed at their corresponding positions to reduce the overall number of parameters while preserving the performance. However, due to the modification of the original strategy employed in DEConv, it is necessary to adapt these changes accordingly. The details of this process are illustrated as follows:

$$\begin{aligned} K_{final} &= K_{VC} - \theta(K_{cdc} + K_{hdc} + K_{vdc} + K_{adc}) \\ &= K_{VC} - \theta K_{AGG} \end{aligned} \quad (10)$$

where K_{cdc} , K_{hdc} , K_{vdc} , and K_{adc} denote the convolutional kernels corresponding to the CDC, HDC, VDC, and ADC, respectively. K_{AGG} is the aggregated kernel of DCs while K_{VC} represents the VC kernel. K_{final} refers to the final aggregated kernel applied directly to the unaltered input, and θ represents the balancing coefficient. Final bias for K_{final} is created using the same process as that for generating K_{final} .

By integrating the modifications detailed in Sec. III-A and Sec. III-B, along with the novel strategy for combining DC and VC, we address the challenges that exist in the DEA-Net model. Furthermore, these modifications contribute to the development of a new model, named DEA-Net-CR-LapSE. This model not only enhances the performance of DEA-Net across multiple datasets and metrics but also ensures the preservation of a low parameter count.

IV. EXPERIMENTAL RESULTS

A. Datasets and Metrics

1) *Datasets*: The Realistic Single Image Dehazing (RESIDE) [17] dataset is used as a comprehensive benchmark for advancing research in the field of the single image dehazing, structured into five distinct subsets. To ensure a fair comparison between baseline and its modified version, DEA-Net-CR-LapSE, we will utilize three subsets from the RESIDE dataset, including Indoor Training Set (ITS), Outdoor Training Set (OTS), and Synthetic Objective Testing Set (SOTS). The ITS subset contains 1399 clear images, with each accompanied by 10 corresponding hazy images. For the OTS subset, we eliminated duplicate images that appear in both training and testing, as in the [18]. In addition, we also remove images with dimensions smaller than 256 pixels in either length or width. This refinement results in a total of 296,520 hazy images in the OTS subset. The ITS and OTS subsets will be employed for training the model under indoor and outdoor conditions, respectively. For evaluation, we will use the SOTS subset, which is divided into two distinct components corresponding to the indoor and outdoor training conditions: SOTS-Indoor (500 hazy images) and SOTS-Outdoor (500 hazy images). Apart from the RESIDE [17] dataset, the Haze4K [19] dataset is utilized to evaluate the models' performance. This dataset comprises a total of 4,000 hazy images, with 3,000 images designated for training and 1,000 images used for testing.

2) *Metrics*: The DEA-Net [2] model utilizes two commonly used metrics, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measurement (SSIM) [20], to evaluate the similarity between the dehazed image and its corresponding clear version. However, these measures fail to fully capture the human visual perception. In particular, SSIM and PSNR primarily focus on pixel-level discrepancies, often overlooking the broader content, structure, and perceptual qualities of an image. As a result, they are less sensitive to factors such as texture distortion, style inconsistencies, and other visual elements that are crucial for human observers. To more accurately assess the similarity, we incorporate the Learned Perceptual Image Patch Similarity (LPIPS) [7] metric alongside PSNR and SSIM. By utilizing features extracted

from deep neural networks, LPIPS can capture and compare high-level information such as textures, structures, and patterns that are more aligned with human visual perception. This characteristic enables LPIPS to perform a more human-like evaluation of image similarity, considering perceptual aspects that traditional pixel-based metrics might ignore.

B. Experimental settings

The DEA-Net-CR-LapSE model was trained on a single computer equipped with an NVIDIA GeForce RTX 3090 GPU and running the Windows 11 operating system. The model's implementation is conducted using the PyTorch deep learning framework. Data augmentation techniques included random cropping followed by random rotations at angles of 0, 90, 180, or 270 degrees. The model is trained for a total of 300 epochs, with each epoch consisting of 5,000 iterations (resulting in a total of 1,500,000 training steps), and the batch size is 16. The coefficients for L1 loss, Contrast Regularization (CR) [13], and Laplacian Pyramid Loss [4] [5] are 1.0, 0.1, and 0.8, respectively. The θ used to balance DC and VC is 0.7. All other training parameters and settings remained consistent with those established for the DEA-Net-CR model¹.

C. Experimental results

The proposed DEA-Net-CR-LapSE model is quantitatively evaluated against its foundational counterpart, DEA-Net-CR. We selected DEA-Net-CR as the primary baseline over DEA-Net because DEA-Net-CR consistently demonstrated superior performance across the majority of evaluation metrics and all datasets. Furthermore, to indicate the performance of our method, it is compared against other current high-performance dehazing algorithms, including MSBDN [14], Dehamer [15], FFA-Net [11], and PMDNet [16]. It is important to note that the evaluation results for these four methods are presented in [2]. For comparative analysis, the results of the DEA-Net-CR model on the Haze4K and SOTS-Indoor datasets are derived from two separate trained models. These models were trained on the Haze4K and ITS datasets, respectively, and are made available on the same link as the source code. Additionally, the DEA-Net-CR will be trained on the OTS dataset and evaluated on the SOTS-Outdoor dataset to ensure a fair comparison because we modified the OTS dataset. Note that the best-performing results are highlighted in bold, and the second-best are underlined in all following tables.

The detailed quantitative results of these methods are presented in Table I. Analysis of the quantitative results reveals that both DEA-Net-CR and the proposed DEA-Net-CR-LapSE models exhibit substantially superior performance in terms of PSNR and SSIM metrics compared to other competitive methods across all three benchmark datasets. Specifically, on the SOTS-Outdoor dataset, DEA-Net-CR-LapSE achieves a PSNR of 36.7875 dB, representing an approximate 1.8% improvement over the result of DEA-Net-CR (36.1294 dB). This result also surpasses the PSNR value of 36.59 dB reported for DEA-Net-CR on SOTS-outdoor in the original paper by a margin of

¹Source code of DEA-Net: <https://github.com/cecret3350/DEA-Net>

TABLE I: Quantitative evaluation results of several dehazing methods across Haze4K, SOTS-Indoor, and SOTS-Outdoor. Performance is assessed using PSNR and SSIM metrics. A hyphen (-) signifies missing data.

Method	Haze4K		SOTS-Indoor		SOTS-Outdoor	
	PSNR $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$
MSBDN [14]	22.99	0.85	32.77	0.9812	34.81	0.9857
Dehamer [15]	-	-	36.63	0.9881	35.18	0.9860
FFA-Net [11]	26.97	0.95	36.39	0.9886	33.57	0.9840
PMDNet [16]	33.49	0.98	38.41	0.9900	34.74	0.9850
DEA-Net-CR [2]	34.25	0.99	41.31	0.9945	36.13	0.9888
Ours	34.34	0.99	41.47	0.9947	36.79	0.9901

0.55%. Similarly, on the Haze4K and SOTS-Indoor datasets, our proposal demonstrates superior PSNR values of 34.3411 dB and 41.4689 dB, respectively, exceeding DEA-Net-CR's scores of 34.2546 dB and 41.3084 dB. Regarding the SSIM metric, the proposed model consistently outperforms DEA-Net-CR across all datasets. These improvements emphasize the effectiveness of the architectural refinements and modifications applied to the DEA-Net model. While other comparative methods, such as MSBDN, Dehamer, FFA-Net, and PMDNet, demonstrate potential performance, the proposed approach consistently maintains a significant performance lead. For instance, our method significantly outperforms the FFA-Net by approximately 27.3% in terms of PSNR and about 4.2% in SSIM. Note that in Table I, PSNR values are rounded to two decimal places and SSIM values to four decimals, except for Haze4K, SSIM value are rounded to two decimals for consistency. In general, the results presented in Table I demonstrate that the proposed method yields superior performance compared to its baseline, DEA-Net-CR, and other techniques. This demonstrates that our modifications are effective and successfully address the remaining limitations of the DEA-Net model.

TABLE II: Comparative evaluation of the DEA-Net-CR and DEA-Net-CR-LapSE model on HazeK, SOTS-Indoor, and SOTS-Outdoor, measured by the LPIPS.

Method	LPIPS $\downarrow$		
	Haze4K	SOTS-Indoor	SOTS-Outdoor
DEA-Net-CR [2]	0.00777	0.00277	0.00508
Ours	0.00750	0.00270	0.00455

Table II presents a comprehensive comparison of the Learned Perceptual Image Patch Similarity (LPIPS) metric between two distinct models: DEA-Net-CR and DEA-Net-CR-LapSE. The LPIPS values are rounded to five decimals. In detail, DEA-Net-CR-LapSE demonstrated performance enhancements, with LPIPS values reduced by 3.5%, 2.5%, and 10.4% compared to DEA-Net-CR on Haze4K, SOTS-Indoor, and SOTS-Outdoor, respectively. It is crucial to note that in the context of image quality assessment, a lower LPIPS value indicates a higher degree of perceptual similarity between the dehazed output and the corresponding ground truth image. This metric is particularly valuable as it aligns more closely with human visual perception of image quality than traditional pixel-wise error metrics. Therefore, the consistent reduction

in LPIPS observed with DEA-Net-CR-LapSE indicates that its dehazing outputs are more visually identical with the true underlying scenes, thereby offering a more faithful dehazing of image clarity.

TABLE III: Assessment of the number of parameters of DEA-Net-CR and DEA-Net-CR-LapSE with and without the re-parameterization (re-pa.) technique.

Method	Number of Parameters $\downarrow$	
	w/ re-pa.	w/o re-pa.
DEA-Net-CR [2]	3653377	7789313
Ours	3652001	7787937

As detailed in Table III, the DEA-Net-CR-LapSE model not only outperforms DEA-Net-CR on all three datasets across all three evaluation metrics (PSNR, SSIM, and LPIPS), but it also achieves this with a reduced number of parameters. When the re-parameterization technique is applied, DEA-Net-CR-LapSE utilizes 3,652,001 parameters compared to DEA-Net-CR's 3,653,377. Without re-parameterization, the difference remains clear, with DEA-Net-CR-LapSE at 7,787,937 parameters versus DEA-Net-CR's 7,789,313. In both scenarios, DEA-Net-CR-LapSE demonstrates a reduction of 1,376 parameters. This reduction in parameters demonstrates a more concise and potentially more efficient model architecture. In addition, Fig. 3 illustrates the model's qualitative performance by comparing it with the ground truth on the Haze4K dataset.

D. Ablation Study

We sequentially apply modifications to the DEA-Net-CR model to evaluate their impact, with results presented in Table IV. The PSNR values are rounded to three decimals, while SSIM and LPIPS values are rounded to five decimals. These modified versions are trained and evaluated on the Haze4K dataset for 225,000 steps, with each adjustment's configuration outlined in Sec. IV-B. Notably, enhancements to the total loss function (DEA-Net-CR_LPL) and channel attention mechanism (DEA-Net-CR_SE) yield superior performance across all metrics. Specifically, the DEA-Net-CR_LPL achieves the highest SSIM value (0.98703) while the DEA-Net-CR_SE obtains the second-highest result in the LPIPS metric (0.00886). Note that the LPL balancing coefficient is set to 1.0 to better highlight impact of LPL in the DEA-Net-CR_LPL model. The combination of both adjustments in the loss function and channel attention mechanism (DEA-Net-CR_LPL_SE) further enhances performance in the PSNR



Fig. 3: Comparison between hazy image, our model outputs, and ground truth on Haze4K dataset.

TABLE IV: Ablation study of modifications applied to DEA-Net-CR on Haze4K dataset.

Method	Haze4K		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
DEA-Net-CR [2]	33.075	0.98634	0.00894
DEA-Net-CR_LPL	33.117	0.98703	0.00896
DEA-Net-CR_SE	33.140	0.98654	0.00886
DEA-Net-CR_LPL_SE	33.151	0.98697	0.00893
DEA-Net-CR-LapSE	33.195	0.98700	0.00881

metric. By integrating all modifications, DEA-Net-CR-LapSE outperforms all other variants, achieving the best results in PSNR (33.195) and LPIPS (0.00881), confirming the effectiveness of our proposed modifications.

V. CONCLUSION

In this study, we address challenges in image dehazing, such as inefficiencies in the loss function, underutilization of channel relationship information, and imbalances in the integration of Difference Convolutions (DCs) and Vanilla Convolution (VC) strategies, with a particular focus on the DEA-Net model. By incorporating Laplacian Pyramid Loss (LPL), refining the attention mechanism, and optimizing the DC-VC combination, we introduce the DEA-Net-CR-LapSE model. These modifications effectively enhance not only the learning process but also the quality of feature maps. Experimental results on the RESIDE and Haze4K datasets show that DEA-Net-CR-LapSE outperforms baseline methods and other approaches across three metrics: PSNR, SSIM, and LPIPS. Future work will focus on improving spatial attention, particularly for bright pixels, where the performance of model

is currently less optimal. In addition, reducing training time is also a key goal for further optimization.

REFERENCES

- [1] S. Narasimhan and S. Nayar, “Contrast restoration of weather degraded images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 713–724, 2003.
- [2] Z. Chen, Z. He, and Z.-M. Lu, “Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention,” *IEEE Transactions on Image Processing*, vol. 33, pp. 1002–1015, 2024.
- [3] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, “Searching central difference convolutional networks for face anti-spoofing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [4] S. Li, X. Xu, L. Nie, and T.-S. Chua, “Laplacian-steered neural style transfer,” in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1716–1724, 2017.
- [5] P. Bojanowski, A. Joulin, D. Lopez-Pas, and A. Szlam, “Optimizing the latent space of generative networks,” in *Proceedings of the 35th International Conference on Machine Learning (J. Dy and A. Krause, eds.)*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 600–609, PMLR, 10–15 Jul 2018.
- [6] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [8] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.
- [9] R. Fattal, “Dehazing using color-lines,” *ACM transactions on graphics (TOG)*, vol. 34, no. 1, pp. 1–14, 2014.
- [10] H. Zhang and V. M. Patel, “Densely connected pyramid dehazing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3194–3203, 2018.
- [11] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, “Ffa-net: Feature fusion attention network for single image dehazing,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 11908–11915, 2020.
- [12] Y. Song, Z. He, H. Qian, and X. Du, “Vision transformers for single image dehazing,” *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, 2023.
- [13] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma, “Contrastive learning for compact single image dehazing,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10551–10560, 2021.
- [14] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, “Multi-scale boosted dehazing network with dense feature fusion,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2157–2167, 2020.
- [15] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, “Image dehazing transformer with transmission-aware 3d position embedding,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5812–5820, 2022.
- [16] T. Ye, Y. Zhang, M. Jiang, L. Chen, Y. Liu, S. Chen, and E. Chen, “Perceiving and modeling density for image dehazing,” in *European conference on computer vision*, pp. 130–145, Springer, 2022.
- [17] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, “Benchmarking single-image dehazing and beyond,” *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2019.
- [18] X. Liu, Y. Ma, Z. Shi, and J. Chen, “Griddehazenet: Attention-based multi-scale network for image dehazing,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7314–7323, 2019.
- [19] Y. Liu, L. Zhu, S. Pei, H. Fu, J. Qin, Q. Zhang, L. Wan, and W. Feng, “From synthetic to real: Image dehazing collaborating with unlabeled real data,” in *Proceedings of the 29th ACM international conference on multimedia*, pp. 50–58, 2021.
- [20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

Enhancing Micro-Expression Recognition via Multi-Task Learning with Demographic-Aware Auxiliary Supervision

Hai Tien Trinh
Faculty of Information Technology
VNU University of
Engineering and Technology
Hanoi, Vietnam
trinhhaitien03@gmail.com

Thi Bich Phuong Man
Faculty of Information Technology
VNU University of
Engineering and Technology
Hanoi, Vietnam
phuongman2004@gmail.com

Vu Tram Anh Khuong
Faculty of Information Technology
VNU University of
Engineering and Technology
Hanoi, Vietnam
21020104@vnu.edu.vn

Luu Tu Nguyen
Faculty of Information Technology
VNU University of
Engineering and Technology
Hanoi, Vietnam
nltu@vnu.edu.vn

Thanh Ha Le
Faculty of Information Technology
VNU University of
Engineering and Technology
Hanoi, Vietnam
ltha@vnu.edu.vn

Thi Duyen Ngo^{*}
Faculty of Information Technology
VNU University of
Engineering and Technology
Hanoi, Vietnam
duyennt@vnu.edu.vn

Abstract—Facial micro-expressions are rapid, involuntary facial movements that can reveal hidden emotions, offering crucial cues for applications in mental health, security, and affective computing. While multi-task learning (MTL) has shown promise in related vision tasks, its application to micro-expression recognition (MER) remains limited. Motivated by psychological studies showing that demographic factors such as race and gender influence how emotions are expressed, this study is the first to explore the role of racial information in an MTL-based MER framework. By incorporating race and gender classification as auxiliary tasks, the model benefits from shared representations that capture both individual-specific and group-specific expressive cues. Evaluations on two publicly available, multi-ethnic datasets, SAMM and SMIC-HS, demonstrate that adding race supervision leads to consistent improvements in MER performance, with gains of up to 10%. These results highlight the potential of demographically-informed learning in enhancing the robustness and fairness of emotion recognition systems.

Index Terms—micro-expression, micro-expression recognition, multi-task learning, demographic information

I. INTRODUCTION

In recent years, emotion recognition has been receiving increased academic and industry interest [1]. Emotion recognition supports applications across crucial industries such as healthcare, education, marketing, public safety, and business strategy. While emotion can be expressed in multiple ways, facial micro-expression stands out as it appears involuntarily, regardless of conscious effort [2]. A facial micro-expression is a brief facial movement, typically lasting for less than half a second [3]. The onset phase of a facial micro-expression can complete in under 260ms [3], and its intensity is subtle

enough to obscure itself in casual human-human interaction. Interpreting and classifying those expressions manually is inherently subjective, as humans perceive emotions in different ways. Fortunately, advances in computer vision and machine learning now enable us not only to capture facial micro-expressions, but also to label them objectively [4].

There have been multiple approaches to micro-expression recognition (MER). Traditional method, such as LBP [5] encode local temporal patterns into a compact code and classify them with SVM. More recently, deep learning methods have demonstrated even stronger performance [6]. To further boost accuracy on video-based MER, researchers have incorporated auxiliary factors, most commonly gender, into a multi-task learning framework [7]. However, other demographic attributes, such as race, are often ignored due to limited annotations. Psychological research provides strong evidence that demographic factors, such as gender, cultural background and race, significantly modulate the intensity and masking of facial micro-expressions [8]–[10]. For example, Japanese, Americans and Russians display different micro-expression intensity under identical conditions [9], and the activation patterns of facial action units during micro-expression vary across gender and racial group. These insights have inspired the use of self-labeling techniques and incorporation of racial information to enhance the performance of MER, highlighting the urgent need to investigate how demographic factors influence expressive behavior in MER research.

To integrate demographic attributes into the main task of MER, deep learning gives a concept of multi-task learning, which extends a single network predicting one task with auxiliary objectives and jointly optimizes them, allowing the

^{*}Corresponding author (duyennt@vnu.edu.vn)

primary task to learn more generalizable features [11]. In practice, gender has been selected as an auxiliary task due to its straightforward annotation, which has led to improved performance in several multi-task models [7], [12]. By contrast, more nuanced demographics, such as racial identity, introduce greater labeling complexity and remain less explored.

This study introduces a novel multi-task learning framework that incorporates both gender and race as auxiliary tasks to enhance micro-expression recognition. To the best of our knowledge, this is the first work to systematically analyze the impact of race on MER performance. To support this investigation, we adopt and expand a multi-task learning model to explicitly incorporate race information, and demonstrate its effectiveness through experimental evaluation. Experimental results on two benchmark datasets demonstrate that adding race as an auxiliary signal leads to consistent improvements in classification accuracy and generalization across demographic groups. These findings suggest that incorporating demographic supervision can play a significant role in building more effective and inclusive micro-expression recognition systems.

The remainder of this paper is organized as follows: Section II provides an overview of existing research in micro-expression and multi-task learning. Section III describes our method in details. Next, Section IV presents our experimental setup and discusses the results. Finally, Section V concludes the paper and discusses future directions.

II. RELATED WORK

Micro-expression recognition aims to classify the underlying concealed emotions. Early MER methods approaching these tasks relied on handcrafted spatiotemporal descriptors paired with classical classifiers. For instance, LBP-TOP extends Local Binary Patterns to three planes to capture changes in facial textures across frames, or the Bi-WOOF uses a bi-weighted oriented optical flow descriptor computed on apex and onset frames [5], [13]. These predefined feature frameworks were typically combined with support vector machine (SVM) to perform the final classification. However, these manual pipelines often lack the flexibility to adapt to wide variations and different datasets. As a result, recent research has shifted to deep learning solutions that automatically learn patterns from raw inputs. TSCNN employs a triple stream 3D CNN to jointly process optical flow, horizontal flow, and vertical flow representations to capture facial movement [14]. OrigNet fine-tunes a CNN pretrained on large datasets using static apex frames to capture subtle micro expression cues [15].

In recent years, researchers have begun integrating demographic attributes into MER deep learning models to enable them to learn additional contextual information. This can be done by employing a variety of multi-task learning techniques, including hard parameter sharing, soft parameter sharing, cross-stitch networks or sluice architecture. The authors of [12] proposed GEME, a Multi-task learning network with primary task of recognizing emotions and subtask of classifying gender. GEME first processes inputs through a unified early feature extractor, then injects gender cues into the emotion pipeline via

a gender branch whose parameters are synchronized with those of the emotion stream. In GEME’s evaluations on benchmark micro-expression datasets, incorporating gender cues boosted recognition accuracy by up to 5 %. This raises the question of whether integrating other demographic cues, specifically race, could further boost MER performance, potentially exceeding the improvements achieved through emotion only and gender-base augmentation baselines.

The inclusion of racial information is supported by psychological research. Ekman et al. [8] reported that, in a spontaneous expression study, Asian participants masked negative expressions significantly more often than American participants. In a broad survey of 32 countries, researchers confirmed systematic differences in expressive intensity across cultural, and often racial, groups [9]. Using Facial Action Coding System on a large dataset, it was found that African American subjects displayed higher Cheek Raiser and Lip Corner Puller intensities when smiling than Asian subjects [10]. In facial emotion recognition (FER), the paper et al. [16] fuses features from a pre-trained racial identity network with expression features and achieves noticeable increase in accuracy. Altogether, these findings demonstrate that gender, cultural background and race can modulate the intensity and reactivity of micro-expression during emotional displays, and thus holds promise for boosting MER model performance.

III. PROPOSED METHOD

This paper presents a multi-task learning framework that models and exploits demographic attributes alongside micro-expression cues to improve performance, and analyzes whether integrating racial information into this framework further boosts micro-expression recognition performance.

The pipeline is shown in Figure 1. First, faces in each sequence are detected, cropped and resized; demographic information is derived via a pre-processing pipeline and manually verified for use as auxiliary targets. Next, each video is compressed into a single dynamic image to capture temporal dynamics. Finally, a multi-task network takes the dynamic image and jointly predicts emotion alongside auxiliary tasks.

A. Pre-processing

The pre-processing pipeline aims to extract standardized and reliable facial inputs from raw video frames for downstream learning tasks. It includes two main stages: facial region extraction and demographic label assignment.

First, each video frame is processed using a pretrained frontal face detector from the DLib library [17], which is known for its robustness under varying illumination, pose, and background conditions. Upon detecting the face, 68 facial landmarks [18] are extracted to identify key structural points, such as the eyes, nose, and mouth. These landmarks are used to construct a tight bounding box around the facial region of interest. The cropped face is then resized to a fixed resolution of 112×112 pixels to satisfy model input requirements.

Since publicly available micro-expression datasets lack demographic annotations, race and gender labels are automatically inferred using the FairFace classifier [19]. For each

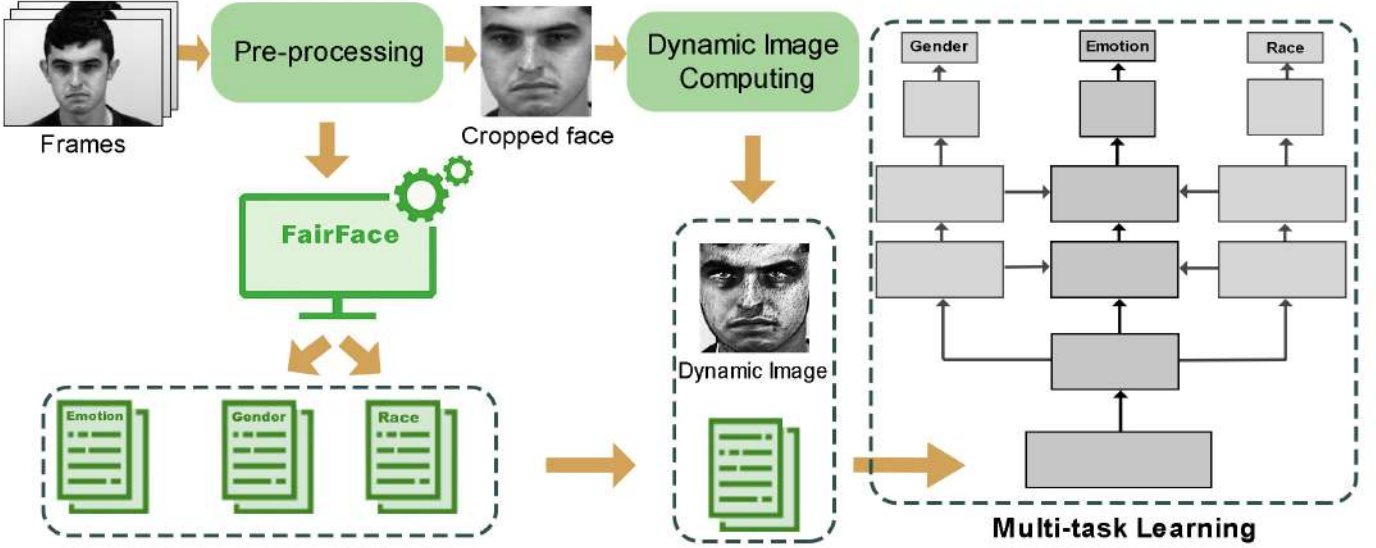


Fig. 1. The proposed multi-task learning pipeline for micro-expression recognition, combining dynamic appearance modeling with demographic supervision

subject, a representative frame (typically the first) is resized and normalized to meet FairFace’s input specifications, then passed through the pretrained model. The predicted race (White, Asian, Black, Indian) and gender (male, female) are used as supervisory signals in auxiliary tasks during training. To ensure label reliability, all predicted demographic attributes are manually verified and corrected when necessary. This step mitigates potential misclassification in low-quality or ambiguous cases and ensures higher annotation fidelity for demographic-aware learning.

B. Dynamic image computing

Better than the apex frame (which corresponds to the moment of peak facial muscle activation), dynamic imaging compresses a video sequence into a single image that preserves temporal ordering and motion evolution. Originally developed for action recognition, dynamic images have proven effective in compressing the dynamics of motion sequences into a static yet informative form. For micro-expression sequences, this representation is particularly useful because micro-expressions are short-lived and subtle, making aggregated temporal motion more informative than individual frames.

The dynamic image construction is built upon the rank pooling framework introduced by Fernando *et al.* [20], [21]. Given a sequence of T frames $F_1, F_2, \dots, F_T$, each frame is mapped to a feature representation $\psi(F_t)$. The temporal evolution is encoded into a parameter vector d by solving the following convex optimization problem:

$$d^* = \arg \min_d \left(\frac{\lambda}{2} \|d\|^2 + \frac{2}{T(T-1)} \sum_{q>t} \max(0, 1 - \langle d, V_q - V_t \rangle) \right) \quad (1)$$

where V_t denotes the time-averaged feature vector up to frame t . The resulting vector d^* aggregates the temporal structure of the entire sequence and can be directly interpreted as a dynamic image.

To reduce computational overhead, Bilen *et al.* [22] proposed an efficient approximation known as Approximate Rank Pooling (ARP). Instead of solving Equation 1 exactly, ARP estimates d^* as a weighted linear combination of frame features in Eq. 2:

$$d^* = \sum_{t=1}^T \alpha_t \psi(F_t), \quad (2)$$

where the temporal weights α_t are defined as follows:

$$\alpha_t = 2t - T - 1. \quad (3)$$

This approximation naturally biases the representation toward later frames, effectively capturing the accumulated motion over time. For micro-expression sequences, where the most salient facial movements often occur near the apex, this weighting is particularly advantageous.

C. Multi-task learning

This paper designs a multi-task learning model that jointly performs micro-expression recognition and demographic attribute classification. Compared with prior work [12], our model introduces configurable auxiliary branches and integrates a more systematic demographic annotation strategy.

The model consists of three key components. First, dynamic images are passed through a shared convolutional backbone that extracts general-purpose visual features. These shared features are then propagated to multiple task-specific branches. The main branch is dedicated to micro-expression classification, while optional auxiliary branches handle race and gender

prediction. Each branch ends with an independent classifier, and the entire network is optimized using a joint loss function that balances the contributions of all active tasks. This modular structure allows flexible adaptation: auxiliary tasks can be selectively enabled to regularize the feature space and improve generalization, or disabled when demographic supervision is unavailable. By leveraging multi-task learning, the model encourages shared representations that are both discriminative for emotion recognition and sensitive to demographic cues.

Specifically, the dynamic image is in RGB format with dimensions of $112 \times 112 \times 3$. A hard parameter sharing module with two CBR blocks (Convolution, Batch normalization and ReLU) that each double the number of channels is applied to produce common feature maps for subsequent branch splitting. From here, three branches for specific tasks are created in parallel: a primary branch for micro-expression recognition, followed by auxiliary branches for gender and race. The shared feature maps generated by the hard parameter sharing module serve as inputs to each branch. In each branch, four ResNet blocks are applied in sequence, each again doubling the feature channels. During this stage, soft parameter sharing is implemented by adding features from auxiliary tasks to the main task at each corresponding step, while leaving the auxiliary maps unchanged. Finally, each branch's output is flattened, passed through a dropout layer, reduced dimension to yield class probabilities. Each auxiliary branch is trained by its own loss function, and the loss of primary branch is influenced by these auxiliary losses through joint optimization.

a) Losses: Due to the imbalance of training datasets, different loss functions are implemented to match each task's properties. For the emotion stream and race stream, which have relatively imbalanced class distribution, focal loss is used to prevent domination of the majority class. The focal loss, introduced by Lin *et al.* [23], is designed to address class imbalance by down-weighting well-classified examples. Its formulation is given in Eq. 4:

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad (4)$$

where p_t denotes the predicted probability for the ground-truth class, α_t is a class-specific weighting factor based on class frequency, and γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted.

In contrast, for the gender classification stream, where the class distribution is approximately balanced, we adopt the standard cross-entropy loss, as shown in Eq. 5:

$$\text{CE}(p_t) = -\log(p_t). \quad (5)$$

The training objective combines the losses from emotion, gender, and race classification, depending on the configuration of the multi-stream architecture. This is summarized in Eq. 6:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{emotion}} + \underbrace{0.5 \times \mathcal{L}_{\text{gender}} + 0.5 \times \mathcal{L}_{\text{race}}}_{(\text{if enabled})}. \quad (6)$$

IV. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the proposed multi-task learning framework that leverages racial demographic information for MER, we conduct experiments on multi-race micro-expression datasets. The goal is to assess whether integrating auxiliary attributes can improve emotion classification performance. Four experimental scenarios are compared: (1) a single-task baseline trained solely for emotion classification, (2) a dual-stream model jointly learning emotion and gender, (3) a dual-stream model jointly learning emotion and race, and (4) a multi-task model that simultaneously learns emotion, gender, and race. All models share the same backbone architecture, training strategy, and evaluation protocol to ensure a fair and controlled comparison.

A. Datasets

Experiments are conducted on two publicly available and widely used micro-expression datasets: the Spontaneous Actions and Micro-Movements (SAMM) dataset [24] and the Spontaneous Micro-expression Corpus (SMIC) dataset [25]. These datasets are selected due to their participant diversity in terms of both gender and ethnicity, as well as their extensive annotations of micro-expressions.

- **SAMM:** The SAMM dataset includes high-speed facial recordings from 29 participants, with 159 labeled micro-expression sequences annotated with onset, apex, and offset frames. It features a demographically diverse pool: 16 White, 7 Asian, 4 Indian, and 2 Black participants, and a near-equal gender distribution (15 males, 14 females). For experiments, only emotion classes with more than 10 samples are kept, resulting in five categories: anger, happiness, surprise, contempt, and others.
- **SMIC:** The SMIC dataset has three recording modalities, with this study focusing on the high-speed (SMIC-HS) at 100 fps. It contains 164 micro-expression clips from 16 subjects (10 males, 6 females), evenly split by ethnicity (8 White, 8 Asian). Clips are categorized into three emotion classes: positive, negative, and surprise. As ground-truth apex frames are not provided, the apex index is approximated as the midpoint between onset and offset frames.

B. Evaluation Protocol

To assess the generalization capability of the proposed method, two widely adopted evaluation strategies in micro-expression recognition are employed: k-fold cross-validation and Leave-One-Subject-Out (LOSO) validation.

In the k-fold cross-validation setting ($k = 5$), the dataset is randomly partitioned into five equal subsets. In each iteration, four folds are used for training while the remaining one is used for testing. This repeats five times so each fold serves as the test set once. The final performance is reported as the average across all folds, which provides a reliable estimate of model performance under balanced data splits.

In the LOSO protocol, data from one subject is held out for testing while the model is trained on the remaining subjects. This process is repeated for each subject in the dataset, and the

TABLE I
COMPARISON OF AUXILIARY TASK CONFIGURATIONS FOR MER UNDER 5-FOLD CROSS VALIDATION

Input	Method	Accuracy		UF1	
		SAMM	SMIC	SAMM	SMIC
Apex Frame	One stream (Emotion)	55.15	59.51	37.14	57.05
Dynamic Image		53.73	57.75	44.87	56.91
Apex Frame	Two streams (Emotion, Gender)	56.61	61.96	45.04	60.34
Dynamic Image		57.33	62.84	38.02	63.80
Apex Frame	Two streams (Emotion, Race)	59.60	63.19	48.54	61.85
Dynamic Image		60.37	64.42	44.83	63.69
Apex Frame	Three streams (Emotion, Gender, Race)	59.60	68.10	44.78	48.80
Dynamic Image		60.34	67.48	45.05	66.89

TABLE II
COMPARISON OF AUXILIARY TASK CONFIGURATIONS FOR MER UNDER LOSO

Input	Method	Accuracy		UF1	
		SAMM	SMIC	SAMM	SMIC
Apex Frame	One stream (Emotion)	49.18	50.31	29.05	50.18
Dynamic Image		50.77	52.15	31.89	52.17
Apex Frame	Two streams (Emotion, Gender)	50.77	57.06	31.30	56.84
Dynamic Image		51.47	57.06	32.12	57.60
Apex Frame	Two streams (Emotion, Race)	51.47	58.19	30.93	56.52
Dynamic Image		53.33	58.28	38.49	58.72
Apex Frame	Three streams (Emotion, Gender, Race)	55.88	58.90	35.68	58.07
Dynamic Image		54.41	63.80	42.87	63.76

overall performance is averaged across all iterations. LOSO is particularly suitable for evaluating subject-independent generalization, which is critical for real-world deployment where the model encounters unseen individuals.

C. Evaluation Metrics

The performance of our proposed method is assessed using two standard metrics:

- **Accuracy:** This measures the overall proportion of correctly classified samples.

$$\text{Acc} = \frac{\text{Number of correct predictions}}{\text{Total number of samples}} \quad (7)$$

In 5-fold cross-validation, the final accuracy is the average of accuracies across all folds. For LOSO, due to the unequal sample sizes between subjects, the overall accuracy is calculated as follows:

$$\text{Acc}_{\text{LOSO}} = \frac{\sum_{i=1}^S T_i}{\sum_{i=1}^S N_i} \quad (8)$$

where S is the number of subjects, T_i is the number of correctly predicted samples for subject i , and N_i is the total number of samples from subject i .

- **Unweighted F1-score (UF1):** To mitigate the effect of class imbalance, UF1 computes the mean F1-score across

all classes, treating each class equally regardless of its frequency.

$$\text{F1}_c = \frac{2 \cdot \text{TP}_c}{2 \cdot \text{TP}_c + \text{FP}_c + \text{FN}_c} \quad (9)$$

$$\text{UF1} = \frac{1}{C} \sum_{c=1}^C \text{F1}_c \quad (10)$$

where TP_c , FP_c , and FN_c are the true positives, false positives, and false negatives for class c ; C is the total number of emotion classes.

D. Implementation Details

All models are implemented in PyTorch [26]. Input images are first resized to 112×112 pixels, histogram equalized, and augmented using rotations from -15° to $+15^\circ$ in 5° steps, along with vertical flipping. Training is conducted for 20 epochs using stochastic gradient descent with an initial learning rate of 0.001, momentum of 0.9, weight decay of 0.001, and a batch size of 64. A dropout rate of 0.3 is applied after each fully connected layer to prevent overfitting.

The gender classification stream uses cross-entropy loss, while focal loss is applied for race classification on the SAMM dataset to address class imbalance, and cross-entropy is used for the more balanced SMIC dataset.

E. Results

In multi-task learning, selecting feature combinations and auxiliary tasks is a critical decision since poor choice can reduce efficiency or fail to improve performance. To assess

if adding a race classification branch to the single-task framework leads to more effective feature learning and how much it boosts accuracy, experiments are conducted under four settings defined above. All models were trained under identical configurations and evaluated on Accuracy and UF1 using two input modalities: a single apex frame and a dynamic image. Results are summarized in Table I and II.

The results show that adding race as an auxiliary task improves performance over the emotion only baseline, and that incorporating both gender and race yields the highest accuracy. Notably, augmenting emotion recognition with race produces better results than with gender information alone, indicating that racial cues contribute more to MER performance than gender cues. This effect is highly evident on the SMIC database, where race labels are balanced, and multi-task models leveraging race information achieve the most gains. On the SAMM dataset, under the k-fold cross validation protocol, models augmented with race alone perform nearly identically to those augmented with both gender and race, likely because the way the folds are constructed allows racial information to implicitly capture much of the gender related variation.

The performance trend when incorporating an auxiliary objective based on race conditioning is consistent across input modalities: apex frame and dynamic image. In addition, dynamic image inputs often produce better accuracy and UF1 than apex frame inputs.

V. CONCLUSION

This study is the first to explicitly incorporate race as an auxiliary task in multi-task learning for micro-expression recognition. By leveraging both race and gender as supervisory signals during training, the proposed framework is able to learn more representative and demographic facial features. Experimental results on SAMM and SMIC datasets reveal that introducing race information alone improves recognition performance by up to 10% compared to single-task baselines. Moreover, jointly modeling both race and gender leads to even greater improvements. These findings reinforce psychological and sociological evidence suggesting that demographic attributes significantly influence expressive patterns. Incorporating such attributes into the training process provides valuable inductive bias, enabling models to generalize better across diverse populations. This work opens new directions for building more versatile emotion recognition systems in heterogeneous real-world environments.

REFERENCES

- [1] R. Mobbs, D. Makris, and V. Argyriou, "Emotion recognition and generation: A comprehensive review of face, speech, and text modalities," 2025.
- [2] E. Rolls, P. Ekman, D. Perrett, and H. Ellis, "Facial expressions of emotion: An old controversy and new findings: Discussion," *Royal Society of London Philosophical Transactions Series B*, vol. 335, pp. 69–, 01 1992.
- [3] W.-J. Yan, Q. Wu, Y.-H. Chen, J. Liang, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, 12 2013.
- [4] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–23, 10 2022.
- [5] Y. Wang, H. Yu, B. Stevens, and H. Liu, "Dynamic facial expression recognition using local patch and lbp-top," in *2015 8th International Conference on Human System Interaction (HSI)*, pp. 362–367, 2015.
- [6] N. G. Matharaarachchi and M. F. Pasha, "Micro-expression recognition using deep learning: A systematic review of literature," *SSRN Electronic Journal*, Nov. 2024.
- [7] X. Z. H. Z. X. D. Y. S. Chunlong Hu, Jianjun Chen, "Gender-specific multi-task micro-expression recognition using pyramid cgbp-top feature," *Computer Modeling in Engineering & Sciences*, vol. 118, no. 3, pp. 547–559, 2019.
- [8] P. Ekman, W. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. LeCompte, T. Pitcairn, P. Ricci Bitti, K. Scherer, M. Tomita, and A. Tzavaras, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of personality and social psychology*, vol. 53, pp. 712–7, 10 1987.
- [9] D. Matsumoto, S. H. Yoo, S. Hirayama, and G. Petrova, "Development and validation of a measure of display rule knowledge: The display rule assessment inventory," *Emotion*, vol. 5, pp. 23–40, 03 2005.
- [10] Y. Fan, J. Lam, and V. Li, "Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness," *Scientific Reports*, vol. 11, 03 2021.
- [11] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, 07 1997.
- [12] X. Nie, M. Takalkar, M. Duan, H. Zhang, and M. Xu, "Geme: Dual-stream multi-task gender-based micro-expression recognition," *Neuro-computing*, vol. 427, pp. 13–28, 02 2021.
- [13] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, p. 82–92, Mar. 2018.
- [14] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. PP, pp. 1–1, 12 2019.
- [15] G. Verma and A. Dhall, "A non-ordinal approach to micro-expression recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (Snowmass Village, CO), pp. 2436–2445, February 2020.
- [16] M. Sohail, A. Ghulam, D.-J. Rashid, I. Ahmad, S. Almotiri, M. Al Ghamdi, A. Nagra, and K. Masood, "Racial identity-aware facial expression recognition using deep convolutional neural networks," *Applied Sciences*, vol. 12, p. 88, 12 2021.
- [17] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [18] Y. Qiu and Y. Wan, "Facial expression recognition based on landmarks," in *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 1356–1360, IEEE, 2019.
- [19] K. Kärkkäinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age," 2019.
- [20] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5378–5387, 2015.
- [21] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 773–787, 2017.
- [22] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3034–3042, 2016.
- [23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," pp. 2999–3007, 10 2017.
- [24] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2018.
- [25] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, 2013.
- [26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019.

A lightweight visual deception detection method using dynamic images and a customized SmallCNN–BiLSTM architecture

Thi Bich Phuong Man

Faculty of Information Technology
VNU University of Engineering and Technology
Hanoi, Vietnam
22021119@vnu.edu.vn

Luu Tu Nguyen

Faculty of Information Technology
VNU University of Engineering and Technology
Hanoi, Vietnam
nltu@vnu.edu.vn

Thanh Ha Le

Faculty of Information Technology
VNU University of Engineering and Technology
Hanoi, Vietnam
ltha@vnu.edu.vn

Thi Duyen Ngo*

Faculty of Information Technology
VNU University of Engineering and Technology
Hanoi, Vietnam
duyennt@vnu.edu.vn

Abstract—Deception detection has been a topic of interest for nearly a century and has applications in many practical areas such as law, security, and recruitment, where accurately assessing human behavior can lead to better decision-making. Traditional deception detection methods often rely on physiological signals, such as polygraphs, which are invasive and susceptible to bias due to individual differences in physiological characteristics. Meanwhile, visual data is a non-invasive source of information, easy to collect, and contains many non-verbal behavioral cues. Most current visual-based deception detection methods process data at the frame-level using features such as action units and eye gaze, which do not fully take advantage of temporal dynamics. This paper has proposed a lightweight and efficient deception detection method based on visual data, utilizing dynamic images and a SmallCNN–BiLSTM ensemble model for feature extraction and time-series modeling. Specifically, dynamic images help synthesize essential motion information into a single image, reduce noise from irrelevant small movements, and optimize computational cost compared to processing each frame individually. Bi-LSTM enables the capture of temporal dependencies in both directions (past and future), allowing the model to learn a more complete context to distinguish deceptive behavior. Experimental results show that the proposed method achieves competitive accuracy, while reducing computational costs and is suitable for deployment in real-world applications.

Index Terms—Deception detection, Bi-LSTM, CNN, Dynamic image

I. INTRODUCTION

Deception is the act of intentionally misleading or causing someone to believe in something untrue, through facial expressions and body gestures. Humans tend to use deception for various reasons. Some lies may be harmless and commonly occur in daily social interactions, not necessarily requiring detection. However, certain deceptive behaviors can lead to serious consequences if not detected in time, such as in law enforcement

(wrongful conviction of innocent individuals), border security (risking lives if dangerous criminals are allowed to enter), or in human resource recruitment [1]. Therefore, detecting high-risk deceptive behaviors is extremely important and should be addressed early. Research on deception detection has been of interest for nearly a century. In while, many studies show that even professionally trained experts such as police officers or psychological counselors detect deception with accuracy close to random chance [2] [3] [4].

Traditional deception detection methods use polygraphs based on physiological signals such as blood pressure, heart rate, brain activity, etc [5]. These methods come with various limitations: they require trained personnel, are invasive, and are prone to misjudgment (e.g., innocent individuals showing stress responses). These drawbacks have driven the development of non-invasive deception detection methods, such as extracting visual or audio features from a person's speech or appearance [6]. Moreover, when lying, people must exert more cognitive effort than when telling the truth, as they need to imagine and prepare content before speaking. This often unintentionally leads to cues such as facial tension, excessive blinking, pupil dilation, or unusual head movements.

With the rapid development of artificial intelligence, especially machine learning and deep learning in recent years, many studies have applied these techniques to deception detection using visual and audio features [6] [7] [8]. Current approaches mainly use input from video, audio, text, physiological signals such as EEG, or combine multiple modalities to increase accuracy [1] [9]. Each data type comes with its own limitations. Physiological data require specialized sensors and are unsuitable for natural environments. Text-based approaches rely on highly accurate speech-to-text systems and usually focus on English, which limits scalability to other languages. Audio data, though capable of providing features like pitch,

*Corresponding author is Thi Duyen Ngo (email: duyennt@vnu.edu.vn)

speech rate, and pauses, are easily affected by environmental noise and can be manipulated by the speaker to conceal their intentions. Multimodal fusion, while improving accuracy, requires synchronization across systems, costly preprocessing, multiple simultaneous sensors, and risks overfitting when data is insufficient.

Meanwhile, visual data is becoming increasingly popular, easy to collect, and rich in non-verbal behavioral information such as facial expressions, gaze, gestures, posture, and head movements. Therefore, it is a promising, efficient, and highly practical approach. Most current visual methods operate at the frame-level, using features like AUs, eye gaze, and head pose [10] [11]. This leads to high computational cost, vulnerability to noise from small irrelevant movements, and poor use of temporal evolution. Many methods only compute averages or extract a few representative frames, resulting in a loss of important temporal dynamics.

Based on these observations, this study proposes a lightweight visual deception detection method using dynamic images and a customized SmallCNN-BiLSTM architecture. Instead of processing individual frames, dynamic image (DI) compresses motion across a sequence into a single image, preserving key spatial-temporal features such as facial expressions and head or eye movement. A sequence of DIs is then passed to an Bi-LSTM to model behavioral dynamics over time. This is the first study to propose the use of DIs for the task of deception detection. This approach reduces computational cost, captures global motion effectively, minimizes irrelevant noise, and requires only standard RGB video. It is practical, non-invasive, and scalable for real-world applications.

The remaining of the paper is organized as follows: Section II reviews related work; Section III describes the proposed method; Section IV presents experiments and evaluation results; and Section V concludes the study.

II. RELATED WORKS

In recent years, researchers have focused on applying machine learning and deep learning techniques to deception detection using various types of data, such as video, audio, and physiological signals [6]. Audio-based methods often rely on features such as Mel spectrogram, MFCC, pitch, energy, or prosody to train classifiers like SVM, Random Forest, or CNN [9] [12]. These acoustic features are highly sensitive to environmental noise, microphone quality, and language variation. Furthermore, most existing research focuses on English, which makes it difficult to generalize to other languages.

Text data, typically transcribed from audio, is also used to analyze semantics and emotional content to detect deception, using features such as n-grams, LIWC, or pronoun usage. This approach depends on automatic speech recognition systems, which are prone to errors. Moreover, linguistic cues of deception are often subtle and ambiguous, making it difficult for models to detect lies based solely on text [13].

Physiological signals such as ECG, GSR, EEG, fMRI, or thermal imaging are also widely used in deception detection, as they can reflect neurological changes associated with lying.

However, collecting data like EEG requires expensive equipment that must be worn directly on the body, which can cause discomfort and limit practical deployment [14].

Among the various approaches, visual-based deception detection has emerged as a non-invasive and widely accessible alternative [5]. These methods analyze visual cues such as facial expressions, eye movements, and body gestures using images or videos to infer deceptive behavior [1]. Micro-expression analysis is a common technique where small and rapid muscle movements on the face are linked to underlying emotions or deceit. Eye gaze analysis including increased blink rate, gaze aversion, pupil dilation, and involuntary eye shifts can signal increased mental stress or cognitive load during deception.

Deep learning models such as 3D-CNN and ConvLSTM are commonly used to model and classify these spatiotemporal patterns in videos [1]. Some advanced systems also adopt multimodal fusion, where visual inputs are combined with audio or text features to improve detection performance by leveraging the complementary nature of different modalities.

Despite their effectiveness, visual-based deception detection approaches still face significant limitations that hinder real-world applicability. First, most models only focus on static frames or local visual features (frame-level) [10] [7], neglecting the temporal dynamics of behavior, which are crucial for distinguishing deceptive cues from normal expressions. This may cause the model to miss important behavioral patterns, such as averting gaze after a question or brief facial changes following a response, which only manifest across time.

Second, many video-based models process entire sequences or individual frames without temporal focus, leading to high computational cost, exposure to irrelevant motions, and difficulty attending to the most informative segments [12] [15]. Third, although models like 2D-CNN or 3D-CNN are effective in spatial feature extraction, they often fail to capture long-term temporal dependencies, especially without memory mechanisms such as LSTM or attention modules. This limits their ability to learn contextual relationships between behaviors that are temporally distant or sparsely distributed.

In addition to traditional hand-crafted and frame-level methods, recent studies have introduced advanced deep learning architectures for deception detection or micro-expression recognition. For example, Bilen et al. [16] and Fernando et al. [17] [18] proposed dynamic image representations for action recognition, suggesting their applicability to other temporal tasks. Gao et al. [11] recently employed capsule networks combined with attention mechanisms for video-based deception detection, while Karnati et al. [15] proposed LieNet, a deep CNN architecture for this task. Compared to these works, our method leverages dynamic images for deception detection, which, to the best of our knowledge, has not been explored before. Unlike heavy backbones such as ResNet or EfficientNet, we present a custom SmallCNN that is lightweight yet effective in representing motion information, combined with Bi-LSTM to capture bidirectional temporal dependencies. This design balances accuracy and computational cost, making it

suitable for real-time applications and resource-constrained environments.

III. METHOD

Due to the limitations of existing visual approaches, this study proposes a lightweight method for deception detection using DIs and SmallCNN-BiLSTM. The method compresses motion information, learns distinctive deceptive patterns, reduces computational cost, and captures temporal dependencies via BiLSTM. This makes it suitable for detecting subtle but important behavioral changes.

As shown in Fig. 1, the workflow includes: splitting video into fixed-length segments, cropping faces from frames, generating dynamic images, extracting features with SmallCNN, modeling temporal sequences using Bi-LSTM, and finally classification.

A. Video segmentation and Face cropping

Videos, typically 3–45 seconds at 30 fps, are divided into fixed-length segments to simplify processing and capture subtle, transient movements indicative of deception. Deceptive behaviors often appear in facial expressions, such as eye movements, raised eyebrows, or micro-expressions. Cropping the face removes background, hands, and clothing, enabling the CNN to focus on informative regions and learn efficiently with fewer resources. Computing dynamic images only on the facial region helps capture fine-grained motion patterns critical for distinguishing deceptive from truthful behavior. These preprocessing steps ensure the model input emphasizes the most relevant information, improving learning efficiency and detection.

B. Dynamic imaging

Previous studies often process video at the frame level, which incurs high computational costs, or simply extract a few representative keyframes for the entire video. This approach prevents models from learning motion-related features that change over time. This study proposes the use of DI to address these limitations, enabling the synthesis of a sequence of frames into a single image while preserving the temporal dynamics of motion.

A DI is a compact representation of a short video sequence, initially introduced and developed by Bilen [16] based on the concept of a ranking function, and widely adopted through the rank pooling technique proposed by Fernando [17] [18], to encode both motion and temporal ordering of frames into a single image. Building on this foundation, Bilen improved the efficiency of DI generation by computing directly on RGB pixel values and introducing approximate rank pooling (ARP), an approximate variant that significantly accelerates computation in practice.

In principle, Fernando consider a video sequence as a collection of frames $I_1, I_2, \dots, I_T$, where each frame is encoded into a feature vector $\psi(I_t)$. At each time step t , they compute the time-averaged feature vector, illustrated in Eq 1:

$$V_t = \frac{1}{t} \sum_{\tau=1}^t \psi(I_\tau), \quad (1)$$

and use a parameter vector $d \in \mathbb{R}^n$ to assign a score to each frame through the function as shown in Eq 2:

$$S(t|d) = \langle d, V_t \rangle. \quad (2)$$

The goal is to ensure that the scores reflect the temporal order of frames, meaning that frames appearing later should receive higher scores: $S(q|d) > S(t|d)$ if $q > t$. The optimization problem is defined to find the vector d^* that satisfies this ranking condition, given by Eq 3:

$$d^* = \arg \min_d \left(\frac{\lambda}{2} \|d\|^2 + \frac{2}{T(T-1)} \sum_{q>t} \max(0, 1 - S(q|d) + S(t|d)) \right) \quad (3)$$

The resulting vector d^* serves as a compact representation of the entire video and can be mapped back into an RGB image, referred to as the DI.

However, solving the above optimization problem is computationally expensive. To address this, Bilen proposed the ARP method, which allows direct estimation of d^* using a weighted sum of frame-level features, without solving the optimization. The ARP formulation is shown in Eq 4:

$$d^* = \sum_{t=1}^T \alpha_t \psi(I_t), \quad (4)$$

$$\alpha_t = 2t - T - 1. \quad (5)$$

The weights α_t in Eq 5 are designed to emphasize the temporal order, giving more importance to later frames. This method is both computationally efficient and capable of preserving the dynamic characteristics of the video sequence.

In our proposed method, we utilize the ARP approach to compute dynamic images for each segment consisting of a sequence of frames.

C. Feature extraction with SmallCNN

In this step, the DIs are processed by a CNN for feature extraction, then modeled temporally using a Bi-LSTM, and finally classified. This study proposes a lightweight CNN model, named SmallCNN, to extract features from DIs. The use of SmallCNN is motivated by the following considerations:

First, a DI temporal information from a video segment. Therefore, feature extraction from DIs does not necessarily require deep CNN models such as ResNet, EfficientNet, or MobileNet, which are primarily optimized for large-scale image classification tasks.

Second, deep networks typically involve a large number of parameters, resulting in high computational costs, longer training times, and an increased risk of overfitting, particularly when the dataset is small. Data for visual deception detection tasks are often characterized by short, low-resolution videos that are susceptible to noise. Hence, a lightweight model is more appropriate.

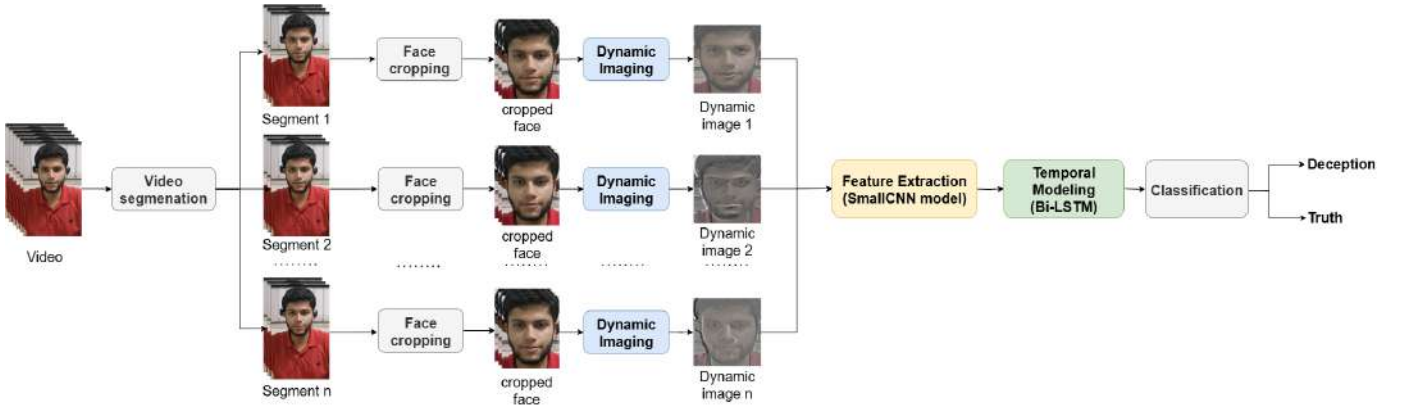


Fig. 1. Pipeline for deception detection using dynamic images and SmallCNN-BiLSTM

Third, SmallCNN is compact and easily adaptable. Its architecture can be adapted to suit the characteristics of visual data in deception detection tasks. Using SmallCNN helps reduce training time and computational resources, making it more suitable for practical deployment scenarios than heavier architectures, while still preserving the ability to extract discriminative features effectively for downstream classification tasks.

For these reasons, SmallCNN is a reasonable and efficient choice for feature extraction from DIs. The architecture of SmallCNN is illustrated in Fig 2.

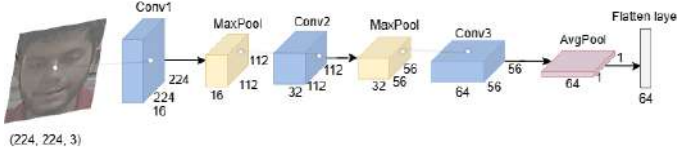


Fig. 2. The architecture of SmallCNN model

The proposed SmallCNN consists of three convolutional blocks. Each block includes a convolutional layer (with kernel size 3×3 and padding 1), followed by a ReLU activation and a downsampling layer. The first two blocks use max pooling layers, while the last block applies an adaptive average pooling to reduce the spatial dimensions to 1×1 . The final output is a 64-dimensional feature vector representing the input image. This architecture ensures simplicity, efficiency, and suitability for training on small datasets while still retaining important features from the DI.

D. Temporal modeling and Classification

In this step, the sequence of features extracted from the DI segments is fed into a Bi-LSTM network to capture temporal dependencies in both forward and backward directions. This is particularly important in short video sequences, as deceptive behavior may depend not only on previous movements but also on subsequent ones. For example, an unusual eye gaze or a hesitant facial expression may only be meaningful when considered in the context of the entire video sequence. The

final hidden states from both directions are concatenated to form a single context vector representing the observed motion segment.

This context vector is then passed to a classifier composed of four fully connected layers, with a decreasing number of neurons in each layer. The first two layers incorporate additional techniques such as batch normalization, ReLU activation, and dropout (with rates of 0.5 and 0.3, respectively) to enhance the model's generalization capability, reduce the risk of overfitting, and promote more stable learning. The final layers perform dimensionality reduction and output class probabilities.

While the Bi-LSTM is responsible for modeling the temporal sequence, the classifier transforms the learned representation into the final prediction. The combination of these two components enables the model to effectively capture temporal dynamics and achieve accurate classification in the deception detection task.

IV. EXPERIMENT AND RESULTS

A. Experiment design

We design two experimental scenarios to evaluate the effectiveness of the proposed DI-based method in the task of deception detection:

Experiment 1: The first experiment investigates how the temporal segment length used to generate each DI affects classification performance. It has been observed that when very short segments (e.g. 0.5 seconds) are used, insufficient motion is captured in the resulting DIs, which may limit the ability of the model to identify relevant behavioral cues. Conversely, when longer segments (e.g. 4–5 seconds) are applied, too much motion may be compressed into a single DI, leading to temporal ambiguity and the inclusion of irrelevant information. We evaluate six different durations: 0.5s, 1s, 2s, 3s, 4s, and 5s per DI. For each case, the sequence of DIs generated from the input video is passed through a two-stage system:

item *Feature extractor (CNN)*: including the proposed SmallCNN and three popular pre-trained models —

ResNet18, EfficientNet-B0, and MobileNetV2, all pre-trained on the ImageNet dataset. *Temporal modeling (RNN)*: including LSTM and Bi-LSTM, to analyze the effect of unidirectional vs bidirectional sequence learning.

Each CNN-RNN combination is trained independently and evaluated using 5-fold cross-validation ($k = 5$).

Experiment 2: In the second experiment, we compare the best-performing configuration from Experiment 1 with several recent baseline methods, including both visual-only and multimodal approaches. This comparison aims to assess whether our DI-based approach, while using video modality only, can compete with or outperform more complex multimodal systems.

B. Dataset

The experiments were performed on two datasets: Bag of Lies (BOL) and MU3D.

Bag of Lies [12]: consists of 325 videos from 35 participants (162 deceptive, 163 truthful), recorded using a phone camera. Participants were asked to describe 6–10 images either truthfully or deceptively. Although the dataset also includes audio, gaze, and EEG signals, this study only uses the video modality.

MU3D [19]: contains 320 videos from 80 speakers of different genders and races. Each participant recorded four videos under the following conditions: truthful–positive, truthful–negative, deceptive–positive, and deceptive–negative. In this paper, only truthful and deceptive labels are used, combining both positive and negative cases. The videos (45 seconds each) were recorded in a controlled environment to study low-stakes deception.

To increase data diversity and reduce overfitting, the dataset was augmented 10-fold using horizontal flipping and slight rotations ($\pm 5^\circ$, $\pm 10^\circ$). The images were then normalized, converted to 3-channel grayscale, resized to 224×224 , and transformed into tensors.

C. Results

TABLE I

CLASSIFICATION ACCURACY (%) ON MU3D DATASET UNDER DIFFERENT DI DURATIONS AND MODEL ARCHITECTURES

Model	Input duration per DI					
	1DI/0.5s	1DI/1s	1DI/2s	1DI/3s	1DI/4s	1DI/5s
<i>Bi-LSTM</i>						
Small-CNN	59.69	68.44	65.32	62.50	62.19	61.87
EfficientNet	50.63	60.63	52.19	51.25	51.25	51.25
MobileNet	51.87	63.75	55.94	55.69	53.75	52.81
ResNet	42.81	58.44	53.75	53.44	50.63	50.62
<i>LSTM</i>						
Small-CNN	59.37	65.62	61.88	61.25	60.94	60.00
EfficientNet	48.44	58.12	52.81	51.87	51.25	49.06
MobileNet	50.00	57.81	53.44	53.12	53.13	52.50
ResNet	49.06	57.19	54.37	53.75	51.87	51.56

1) *Experiment 1 result:* Tables I and II present the classification accuracy on the two datasets, MU3D and BOL, under different DI segment durations, combined with various CNN + RNN model architectures. From the results, it is evident that the 1s/1DI configuration consistently yields the highest

TABLE II

CLASSIFICATION ACCURACY (%) ON BOL DATASET UNDER DIFFERENT DI DURATIONS AND MODEL ARCHITECTURES

Model	Input duration per DI					
	1DI/0.5s	1DI/1s	1DI/2s	1DI/3s	1DI/4s	1DI/5s
<i>Bi-LSTM</i>						
Small-CNN	58.09	71.43	63.35	60.26	59.01	58.34
EfficientNet	59.63	64.92	60.86	59.00	58.39	59.63
MobileNet	56.83	66.17	60.57	59.32	58.70	57.76
ResNet	54.97	64.90	60.26	57.77	56.21	57.13
<i>LSTM</i>						
Small-CNN	58.70	64.60	62.72	60.58	57.44	59.01
EfficientNet	58.44	61.50	59.94	59.33	59.46	59.31
MobileNet	59.02	63.35	59.32	59.64	59.94	59.62
ResNet	57.48	62.12	59.97	59.02	58.71	58.38

performance across all model combinations, achieving up to 68.44% on MU3D and 71.43% on BOL with the SmallCNN + Bi-LSTM architecture.

In contrast, configurations such as 0.5s/1DI, 4s/1DI, and 5s/1DI generally result in lower accuracy across architectures. Specifically, with the 0.5s/1DI setting, the accuracy fluctuates only around 58%–59% for SmallCNN + LSTM/Bi-LSTM on both datasets, and drops as low as 49.06% for ResNet + LSTM on MU3D. These findings indicate that the temporal length of each DI segment directly impacts the quality of extracted features and model performance. Segments that are too short or too long tend to degrade representation learning and classification accuracy. The 1s/1DI setting offers an optimal balance long enough to capture motion variation, yet short enough to avoid redundancy or noise.

Furthermore, the proposed SmallCNN consistently outperforms pre-trained backbones such as EfficientNet, MobileNet, and ResNet. This suggests that deeper CNNs like ResNet are more prone to overfitting on short video-based DI representations with limited content. In contrast, SmallCNN, with its lightweight structure, is more suitable for compressed motion representations, allowing more effective feature learning. For example, on BOL with Bi-LSTM, SmallCNN achieves 71.43%, while ResNet only reaches 64.9%.

Moreover, Bi-LSTM consistently outperforms standard LSTM in most scenarios, especially when combined with SmallCNN: on MU3D, 68.44% (Bi-LSTM) vs 65.62% (LSTM); and on BOL, 71.43% (Bi-LSTM) vs 64.60% (LSTM). This highlights the benefit of bidirectional temporal modeling, which enables the model to capture contextual information from both past and future frames in the DI sequence, thus improving the ability to detect deceptive behaviors.

2) *Experiment 2:* Table III presents the classification accuracy results of different lie detection methods with various data types and evaluation settings (10-folds, 5-folds, and 3-folds) on BOL dataset. The proposed method, which uses Dynamic Images (DI) combined with the SmallCNN + Bi-LSTM architecture, consistently outperforms the previous best model, DDCapsNet (71.16%), which uses a combination of features such as facial expressions (AU), gaze, and head motion. This result shows that dynamic images are effective at encoding both motion and temporal information from videos,

TABLE III
COMPARISON OF MODALITIES AND METHODS ON DECEPTION DETECTION
ACCURACY (%) ON BOL DATASET

Modality & Inputs	Method	Accuracy
10-folds		
Visual (AU + Gaze + Head)	DDCapsNet [11]	71.16
Visual (AU)	LSTM [11]	67.11
Visual (DI)	SmallCNN + LSTM (ours)	72.95
5-folds		
Visual (Face frames) + Visual (AU + Gaze + Affect)	Atten-Mixer [9]	55.45
Visual (Face frames) + Visual (AU + Gaze + Affect) + Audio (Mel spectrogram)	Atten-Mixer [9]	59.41
Visual (LBP)	LieNet	46.43
Visual (DI)	SmallCNN + LSTM (ours)	71.43
3-folds		
Gaze + Video + Audio	Score-level fusion of best performing algorithms on various modalities [12]	60.09
Visual (DI)	SmallCNN + LSTM (ours)	65.84

while SmallCNN + Bi-LSTM still ensures strong feature learning capabilities.

In the 5-fold setting, our method continues to outperform with an accuracy of 71.43%, which is significantly higher than multi-modal methods such as Atten-Mixer using visual and audio features (59.41%) or visual features only (55.45%). This shows that adding more data types does not necessarily improve performance if the relationships between them are not optimally exploited.

Similarly, in the 3-fold setting, the proposed method achieves an accuracy of 65.84%, which is higher than the score-level fusion of the best multi-modal models (60.09%). This confirms that the DI-based approach not only simplifies processing but also provides effective features for the deception detection task.

Meanwhile, models using hand-crafted features such as LieNet (based on LBP) only achieved 46.43%, showing that modern deep learning methods combined with motion encoding significantly outperform traditional approaches.

V. CONCLUSION

This paper has proposed a lightweight and efficient visual deception detection method that combines DI with the SmallCNN-BiLSTM model. By compressing motion information into a compact representation and exploiting bidirectional temporal dependencies, our method achieves a balanced trade-off between accuracy and computational cost. A key advantage of the proposed SmallCNN-BiLSTM architecture is its lightweight design with fewer parameters, making it suitable for real-time deployment and resource-constrained environments. This not only ensures scalability but also opens up the potential for practical applications of vision-based deception detection beyond laboratory settings, such as on mobile or embedded devices. Experimental results on two datasets, MU3D and BOL, demonstrate that our method achieves competitive or superior performance compared to existing approaches while maintaining lower computational cost. These findings

highlight the potential impact of our research in advancing practical and accessible deception detection solutions.

REFERENCES

- [1] S. L. King and T. Neal, "Applications of ai-enabled deception detection using video, audio, and physiological data: A systematic review," *IEEE Access*, vol. 12, pp. 135207–135240, 2024.
- [2] P. Ekman, M. O'Sullivan, and M. G. Frank, "A few can catch a liar," *Psychological Science*, vol. 10, no. 3, pp. 263–266, 1999.
- [3] S. L. King, N. Johnson, K. Kosyluk, and T. Neal, "Therapist perceptions of automated deception detection in mental health applications," in *Artificial Intelligence in HCI* (H. Degen and S. Ntoa, eds.), (Cham), pp. 83–97, Springer Nature Switzerland, 2023.
- [4] E. Garrido, J. Masip, and C. Herrero, "Police officers' credibility judgments: Accuracy and estimated ability," *International Journal of Psychology*, vol. 39, pp. 254–275, 08 2004.
- [5] G. Joshi, V. Tasgaonkar, A. Deshpande, A. Desai, B. Shah, A. Kushawaha, A. Sukumar, K. Kotecha, S. Kunder, Y. Waykole, H. Maheshwari, A. Das, S. Gupta, A. Subudhi, P. Jain, N. K. Jain, R. Walambe, and K. Kotecha, "Multimodal machine learning for deception detection using behavioral and physiological data," *Scientific Reports*, vol. 15, no. 1, p. 8943, 2025.
- [6] S. A. Prome, N. A. Ragavan, M. R. Islam, D. Asirvatham, and A. J. Jegathesan, "Deception detection using machine learning (ml) and deep learning (dl) techniques: A systematic review," *Natural Language Processing Journal*, vol. 6, p. 100057, 2024.
- [7] S. Islam, P. Saha, T. Chowdhury, A. Sorowar, and R. Rab, "Non-invasive deception detection in videos using machine learning techniques," in *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1–6, 2021.
- [8] F. M. Talaat, "Explainable enhanced recurrent neural network for lie detection using voice stress analysis," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 32277–32299, 2024.
- [9] X. Guo, Z. Yu, N. M. Selvaraj, B. Shen, A. W.-K. Kong, and A. C. Kot, "Benchmarking cross-domain audio-visual deception detection," *arXiv preprint arXiv:2405.06995*, p. 15, 2024.
- [10] H. U. D. A. Khan, U. I. Bajwa, N. I. Ratyal, F. Zhang, and M. W. Anwar, "Deception detection in videos using the facial action coding system," *Multimedia Tools and Applications*, vol. 84, no. 9, pp. 6429–6443, 2025.
- [11] S. Gao, L. Chen, Y. Fang, S. Xiao, H. Li, X. Yang, and R. Song, "Video-based deception detection via capsule network with channel-wise attention and supervised contrastive learning," *IEEE Open Journal of the Computer Society*, vol. 5, pp. 660–670, 2024.
- [12] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-lies: A multimodal dataset for deception detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 83–90, 2019.
- [13] P. Li, M. Abouelenien, R. Mihalcea, Z. Ding, Q. Yang, and Y. Zhou, "Deception detection from linguistic and physiological data streams using bimodal convolutional neural networks," in *2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPPDS)*, p. 263–267, IEEE, May 2024.
- [14] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Detecting deceptive behavior via integration of discriminative features from multiple modalities," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1042–1055, 2017.
- [15] M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "Lienet: A deep convolution neural network framework for detecting deception," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 971–984, 2022.
- [16] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," 2017.
- [17] B. Fernando, E. Gavves, M. José Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5378–5387, 2015.
- [18] B. Fernando, E. Gavves, J. Oramas M., A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, p. 773–787, Apr. 2017.
- [19] E. P. Lloyd, J. C. Deska, K. Hugenberg, A. R. McConnell, B. T. Humphrey, and J. W. Kunstman, "Miami university deception detection database," *Behavior Research Methods*, vol. 51, no. 1, pp. 429–439, 2019.

Deep learning-based segmentation of Rooftop photovoltaic panels in Ho Chi Minh City, Vietnam

1st Hai N. Nguyen

*Faculty of Computer Science
University of Information Technology
Ho Chi Minh City, Viet Nam
23520418@gm.uit.edu.vn*

2nd Tran Thi Thanh Dung

*Faculty of Environmental Science
VNUHCM-University of Science
Ho Chi Minh City, Viet Nam
trttdung@hcmus.edu.vn*

3rd Truong Dinh Hung

*Faculty of Mathematics and Computer Science
VNUHCM-University of Science
Ho Chi Minh City, Viet Nam
24280068@student.hcmus.edu.vn*

4th Ha Minh Tuan

*Faculty of Mathematics and Computer Science
VNUHCM-University of Science
Ho Chi Minh City, Viet Nam
hmtuan@hcmus.edu.vn*

Abstract—As the global transition toward renewable energy accelerates, rooftop photovoltaic (PV) systems have emerged as a critical solution for sustainable electricity generation in urban environments. This study investigates the application of deep learning-based semantic segmentation to detect and delineate rooftop photovoltaic (PV) installations in Ho Chi Minh City, Vietnam. To support this investigation, we curated a dataset consisting of 804 annotated satellite images of rooftop PV installations using our proposed pipeline and used it to evaluate three segmentation models: U-Net, Feature Pyramid Network (FPN), and DeepLabv3+. The evaluation highlights domain-specific challenges such as rooftop heterogeneity, panel size variation, and the impact of low-resolution imagery. The results show that U-Net achieved the best performance, with an Intersection over Union (IoU) of 0.8532 and F1-score of 0.9201, outperforming the other models in preserving panel boundaries. Moreover, the dataset can be further used for training better deep learning models to extract PV area, enhancing future comparative studies and further developments.

Index Terms—Photovoltaic panels, Remote sensing, GIS systems, Semantic segmentation.

I. INTRODUCTION

The rapid global shift toward renewable energy has positioned solar photovoltaic (PV) technology as one of the most promising solutions for sustainable electricity generation. According to the International Renewable Energy Agency (IRENA), global installed PV capacity surpassed 1,177 GW in 2022, contributing significantly to the decarbonization of the energy sector [1]. In Vietnam, the combination of favorable location and many hours of sunlight makes solar power, especially rooftop systems, an inevitable part of urbanization. Monitoring and accurately mapping the location and scale of photovoltaic (PV) panels is very important. It helps policymakers assess the potential of renewable energy and supports power companies in managing and operating the grid effectively, especially in dynamic economic centers like Ho Chi Minh City.

To automate this monitoring process, deep learning methods, especially semantic segmentation, have shown strong effectiveness. Recent advances in computer vision, particularly deep learning-based semantic segmentation, have created new opportunities for large-scale PV detection using high-resolution satellite and aerial imagery. Unlike traditional object detection, semantic segmentation classifies each pixel within an image, enabling precise delineation of PV panel boundaries [2]. This approach has been successfully applied in various geographic contexts and has shown robustness in complex urban environments with heterogeneous roof materials and installation patterns [3]. Nevertheless, model performance is strongly influenced by the representativeness of the training datasets. The automation and high accuracy of deep learning have opened a new era in the management and exploitation of large-scale solar energy.

However, the performance of deep learning models largely depends on the availability of large, diverse, high-resolution training datasets. For example, a study in China [4] published a diverse dataset comparing images taken over fields, cropland, and water, with high spatial resolutions of 0.8, 0.3, and 0.1 m, respectively. Another study in Germany [5] used images with resolution of 0.15 and 0.31 m. Similarly, a US dataset had a spatial resolution from 0.3 to 0.8 m [6]. The common feature of these datasets is their high resolution. In those images, solar panels were often placed on rooftops in areas where the buildings were simple and looked similar. As a result, the solar panels are easily visible and distinguishable from the environment.

This contrasts with the data we collected in Ho Chi Minh City. Our data has a lower resolution, making it difficult at times to separate the solar panels from surrounding metal roofs. Most publicly available datasets are collected in North America, Europe, and China, which do not capture the specific challenges of urban areas in Vietnam. Identifying rooftop solar panels in Ho Chi Minh City is particularly challenging due to

several local factors. Roof structures are highly diverse, mixing materials such as corrugated metal, tiles, and concrete, which creates a visually complex background. Panel sizes also vary greatly, from a few small household units to large commercial systems. These difficulties are exacerbated by low-resolution imagery, where small panels are difficult to distinguish and can be mistaken for reflective roof surfaces. The biggest barrier is the lack of any publicly available labeled dataset for this area, making it difficult to develop and validate machine learning models. To address this research gap, our study introduces a new dataset specifically created and carefully annotated for rooftop solar panel segmentation in Ho Chi Minh City. Our main contributions are: (1) We explored a new path towards photovoltaic panels segmentation in urban area, beginning with Tan Phu and Tan Binh districts in Ho Chi Minh City; and (2) establishing baseline results by evaluating standard segmentation models on the dataset, enabling future comparative studies and further developments. To collect data from these districts, we conducted a manual visual survey of the entire area using Google Maps. We scanned each location and manually annotated all coordinates that contained a solar panel installation. This method is a form of exhaustive sampling for a targeted area. Therefore, our dataset is highly representative of solar panels in dense urban residential areas, which is the focus of our study.

II. RELATED WORKS

PV detection from remote sensing imagery has advanced rapidly, moving from simple classification models to more complex segmentation architectures. Early studies often used Convolutional Neural Networks (CNNs) to classify whether an image patch contained solar panels. These methods worked well on low-quality imagery but could not provide precise location information [7]. Today, semantic segmentation is considered the standard approach because it can pinpoint objects at the pixel level. Many modern network architectures have been applied. DeepLabv3+, for example, captures multi-scale context [8]. PVNet takes a more specialized route, using a contour refinement module to sharpen panel boundaries [9]. More recently, systems like SolarDetector combine Mask R-CNN with GAN-based data generation to segment and analyze each panel in detail [6]. Some studies go the opposite way, choosing simpler machine learning methods to cut computational cost, but this often comes with a drop in accuracy [10].

The performance of all these methods depends heavily on the quality and diversity of training data. This is also the main weakness of the field today. Many early studies built their own datasets for validation. For example, one dataset of high-resolution aerial images (0.3 m) was created in California [11]. While the results were strong, such datasets usually represent the uniform urban layouts of North America [11], Europe [12], [13], and China [9], [14], which have roof structures and planning patterns very different from many other regions.

Many studies use images from unmanned aerial vehicles (UAVs or drones) to detect solar panels [15]–[17]. Drone images often have higher resolution than satellite images. This

helps to see more details. It is possible to locate panels and also find small defects on their surface. Deep learning models are often used for this task. Convolutional Neural Networks (CNNs) are common. They can be trained on RGB or thermal images from drones. The models can detect dust, cracks, cell faults, and hotspots. Modern architectures such as YOLO or Mask R-CNN can detect and classify defects at the same time. They can work in real time. This saves time and reduces cost for maintaining large solar farms.

Some studies show that PV data has very unique characteristics. There is a strong class imbalance. The texture is uniform, but the color can vary a lot. Model performance drops sharply when image resolution is below 0.3 m [18]. This means a model trained in one region may not work well in another region. Extra tuning or new data is often needed.

These models have shown very good results in detecting solar panels in different datasets [4], [9], [19]. However, applying them to satellite imagery in Vietnam still faces several challenges. These include low-quality satellite images, a lack of labeled datasets, and the difficulty of training machine learning models.

III. DATA ACQUISITION AND SOURCES

The study area is located in Ho Chi Minh City, recognized as the principal economic hub of Vietnam, located in the southern region of the country (Figure 1). The city experiences a tropical monsoon climate characterized by elevated levels of solar radiation, which creates optimal conditions for the proliferation of rooftop photovoltaic (PV) systems. Due to the elevated population density coupled with restricted land availability, the majority of PV systems are mounted on building rooftops, as opposed to extensive ground-mounted solar farms. This investigation concentrates on two urban districts: Tan Phu and Tan Binh, which have been selected as representative zones owing to their varied rooftop typologies, disparate building densities, and distinct patterns of PV installations. The prevalent roofing materials within these districts include corrugated metal, clay tiles, and flat concrete, which contribute to substantial spectral and textural variability in satellite imagery. Such conditions present both challenges and opportunities for the development and assessment of PV segmentation models. High-resolution imagery was acquired via the Google Maps Static API, and all images underwent visual inspection to ensure minimal cloud coverage and satisfactory radiometric quality prior to further processing. Potential rooftop PV sites were manually identified, with their geographic coordinates recorded to facilitate annotation and dataset preparation.

IV. METHODOLOGY

The schematic pipeline to generate PV samples in Ho Chi Minh city employed by this study is illustrated in Figure 2. The procedure integrates manual inspection, automated image retrieval, polygon-based annotation, and rigorous quality control to ensure dataset reliability. This structured approach ensures the generation of a high-fidelity dataset tailored to

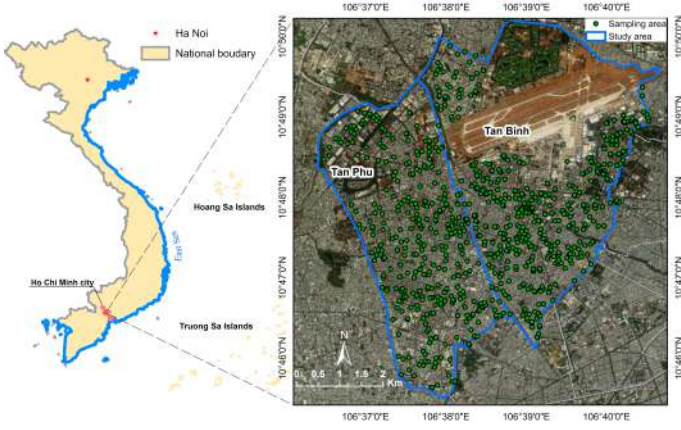


Fig. 1. Study area and rooftop PV sample locations in Tan Phu and Tan Binh districts, Ho Chi Minh City, Vietnam

the unique urban characteristics of Ho Chi Minh City, Vietnam. The methodology aligns with conventional protocols in remote sensing and computer vision for the management of geospatial data, encompassing multi-resolution image capture and polygon-based annotation. This approach underscores the importance of reproducibility, precision, and robustness in addressing domain-specific challenges such as diverse roofing materials and low-resolution images.

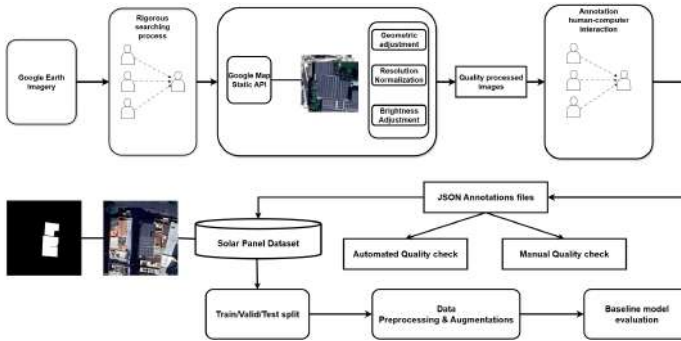


Fig. 2. Workflow for dataset creation and baseline evaluation.

A. Manual searching

Potential PV systems were identified based on distinct visual indicators observable in high-resolution Google Earth imagery. The identification process followed a standardized protocol to minimize subjectivity and enhance reproducibility. An object was flagged as a rooftop PV installation only if it satisfied at least three of the following criteria: (i) geometric regularity, such as rectangular modules arranged in consistent orientations; (ii) spectral contrast with adjacent surfaces, characterized by darker tones against lighter roof backgrounds and quantified by a noticeable difference in RGB channel values; (iii) contextual placement on elevated rooftop structures rather than ground-mounted configurations, verified through shadow pattern analysis; and (iv) absence of confounding features such as heavy shading, vegetation cover, or skylights.

Areas of interest are first identify through a rigorous searching process, wherein multiple people collaborate to manually screen through locations. For each confirmed location, precise geographic coordinates (latitude and longitude) were recorded. The decision on whether an object qualified as a PV installation was made individually by each annotator during the scanning process, thereby reducing oversight bias. All compiled coordinates were subsequently reviewed by a final supervisor, who verified detections, corrected misclassifications, and ensured that only valid PV instances were retained. This multi-stage verification process ensured the creation of a high-quality dataset with minimal false positives.

The completed coordinates file subsequently serves as the input for the Google Maps Static API systems. This methodology ensures reproducibility, as well as uniform and controlled image acquisition, based on the specified longitudes and latitudes. Image requests were automated through a script that utilized the API's query parameters, specifying the central viewpoint (latitude and longitude), zoom level (set to 21), and image dimensions (640×640 pixels). The images then undergo a comprehensive data cleaning process. Geometric corrections are implemented to rectify any spatial distortion present. Resolution normalization is conducted to standardize pixel density across all images, thereby mitigating the effects of inconsistent ground sample distances. Finally, the brightness distribution is balanced accordingly.

B. Annotation process

Images are then split and distributed between annotators for efficiency. In account for different shape of solar panels, we decided to use polygon segmentation. The annotation process is carried out between each individual annotator, before passing to another person, make sure each PV panel are checked twice. Finally, all images are collected and sent to a supervisor responsible for removing any mistakes while filling in 'ing spots. Figure 3. shows examples of panels alongside with their annotations.

C. Creating samples

All annotations are stored in JavaScript Object Notation (JSON) format, which then undergo two different type of quality checks: Automated quality check is done by utilizing a script that scans through the annotation file for empty masks and overlapping masks. Manual quality checking require the supervisor to double check each image-annotation pairs to identify and fix false positives caused by rooftop textures or shadows, potentially removing non solar panel objects. The combination of automated and manual checks substantially reduces noise in the dataset, thereby improving the reliability of subsequent model training. The final result is the solar panel dataset, which can be used for training deep learning models.

V. EXPERIMENTS

A. Dataset

In order to examine the potential of our dataset, a series of experiments of using deep learning to extract accurate PV area

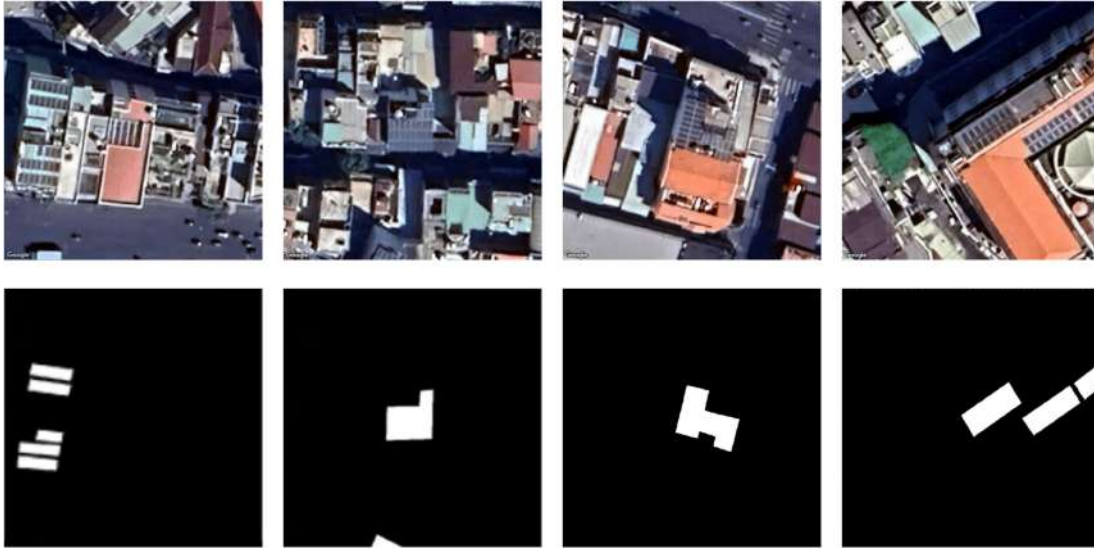


Fig. 3. Sample images with annotations from our dataset

is carried out. We compared the performance of three different deep learning models for this task: U-Net [20], FPN [21] and DeepLabv3+ [22]. U-Net employs an encoder-decoder pattern, going from a contracting path (encoder), allowing better context capturing, before going in a symmetric expanding path (decoder), enabling high localization capabilities. Feature maps from each encoder stage are concatenated with the corresponding decoder stage so the decoder gets direct access to fine-grained, high-resolution cues unaffected by pooling. Whereas FPN constructs a multi-scale feature pyramid on top of the backbone by adding a light top-down pathway and lateral connections. High-level, semantically rich feature maps are upsampled and merged with lower-level, high-resolution maps so that every pyramid level carries both strong semantics and precise spatial detail. DeepLabv3+ utilizes atrous spatial pyramid pooling to better capture multi-scale context, while also adapting the encoder-decoder architecture, enhancing the ability to capture object boundaries.

Our data set, consisting of 804 samples, is split into training set, validation set, and test set with a ratio of 7:2:1, respectively. The validation set is used to calculate metrics for comparison, while the test set is used to visualize the performance of models on various samples. All images go through a pre-processing step, resizing it to a uniform 512×512 , before applying data augmentations like horizontal flip, vertical flip and random rotations.

All models were implemented with stochastic gradient descent (SGD) optimization strategy in order to achieve a fast convergence speed, with an initial learning rate set to 0.001, momentum at 0.9 and decaying rate at 0.00001. All model checkpoints are saved at 50 epochs for further evaluation. In all training configurations, a common encoder backbone was used across models, with weights pretrained on the ImageNet-1k dataset, leveraging transfer learning for enhanced feature representation and improved performance. This study adopts

F1 score, Intersection over Union (IoU), precision, and recall as its main metrics for evaluating model performance.

VI. RESULTS AND DISCUSSION

U-Net demonstrated superior overall performance, characterized by an Intersection over Union (IoU) of 0.8532 and an F1-score of 0.9201, surpassing other models in both segmentation accuracy and the preservation of intricate object boundaries. Notably, FPN and DeepLabv3+ also yielded competitive outcomes, achieving IoU scores of 0.8406 and 0.8380, respectively, indicating that these models are effective at detecting rooftop photovoltaic systems within intricate urban environments. As illustrated in Table I, U-Net effectively maintains the most precise object boundaries, in contrast to other models, which tend to cause boundary bleeding. This phenomenon is clearly depicted in Figure 4, where row 2 and row 4 demonstrate how U-Net’s skip connections facilitate the retention of finer structural details. Specifically, the skip pathways in U-Net enhance detail conservation, whereas FPN’s multi-scale blending, while aiding in detection, may induce object shrinkage, and the atrous convolution in DeepLabv3+ might occasionally smooth out thin structures.

TABLE I
SEGMENTATION PERFORMANCE OF TESTED MODELS

Models	IoU	F1-Score	Precision	Recall
U-Net	0.8532	0.9201	0.9136	0.9278
FPN	0.8406	0.9130	0.9025	0.9243
DeepLabv3+	0.8380	0.9115	0.9061	0.9175

Furthermore, we prove the necessity of our dataset by training a DeepLabv3+ model on another dataset, namely the PVP dataset [9]. The PVP consists of high resolution satellite imagery in Yizhou District in Xinjiang, Dazi District in Tibet, Xintai City in Shandong province and Ruichang City Jiangxi province. Covering solar panels within desert, plateau,

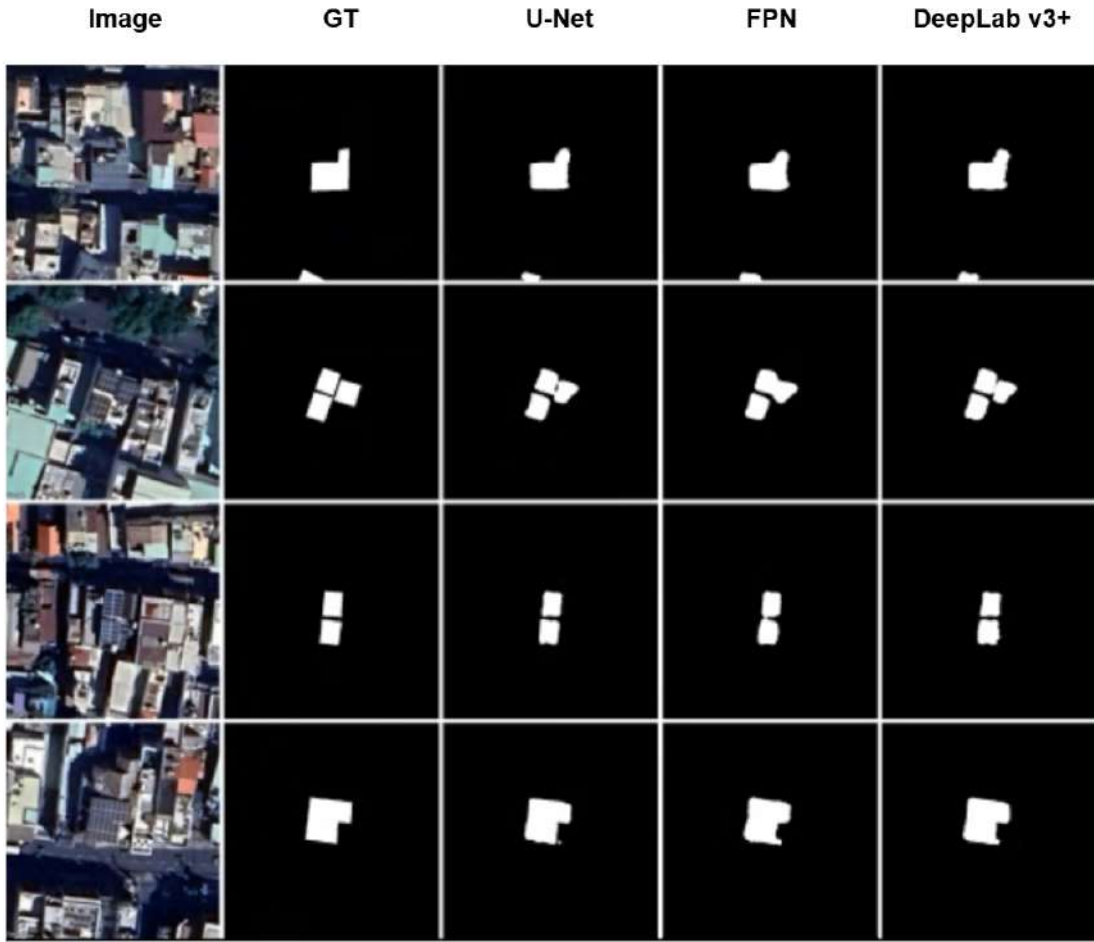


Fig. 4. Visualizing predictions of tested model

agricultural landscape, and water body environments. Making this a highly diverse dataset, pushing the model generalization capabilities higher. We proceed to test whether or not this dataset covers our cases.

We ran an experiment using the DeepLabv3+ model on the PVP dataset, with the exact same training configurations used above to ensure a fair comparison. The results have shown that while both are concerned about the segmentation of photovoltaic panel, there exists notable contrasts in panel structure, texture, and surrounding context. The differences in geographic location and environmental conditions heavily influence the model's ability to predict accurately. This is clearly indicated in the result, where the model after training only achieved a low overall metrics of 0.1318 IoU, 0.2298 F1-score, 0.8055 precision and only 0.1369 recall, demonstrating that each dataset exhibits vastly different attributes.

VII. LIMITATION

While overall performance was high, several limitations remain. All models exhibited difficulties in providing accurate boundaries of very small PV panels, and frequent misclassifications were observed for vents or rooftops with grid-like patterns. The challenges arise from the visual similarity

between certain roofing materials and PV panels, as well as limited resolution for smaller installations. The models also had problem with reflective rooftops as their strong glint reduces local contrast, leading to lower confidence near boundaries. These issues can be mitigated through a larger and more diverse dataset, along with targeted data augmentation techniques.

VIII. CONCLUSION

This research explores the application of deep learning-based semantic segmentation in the detection of rooftop photovoltaic (PV) systems within Ho Chi Minh City, Vietnam. A novel dataset comprising 804 high-resolution images was developed through a systematic workflow that encompasses manual inspection, image acquisition via the Google Maps Static API, polygon-based annotation, and multi-stage quality control processes. Three segmentation architectures were assessed: U-Net, Feature Pyramid Network (FPN), and DeepLabv3+. Among these, U-Net attained the highest accuracy, evidenced by an Intersection over Union (IoU) score of 0.8532 and an F1-score of 0.9201, surpassing the other models in terms of accurately delineating panel boundaries. Although FPN and DeepLabv3+ produced comparably competitive outcomes,

they occasionally exhibited under-segmentation in smaller arrays and misclassification of reflective roof surfaces. A cross-domain evaluation was conducted, wherein a DeepLabv3+ model, initially trained on an external dataset, was tested on the Ho Chi Minh City dataset. This resulted in a significant decline in performance (IoU 0.1318, F1-score 0.2298), emphasizing that model accuracy largely hinges on localized, high-quality datasets due to domain-specific variations in rooftop structures, materials, and urban configurations. Collectively, the findings underscore the necessity of integrating sophisticated deep learning models with representative local datasets to achieve reliable rooftop PV detection in rapidly urbanizing locales. The methodology and dataset introduced in this study offer a robust foundation for future advancements, which may include multi-sensor data integration, domain adaptation strategies, and operational-scale PV monitoring systems aimed at enhancing renewable energy planning and management. We are planning to expand the dataset to cover the entire urban area of Ho Chi Minh City in order to better capture the heterogeneity of rooftop structures. This will enable the evaluation of scalability at the city level and support the development of operational systems for large-scale PV detection.

ACKNOWLEDGMENT

This research is funded by University of Science, VNU-HCM under grant number T2024-93.

REFERENCES

- [1] International Renewable Energy Agency, "Renewable capacity statistics 2023," <https://www.irena.org/publications/2023/Mar/Renewable-Capacity-Statistics-2023>, 2023, accessed: 2025-08-15.
- [2] Q. Li, S. Schott, and D. Chen, "Solardetector: Automatic solar pv array identification using big satellite imagery data," in *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, 2023, pp. 117–129.
- [3] J. Wang, X. Chen, W. Jiang, L. Hua, J. Liu, and H. Sui, "Pvnet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 119, p. 103309, 2023.
- [4] H. Jiang, L. Yao, N. Lu, J. Qin, T. Liu, Y. Liu, and C. Zhou, "Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery," *Earth System Science Data Discussions*, vol. 2021, pp. 1–17, 2021.
- [5] C. N. Clark and F. Pacifici, "A solar panel dataset of very high resolution satellite imagery to support the sustainable development goals," *Scientific Data*, vol. 10, no. 1, p. 636, 2023.
- [6] Q. Li, S. Schott, and D. Chen, "Solardetector: automatic solar pv array identification using big satellite imagery data," in *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, 2023, pp. 117–129.
- [7] V. Golovko, S. Bezobrazov, A. Krushchanka, A. Sachenko, M. Komar, and A. Karachka, "Convolutional neural network based solar photovoltaic panel detection in satellite photos," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 1. IEEE, 2017, pp. 14–19.
- [8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
- [9] J. Wang, X. Chen, W. Jiang, L. Hua, J. Liu, and H. Sui, "Pvnet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 119, p. 103309, 2023.
- [10] F. Giussani, E. Wilczynski, C. Zandonella Callegher, G. Dalle Nogare, C. Pozza, A. Novelli, and S. Pezzutto, "Use of machine learning techniques on aerial imagery for the extraction of photovoltaic data within the urban morphology," *Sustainability*, vol. 16, no. 5, p. 2020, 2024.
- [11] J. Yuan, H.-H. L. Yang, O. A. Omitaomu, and B. L. Bhaduri, "Large-scale solar panel mapping from aerial images using deep convolutional networks," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 2703–2708.
- [12] C. Jörges, H. S. Vidal, T. Hank, and H. Bach, "Detection of solar photovoltaic power plants using satellite and airborne hyperspectral imaging," *Remote Sensing*, vol. 15, no. 13, p. 3403, 2023.
- [13] K. Bódis, I. Kougias, A. Jäger-Waldau, N. Taylor, and S. Szabó, "A high-resolution geospatial assessment of the rooftop solar photovoltaic potential in the european union," *Renewable and Sustainable Energy Reviews*, vol. 114, p. 109309, 2019.
- [14] H. Zhang, P. Tian, J. Zhong, Y. Liu, and J. Li, "Mapping photovoltaic panels in coastal china using sentinel-1 and sentinel-2 images and google earth engine," *Remote Sensing*, vol. 15, no. 15, p. 3712, 2023.
- [15] G. Francesco, L. Sonia, and N. Alessandro, "A semi-automated method for defect identification in large photovoltaic power plants using unmanned aerial vehicles," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2018, pp. 1–5.
- [16] U. Pruthviraj, Y. Kashyap, E. Baxevanaki, and P. Kosmopoulos, "Solar photovoltaic hotspot inspection using unmanned aerial vehicle thermal images at a solar field in south india," *Remote Sensing*, vol. 15, no. 7, p. 1914, 2023.
- [17] M. Sherman, M. Gammill, A. Raissi, and M. Hassanalian, "Solar uav for the inspection and monitoring of photovoltaic (pv) systems in solar power plants," in *AIAA Scitech 2021 Forum*, 2021, p. 1683.
- [18] P. Li, H. Zhang, Z. Guo, S. Lyu, J. Chen, W. Li, X. Song, R. Shibasaki, and J. Yan, "Understanding rooftop pv panel semantic segmentation of satellite and aerial images for better using machine learning," *Advances in applied energy*, vol. 4, p. 100057, 2021.
- [19] A. Greco, C. Pironti, A. Saggese, M. Vento, and V. Vigilante, "A deep learning based approach for detecting panels in photovoltaic plants," in *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, 2020, pp. 1–7.
- [20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [21] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016. [Online]. Available: <http://arxiv.org/abs/1612.03144>
- [22] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: <http://arxiv.org/abs/1802.02611>

HySUP-F: Face-Enhanced Full-Body Reconstruction from a Single Image

Le Cong Thuong*, Quang Nguyen Viet*, Hai-Chau Nguyen-Le*, Thanh Ha Le*§

*Faculty of Information Technology, University of Engineering and Technology, Hanoi, Vietnam

Abstract—Current 3D human reconstruction faces a persistent trade-off: holistic SMPLX-based methods (e.g., SMPLer-X, SMPLest-X) effectively capture full-body kinematics but often produce weak facial expressions, while specialized FLAME-based reconstructors (e.g., SMIRK, TEASER) deliver high-fidelity facial details yet remain parametrically incompatible with full-body models. To address this challenge, we introduce HySUP-F, a lightweight optimization framework that operates as a post-processing step, enabling the outputs of FLAME-based facial models to be integrated with SMPLX-based body meshes, without requiring retraining or architectural modifications. Our experiments show that this approach consistently reduces the Procrustes-Aligned Mean Per-Vertex Position Error (PA-MPVPE) by 2.6 % to 3.3 % across several state-of-the-art models. More importantly, this quantitative improvement leads to a noticeable qualitative enhancement in facial expressiveness, especially around the mouth region. Overall, our framework helps mitigate the limitations of holistic body models by enhancing their facial expressiveness, resulting in more compelling and coherent 3D digital humans. Code and more results can be found on our project page.

Index Terms—Expressive human pose estimation, SMPL-X, FLAME, 3D reconstruction

I. INTRODUCTION

The creation of photorealistic and expressive 3D digital humans is fundamental to modern computer graphics, supporting applications from immersive telepresence to next-generation virtual reality. As illustrated in Fig. 1, progress in this field has produced two powerful, complementary, yet incompatible research tracks. The first, holistic body reconstruction, uses parametric models such as SMPL-X [1] and SMPLest-X [2] to deliver state-of-the-art full-body kinematics, often at the expense of detailed facial expression. The second, specialized facial reconstruction, relies on dedicated head models like FLAME [3]; techniques such as SMIRK [4] and TEASER [5] capture high-fidelity facial detail but lack compatibility with full-body representations. Consequently, this divide enforces a persistent trade-off between whole-body coherence and facial realism, motivating the approach taken in this work. A solution is needed that can combine the expressive fidelity of specialized facial models with the comprehensive coverage of holistic body models, without retraining or architectural changes.

To address this gap, we introduce HySUP-F, a lightweight, model-agnostic optimization framework that integrates outputs from specialized facial models (e.g., FLAME) and holistic body models (e.g., SMPL-X) as a post-processing step.

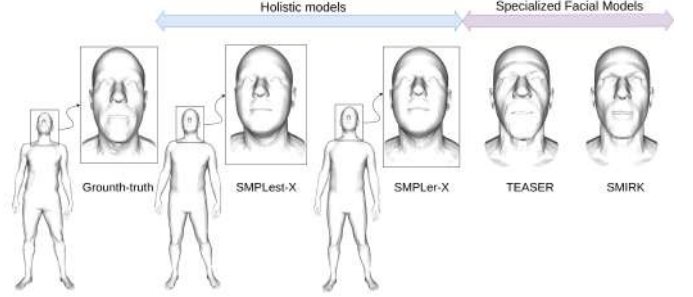


Fig. 1. Holistic models capture full-body pose but struggle with detailed facial and mouth expressions. Specialized models are generally more effective at facial detail, but cannot represent the whole body.

HySUP-F does not require retraining or architectural modification, enabling the detailed facial geometry of FLAME-based models to enhance any SMPL-X-based reconstruction. The framework uses two stages: first, it optimizes global shape to match facial identity; second, it refines facial expression and joint pose to better capture detailed facial features.

Experiments on the Expressive Hands and Faces (EHF) dataset [6], using multiple state-of-the-art model pairings, show that our method consistently reduces Procrustes-Aligned Mean Per-Vertex Position Error (PA-MPVPE) by 2.6 % to 3.3 %. It also yields noticeable improvements in facial expressiveness, especially around the mouth. These results demonstrate that HySUP-F effectively addresses the trade-off between full-body accuracy and facial detail.

II. RELATED WORK

a) Holistic 3D Human Reconstruction: The reconstruction of 3D human bodies from images was transformed by parametric models, beginning with the Skinned Multi-Person Linear model (SMPL) [7], which factorized shape and pose into a low-dimensional space. This foundation was extended by SMPL-X [1], which critically integrated fully articulated hands and expressive facial features. Early fitting methods like SMPLify [8] relied on iterative optimization against 2D evidence. However, the paradigm shifted to end-to-end regression, pioneered by HMR [9]. Subsequent work refined this approach with better supervision (SPIN [10]) and more powerful architectures using attention (PARE [11]). Recently, the field has scaled dramatically with Vision Transformer (ViT) models like SMPLer-X [12] and SMPLest-X [2] trained on massive datasets. Despite achieving impressive full-body

§ Corresponding author: ltha@vnu.edu.vn

kinematic accuracy, these state-of-the-art methods still produce muted or generic facial expressions.

b) *Expressive 3D Facial Reconstruction*: A parallel research track focuses on high-fidelity facial reconstruction, primarily using the FLAME head model [3]. Landmark methods like DECA [13] demonstrated robust regression of detailed FLAME parameters from a single image. Further research has targeted specific attributes, such as emotion-aware expression spaces (EMOCA [14]) and robust facial identity estimation (MICA [15]). While these techniques generate vivid and accurate faces, their fundamental limitation is their parametric isolation. The geometry and parameter spaces of FLAME-based models are incompatible with holistic SMPL-X models, preventing direct integration. Our work, HySUP-F, directly addresses this critical gap by proposing a universal bridge between these two powerful but complementary research domains.

III. HYSUP-F: FLAME-DRIVEN FACIAL EXPRESSION FUSION FOR SMPL-X BODIES

A. Overview

The core of our method is a two-stage optimization pipeline that robustly transfers facial details from the specialized FLAME head model to the holistic SMPL-X body model. To ensure robustness, the process is deliberately decoupled into two sequential steps: **Shape Optimization** followed by **Expression-Pose Optimization**. This separation isolates identity from expression, preventing the optimizer from erroneously altering a person’s underlying facial structure to explain a transient expression. The robustness gained from this design is validated in our ablation study (Table IV).

B. Formal Formulation

We denote the holistic SMPL-X model as M_S , defined by:

$$M_S(\beta, \theta, \psi) \rightarrow V_S, \quad (1)$$

where $\beta \in \mathbb{R}^{|\beta|}$ are global shape parameters, $\theta \in \mathbb{R}^{3J}$ represent joint rotations, and $\psi \in \mathbb{R}^{|\psi|}$ encode facial expressions. The global translation parameter, t , is omitted as our fitting process is invariant to global position due to the Procrustes alignment step described below. A specialized facial model, such as FLAME, is defined as:

$$M_F(\beta_F, \psi_F, \theta_F) \rightarrow V_F, \quad (2)$$

where $\beta_F \in \mathbb{R}^{|\beta_F|}$ represents facial shape, $\psi_F \in \mathbb{R}^{|\psi_F|}$ represents facial expressions, and θ_F represents local facial poses. We define a set of key facial regions, $\mathcal{R} = \{\text{face, eyes, lip}\}$, using vertex masks from SMIRK [4]. This selection allows our optimization to capture the overall facial structure (the “face” region) along with the most expressive features (the “eyes” and “mouth”). For each region $r \in \mathcal{R}$, let $\mathcal{I}_r \subseteq \{1, \dots, |V_F|\}$ be the *set of FLAME vertex indices* that belong to that region. Because FLAME and SMPL-X have different vertex orderings, a mapping function $f_{F \rightarrow S} : \mathcal{I}_r \mapsto \{1, \dots, |V_S|\}$ converts those indices to the holistic mesh. This mapping is

predefined using the official vertex correspondences provided by the SMPL-X and FLAME authors. Accordingly, we define

$$V_F^r = V_F[\mathcal{I}_r], V_S^r = V_S[f_{F \rightarrow S}(\mathcal{I}_r)], \quad (3)$$

which are the vertex subsets used in Eqs. (4)–(5).

C. Algorithm Details

Our optimization process minimizes a loss function that measures the vertex-level discrepancy between our deforming SMPL-X model and the target FLAME model. This loss calculation is consistent across both optimization stages: to ensure the error metric reflects only differences in local geometry—not global translation, rotation, or scale—we first perform a Procrustes alignment. Specifically, we find the optimal similarity transformation between the two vertex sets using the Umeyama algorithm [16] before calculating the final mean squared error. Regularization is also applied in both stages to ensure plausible results. The complete pipeline is summarized in Algorithm 1.

The two stages apply this fitting process to solve different sub-problems:

- 1) **The Shape Stage** isolates identity. It canonicalizes both models to a neutral T-pose to compute the loss between the two pose-invariant shapes, optimizing only the shape parameters (β).
- 2) **The Expression-Pose Stage** transfers the expression. Using the optimized shape as a fixed base, it computes the loss against the full, expressive FLAME target to fit the facial articulation parameters (ψ, θ_E).

The shape optimization loss is:

$$\mathcal{L}_{\text{shape}} = \kappa \sum_{r \in \mathcal{R}} \lambda_r \mathcal{E}(V_S^r, V_F^r) + \gamma_{\text{shape}} \|\beta\|_2^2, \quad (4)$$

where $\mathcal{E}(V_S^r, V_F^r)$ is the mean squared error after Umeyama alignment, λ_r weighs each region $r \in \mathcal{R}$, and κ is a loss-scaling factor. The term γ_{shape} enforces regularization on the shape parameters.

The expression-pose optimization loss is:

$$\mathcal{L}_{\text{expr, pose}} = \kappa \sum_{r \in \mathcal{R}} \lambda_r \mathcal{E}(V_S^r, V_F^r) + \gamma_{\text{expr}} \|\psi\|_2^2 + \gamma_{\text{pose}} \|\theta_E\|_2^2, \quad (5)$$

where γ_{expr} and γ_{pose} enforce regularization. We define the set of optimizable facial pose parameters as $\theta_E \subset \theta$, which includes the jaw and eye joints, allowing for fine-grained control over all facial articulations.

IV. EXPERIMENTS AND RESULTS

A. Setup

a) *Dataset*: We evaluate our method on a subset of 88 samples from the Expressive Hands and Faces (EHF) dataset [6]. EHF provides high-resolution RGB frames of a single subject performing diverse body poses, hand gestures, and facial expressions. The ground-truth SMPL-X meshes are generated by fitting the SMPL-X model to multi-view scans,

Algorithm 1 Two-stage Facial Parameter Transfer Optimization

Require: Paired holistic params $(\theta_{base}, \psi_{base})$ and specialized facial params $(\beta_F, \psi_F, \theta_F)$.

- 1: Initialize global shape $\beta_{opt} \leftarrow \mathbf{0}$ and expressive params $\psi_{opt}, \theta_{opt} \leftarrow \mathbf{0}$.
- 2: Load vertex masks and mapping $f_{F \rightarrow S}$.
- 3: **procedure** SHAPE OPTIMIZATION (IN T-POSE)
- 4: **for** $i = 1, \dots, N_{shape}$ iterations **do**
- 5: Compute $V_S = M_S(\beta_{opt}, \theta_{T-pose}, \mathbf{0})$
- 6: Compute $V_F = M_F(\beta_F, \theta_{neutral}, \mathbf{0})$
- 7: Compute shape loss $\mathcal{L}_{shape}$ using Eq. (3).
- 8: Update β_{opt} using an optimization step.
- 9: **end for**
- 10: **end procedure**
- 11: **procedure** EXPRESSION-POSE OPTIMIZATION
- 12: **for** $i = 1, \dots, N_{expr}$ iterations **do**
- 13: Compute $V_S = M_S(\beta_{opt}, \theta_{base} \cup \theta_{opt}, \psi_{opt})$
- 14: Compute $V_F = M_F(\beta_F, \psi_F, \theta_F)$
- 15: Compute expr, pose loss $\mathcal{L}_{expr, pose}$ using Eq. (4).
- 16: Jointly update ψ_{opt}, θ_{opt} using an optimization step.
- 17: **end for**
- 18: **end procedure**
- 19: **return** Optimized holistic parameters $(\psi_{opt}, \theta_{opt})$.

resulting in millimetre-level vertex accuracy for the body, hands, and face.

b) Metric: We report the Procrustes-aligned Mean Per-Vertex Position Error (PA-MPVPE). This metric measures the Euclidean distance (in mm) between predicted and ground-truth vertices after a similarity alignment that removes differences in global translation, rotation, and scale. A lower PA-MPVPE value indicates higher accuracy.

c) Hardware: All experiments are conducted on a workstation with an Intel Core i9-10900 CPU (2.80 GHz) and a single NVIDIA GeForce RTX 3090 GPU.

d) Implementation details: Table I lists all hyperparameters used in our experiments. These values were chosen based on preliminary experiments for stable convergence.

B. Results and analysis

Table II and Fig. 2 present a comprehensive evaluation on the EHF dataset, demonstrating the clear advantage of augmenting holistic body models with specialized facial components.

Quantitatively, integrating a dedicated head model (SMIRK or TEASER) with the SMPLest-X body achieves the lowest overall error, reducing the PA-MPVPE from 4.78 mm to 4.62 mm—a 3.3% improvement. For SMPLer-X, fusing a specialized head lowers the error from 5.07 mm to 4.94 mm (2.6% gain). These results confirm that specialized heads consistently recover fine-grained facial details lost by global alignment alone.

TABLE I
HYPERPARAMETER SETTINGS USED IN ALL EXPERIMENTS.

Category	Parameter	Value
<i>LBFGS Solver^a</i>	Iterations per stage ($N_{shape}, N_{expr, pose}$)	200, 400
	Learning rate (lr)	1.0
	Line search method	Strong Wolfe
	Max. iterations per step	10
	History size (m)	10
	Change tolerance (ε)	1×10^{-9}
<i>Loss scaling</i>	Global factor (κ)	1000
	Region weights (λ_r)	
	Overall face region	1.0
	Eyes regions	1.0
	Lips region	1.0
<i>L2 regularisation</i>	Shape (γ_{shape})	1×10^{-4}
	Expression (γ_{expr})	1×10^{-4}
	Jaw pose (γ_{jaw})	1×10^{-3}

^a Our implementation uses the LBFGS optimizer from the PyTorch framework.

TABLE II
PA-MPVPE↓ ON THE EHF DATASET. “BASELINE” IS THE PLAIN BODY MODEL; THE THREE ROWS BELOW ADD A SPECIALISED HEAD. LOWER ERROR IS BETTER; BEST RESULT PER BODY MODEL IS **BOLD**.

Body model	Configuration	PA-MPVPE (mm)	Improv. (%)
SMPLest-X	Baseline	4.78	—
	+ EMICA	4.68	2.1
	+ SMIRK	4.62	3.3
	+ TEASER	4.62	3.3
SMPLer-X	Baseline	5.07	—
	+ EMICA	4.98	1.8
	+ SMIRK	4.94	2.6
	+ TEASER	4.94	2.6

Qualitative results (Fig. 2) make this impact tangible: Compared to baseline body models, the fused meshes exhibit noticeably richer facial expressions and more accurate mouth shapes. In particular, our method restores subtle features and realistic deformations in the mouth region, which remains the most challenging area for holistic models.

In summary, our method mitigates the limitations of holistic body models by enhancing facial expressiveness and mouth realism, thereby improving the overall quality of digital humans both quantitatively and visually.

C. Ablation Study

We conduct ablation studies to validate two key aspects of our method: the inclusion of fine-grained facial constraints and the critical role of our two-stage optimization pipeline.

a) Impact of Lip and Eye Losses: First, we analyze the effect of adding specific loss terms for the lip and eye regions. Table III shows this provides a modest improvement for EMICA and TEASER (0.01 mm to 0.02 mm PA-MPVPE reduction), while having a negligible effect on the already

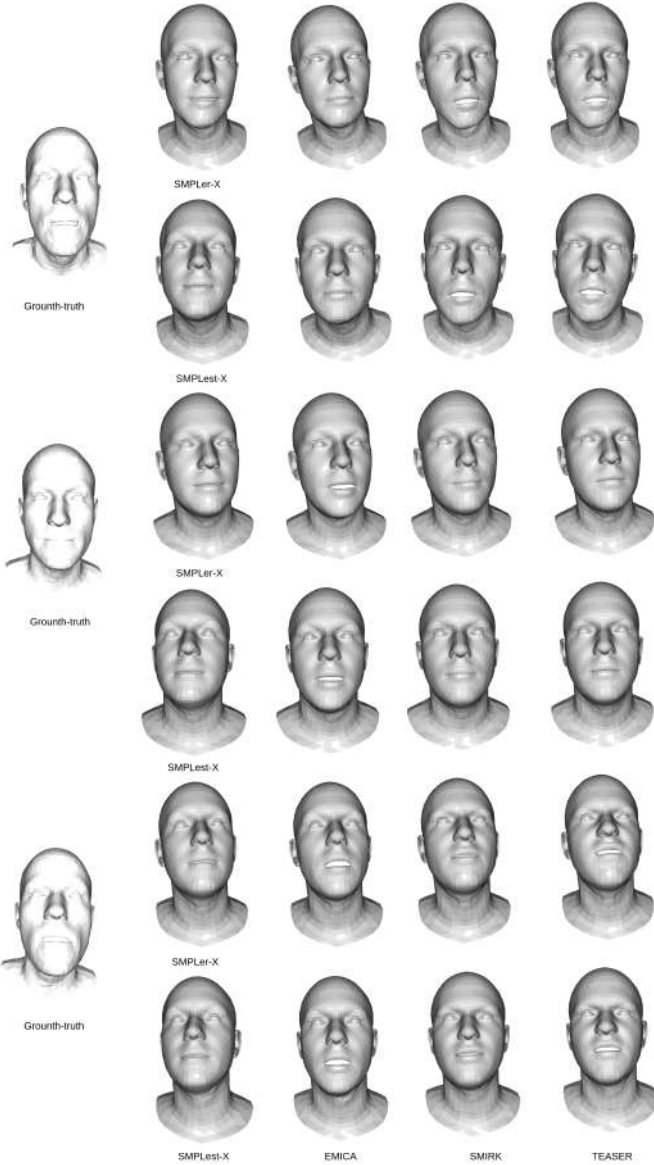


Fig. 2. For each identity, our method yields noticeably improved facial expressions and mouth shapes compared to baseline body models, highlighting the effectiveness of integrating a dedicated facial component.

TABLE III
EFFECT OF ADDING LIP AND EYE REGIONS TO THE LOSS. NUMBERS ARE PA-MPVPE ↓ (MM).

Loss setting	EMICA	SMIRK	TEASER
Face + lips + eyes	4.68	4.62	4.62
Face only	4.69	4.61	4.64

strong baseline of SMIRK. Although the absolute gains are small, these targeted constraints help ensure that highly deformable expressive regions are accurately modeled without harming overall performance. This justifies their inclusion in our final loss configuration to maximize facial detail.

b) Validating the Two-Stage Pipeline: Next, we validate our central hypothesis: that decoupling shape from expression is essential for robust performance. We tested a variant of our method that removes the initial shape-only stage and performs a single optimization for all parameters. The results in Table IV are unambiguous: removing the first stage causes a dramatic performance drop, increasing error by 0.26 mm to 0.31 mm (a 5 % to 6 % relative increase).

This powerfully confirms that first resolving the global facial identity (β) provides an essential geometric foundation. Without it, the optimizer erroneously distorts the facial structure to compensate for identity mismatch, leading to significantly less accurate results. This validates that our two-stage approach is critical for successfully fusing identity and expression.

TABLE IV
IMPACT OF THE SHAPE-ONLY STAGE (β -OPTIMIZATION). METRIC: PA-MPVPE ↓ (MM).

Setting	EMICA	SMIRK	TEASER
Stage-2 only (expr/pose)	4.99	4.88	4.89
Both stages (shape + expr/pose)	4.68	4.62	4.62

V. CONCLUSION

In this work, we presented HySUP-F, a lightweight optimization framework that successfully resolves the long-standing trade-off between holistic body coherence and specialized facial realism. Our two-stage pipeline robustly fuses high-fidelity, FLAME-based facial geometry onto any SMPL-X body model, acting as a universal post-processing step that requires no retraining. Experiments show this approach consistently reduces PA-MPVPE by 2.6 % to 3.3 %, which translates into significant qualitative improvements in facial expressiveness, particularly around the mouth.

While effective, our framework’s primary limitation is its dependence on the quality of upstream models, and its performance is inherently capped by the accuracy of the initial reconstructions. We acknowledge that our validation was conducted on a limited dataset; therefore, future work must extend experiments to larger, multi-subject datasets and rigorously investigate failure cases, such as scenarios with mismatched or low-quality facial inputs, to establish robust generalizability. A key avenue for future technical research is the direct integration of our optimization strategy into the training loop of foundational models. Crucially, as we advance this technology, we must also address ethical considerations by discussing potential risks and developing safeguards to ensure responsible use. Beyond these immediate steps, the enhanced articulation provided by our method opens promising research directions. Extending this fusion principle to incorporate specialized hand models, for example, could create a new standard for fully expressive digital humans, unlocking new potential in applications requiring nuance, from generating accurate visemes for speech-to-animation systems to automatic sign language translation.

Ultimately, HySUP-F provides a practical and effective bridge between disparate domains, paving the way for more compelling, coherent, and human-like digital characters.

ACKNOWLEDGMENT

Our sincere gratitude goes to the entire Vietnamese Sign Language group at the University of Engineering and Technology, Vietnam National University, Hanoi, for their generous contributions of time and effort. We would particularly like to thank Dr. Ngo Thi Duyen, B.Sc. Minh Hang Dang, B.Sc. Son Tung Ha, B.Sc. Thu Phuong Le, and B.Sc. Hai Long Luong for their valuable suggestions.

REFERENCES

- [1] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.
- [2] W. Yin, Z. Cai, R. Wang, A. Zeng, C. Wei, Q. Sun, H. Mei, Y. Wang, H. E. Pang, M. Zhang *et al.*, "Smplest-x: Ultimate scaling for expressive human pose and shape estimation," *arXiv preprint arXiv:2501.09782*, 2025.
- [3] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.
- [4] P. Filntisis, G. Retsinas, R. Danecek, V. F. Abrevaya, A. Roussos, T. Bolkart, and P. Maragos, "Smirk: 3d facial expressions through analysis-by-neural-synthesis," in *Greeks in AI Symposium 2025*, 2024.
- [5] Y. Liu, L. Zhu, L. Lin, Y. Zhu, A. Zhang, and Y. Li, "Teaser: Token enhanced spatial modeling for expressions reconstruction," *arXiv preprint arXiv:2502.10982*, 2025.
- [6] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.
- [7] —, "Smpl: a skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, Oct. 2015. [Online]. Available: <https://doi.org/10.1145/2816795.2818013>
- [8] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European conference on computer vision*. Springer, 2016, pp. 561–578.
- [9] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.
- [10] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2252–2261.
- [11] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3d human body estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 127–11 137.
- [12] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang *et al.*, "Smpler-x: Scaling up expressive human pose and shape estimation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 11 454–11 468, 2023.
- [13] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.
- [14] R. Daněček, M. J. Black, and T. Bolkart, "Emoca: Emotion driven monocular face capture and animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 311–20 322.
- [15] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," in *European conference on computer vision*. Springer, 2022, pp. 250–269.
- [16] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 4, pp. 376–380, 2002.

ZenID: FaceFusionNet-Guided Diffusion for Parent-to-Child Facial Prediction

Nguyen Hoang Lam[†], Vuong Dinh Minh[†], Nguyen Trong Tuan[†],
Nguyen Quang Tri, Nguyen Vo Ngoc Khue, Dinh Viet Sang^{*}
School of Information and Communication Technology
Hanoi University of Science and Technology, Hanoi, Vietnam

Abstract—Predicting the facial features of a child from images of their biological parents is a complex and underexplored challenge that intersects deep learning, generative modeling, and facial genetics. While traditional methods depend on handcrafted morphing or heuristic rules, recent advances in neural embeddings and diffusion-based synthesis have enabled more expressive modeling of familial traits. In this work, we introduce *ZenID*, a framework for synthesizing plausible child faces given parent images. *ZenID* encodes each parent’s features using a face recognition backbone, then fuses them via a learnable network to produce a child-specific embedding. This embedding conditions a Stable Diffusion model enhanced by a fine-tuned IP-Adapter, enabling high-resolution, identity-consistent generation. Our key contributions include: (i) a proposed fusion network that captures biologically plausible facial trait inheritance; (ii) an efficient adapter-tuning mechanism for identity-aware diffusion synthesis. Empirical evaluations demonstrate that *ZenID* produces high-quality, realistic, and genetically coherent child faces that inherit visual traits from both parents in a convincing manner. Experiments demonstrate that our *ZenID* outperforms prior methods in both qualitative and quantitative evaluations, while uniquely enabling controllable generation across gender, age, resolution, and background. This work suggests a potential direction for controllable facial prediction, with applications spanning digital art, family visualization, and genetic education.

Index Terms—Face synthesis, Identity fusion, Conditional diffusion, Kinship modeling, Generative AI

I. INTRODUCTION

Facial technologies are central to identity recognition, biometric authentication, and digital character synthesis. Advances in generative models, especially GANs [1] and Diffusion Models [2], have enabled photorealistic face generation. Within this progress, predicting a child’s face from parental portraits presents a novel challenge that merges deep learning with genetic inheritance. Despite its creative potential, the task is difficult due to the complexity of trait transmission, environmental influences, and the lack of large-scale family datasets.

Previous works such as ChildPredictor [3] and BaByGAN [4] have attempted to address this task using GAN-based architectures. While promising, these approaches are constrained by limited generalization

capacity due to narrow training data, lack of conditional control (i.e. age, gender), and insufficient biological plausibility in generated faces. Moreover, these methods often operate as black-box systems with little support for user interaction or customization.

To address these existing limitations, we introduce ZenID, a framework for predicting child faces from parental images. At its core is a novel neural network that fuses parent embeddings into a genetically coherent child representation, guiding a diffusion model for high-quality image synthesis. We also construct a diverse parent-child dataset to facilitate robust training and evaluation. In summary, our contributions to the task of child face prediction from parental portraits are:

- 1) **Latent-Space Mapping Approach**, which predicts child features by mapping parent embeddings to a child embedding, modeling genetic inheritance in latent space instead of direct image generation.
- 2) **Enhanced Conditioning via Fine-Tuned IP-Adapter**, improving image quality and controllability by fine-tuning IP-Adapter [5] in the diffusion model, enabling better generation with controllable attributes (e.g., age, gender, expression).
- 3) **Curated Parent-Child Dataset**, a diverse and balanced dataset of parent-child facial images from public sources, supporting robust training and generalization.

The remainder of this article is organised as follows: Section II reviews the relevant literature; Section III describes model architecture, and training methodology; Section IV details the dataset curation process and presents comprehensive experimental results and Section V offers concluding remarks and potential directions for future work.

II. RELATED WORK

This section highlights key advancements in kinship face synthesis, from early embedding-based GANs to StyleGAN-driven latent fusion, recent diffusion models with adapter tuning, and a motivation for the development of our approach.

[†] These authors contributed equally to this work.

^{*} Corresponding author.

A. Face Embedding and Kinship GANs

Early works in child face synthesis focused on modeling kinship using discriminative face embeddings and conditional generative networks. ArcFace [6] and InsightFace [7] introduced strong identity-preserving embeddings by applying angular margin constraints, forming the basis for parent-child representation. KinshipGAN [8] used a pre-trained embedding model within an adversarial setup to generate child faces from parent images, but its outputs were low-resolution and lacked detail. DNA-Net [9] improved this by using a conditional autoencoder that included biological cues like age and gender, resulting in more realistic outputs. However, these early methods were limited by poor image quality, weak generalization due to small datasets, and simple blending strategies that failed to capture complex inheritance patterns.

B. StyleGAN-Based and Hybrid Synthesis Approaches

The introduction of StyleGAN [10] enabled fine-grained control in face synthesis through latent space manipulation. Methods like StyleDNA [11] and StyleGene [12] use the disentangled space to interpolate specific facial regions between parent images, allowing partial inheritance of traits such as jawline or eyes. StyleGene adds mutation-inspired operations to improve diversity. While effective, these models involve complex encoding and often require additional tuning, limiting scalability. Hybrid models like StyleDiT [13] combine diffusion transformers with StyleGAN decoders, offering high-quality outputs and sample diversity. However, these approaches mainly target general face generation and lack explicit modeling of hereditary identity blending needed for bi-parental synthesis.

C. Diffusion-based Conditioning and Adapter Tuning

Recent progress in diffusion models has led to significant advances in image generation quality and flexibility. Methods such as DreamBooth [14] and LoRA [15] enable fine-tuning diffusion pipelines on subject-specific instances with minimal training steps, often utilizing cross-modal embeddings from CLIP [16] or semantic representations derived from Vision Transformers [17]. While effective for personalizing generation, these approaches are resource-intensive and prone to overfitting, especially when used on small or imbalanced datasets. IP-Adapter [5] introduces a cross-attention mechanism that conditions frozen diffusion models on external image features without retraining the backbone. This approach provides efficient multimodal synthesis with strong identity preservation, making it particularly suited for face-aware applications. However, IP-Adapter alone does not model structured relationships such as kinship, nor does it provide a mechanism for controlled identity fusion across multiple sources. ZenID builds on these foundations by modeling how facial traits are inherited from both parents, enabling more consistent and realistic child face generation than previous approaches.

III. METHODOLOGY

This section presents the architecture and methodology of ZenID. The system comprises main modules: (1) *FaceFusionNet* for embedding fusion, (2) a fine-tuned *IP-Adapter* for diffusion guidance.

A. System Overview

The system of the proposed ZenID is composed of tightly integrated modules, designed to simulate genetic inheritance while allowing context-aware customization. The overall architecture is illustrated in Fig 1. First, the portraits of the parents are processed by the FaceEncoder, which extracts two latent vectors: *Dad Embedding* and *Mom Embedding*. This module uses state-of-the-art pretrained models such as ArcFace [6] or InsightFace [7] to ensure stable and discriminative representations. The embeddings are passed to the FaceFusionNet, a trainable network that learns to synthesize a *Child Embedding* capturing plausible genetic traits. This *Child Embedding* serves as a conditioning signal for a diffusion-based image generation pipeline. The backbone UNet [18] integrates the embedding via *Cross-Attention* layers, ensuring the generated image aligns with inherited facial features. To preserve facial geometry, the IdentityNet module [19] processes input *facial landmarks* (e.g., eyes, nose, mouth positions), which guide structural consistency in the output image. Additionally, ZenID supports user-driven customization through natural language prompts. A *text prompt* describing attributes such as age, expression, hairstyle, or image style is encoded into a *Text Embedding*, injected into a separate CrossAttention stream, allowing semantic-level control over the generated output.

B. FaceFusionNet: Genetic embedding aggregation network

FaceFusionNet serves as the core module in ZenID, responsible for synthesizing a child face embedding from two input embeddings representing the parents. It learns a nonlinear mapping $f : \mathbb{R}^{512} \times \mathbb{R}^{512} \rightarrow \mathbb{R}^{512}$ such that $f(x_{\text{father}}, x_{\text{mother}}) \approx x_{\text{child}}$ in the identity embedding space. These two input embeddings x_{father} and x_{mother} from parent images are 512-dimensional vectors extracted using a pretrained face recognition model (e.g., InsightFace AntelopeV2 [7]).

1) *Architecture*: The architecture of FaceFusionNet consists of three components: two symmetric EncoderNet branches, a fusion Fully Connected layer, and a sequence of ResidualBlocks. Each parent embedding is independently processed through the same EncoderNet architecture to normalize and reproject into a shared latent space. In details, each EncoderNet comprises M stacked blocks, each with a Linear layer followed by BatchNorm, maintaining the output dimensionality at 512. The outputs from both EncoderNets are concatenated to form a 1024-dimensional vector, then passed through an FC layer to fuse information and

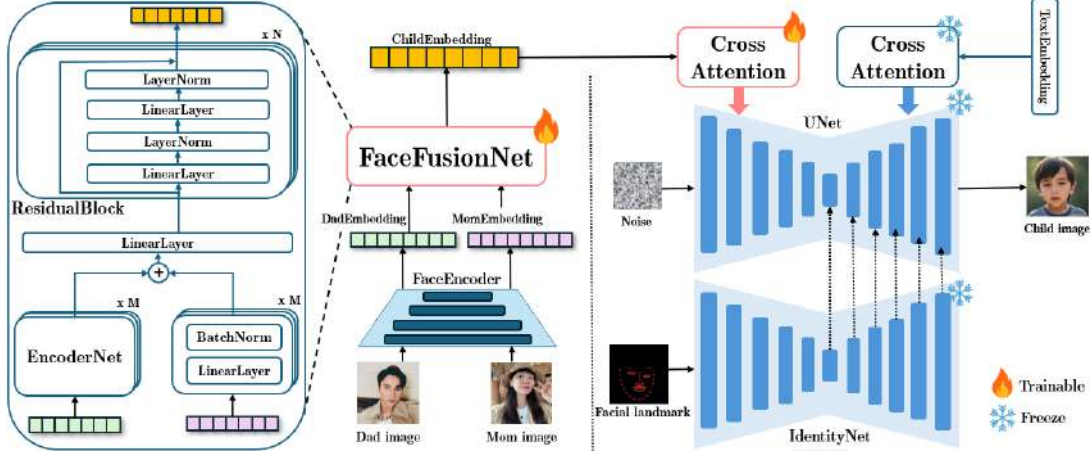


Figure 1: Overall architecture of ZenID. Parent images are encoded into identity embeddings by a pre-trained FaceEncoder, then fused via FaceFusionNet to produce a child embedding. This embedding guides a diffusion model through CrossAttention. Fine-tuned IP-Adapter layers inject facial landmark cues for identity preservation, while text prompts remain frozen.

project it into an intermediate 512-dim representation. This intermediate vector is then refined via N ResidualBlocks, which capture higher-order interactions between parental traits while preserving identity-specific signals. Each ResidualBlock consists of two Linear layers interleaved with LayerNorm, followed by a skip connection to stabilize training and retain input features.

2) *Training strategy*: FaceFusionNet is trained on a dataset of parent-child triplets. For each triplet, parent embeddings are passed into the network to predict the child’s embedding $\hat{y}$. The model is optimized to minimize a joint loss combining Mean Squared Error (MSE) and Cosine Similarity Loss:

$$\mathcal{L} = \lambda_1 \cdot \underbrace{\|\hat{y} - y\|^2}_{\mathcal{L}_{\text{MSE}}} + \lambda_2 \cdot \underbrace{\left(1 - \frac{\hat{y} \cdot y}{\|\hat{y}\| \|y\|}\right)}_{\mathcal{L}_{\text{cosine}}} \quad (1)$$

Here, λ_1 and λ_2 control the relative weights of the two loss terms. The MSE ensures magnitude alignment, while the cosine loss emphasizes angular similarity, which is crucial for face embedding consistency. This dual-loss strategy encourages the model to generate child embeddings that are both numerically and directionally aligned with the true child identity in the embedding space, thereby enhancing the biological plausibility and fidelity of the synthesized face.

C. Fine-tuning IP-Adapter: Bridging genetic embeddings with diffusion synthesis

In addition to FaceFusionNet, ZenID utilizes a fine-tuned IP-Adapter [5] to enhance the quality and identity alignment of images generated by the diffusion model. Acting as a bridge, the IP-Adapter transfers the child embedding produced by FaceFusionNet into the diffusion model’s conditioning space. Inspired by

InstantID [19], ZenID incorporates a similar conditioning pipeline, where child embeddings are injected into a frozen Stable Diffusion model via the IP-Adapter. To enrich semantic supervision, ZenID further utilizes BLIP [20] to generate textual captions from child images, forming multimodal input pairs (embedding and caption). During training, the Stable Diffusion weights are kept frozen. Only the IP-Adapter parameters are updated. For each training sample, a triplet image set (father, mother, child) is processed: FaceFusionNet generates the child embedding c_i , and BLIP provides a caption c_t from the child image. The target image x_0 is noised with Gaussian noise ϵ to obtain x_t at timestep t , and the diffusion model learns to denoise it under both conditions:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{x_0, \epsilon, c_t, c_i, t} \|\epsilon - \epsilon_\theta(x_t, c_t, c_i, t)\|^2 \quad (2)$$

Here, ϵ_θ is the predicted noise conditioned on the embedding c_i and caption c_t . To improve generalization, we adopt random condition dropout [5], where the embedding condition c_i is zeroed out with a certain probability. This encourages robustness in both conditional and unconditional scenarios. The final predicted noise is:

$$\hat{\epsilon}_\theta(x_t, c_t, c_i, t) = w \cdot \epsilon_\theta(x_t, c_t, c_i, t) + (1 - w) \cdot \epsilon_\theta(x_t, t) \quad (3)$$

Here, $w \in [0, 1]$ controls the influence of the embedding. When $w = 1$, the model fully conditions on c_i ; when $w = 0$, it behaves unconditionally. Through this training strategy, the IP-Adapter learns to map genetic embeddings into a semantic space interpretable by the diffusion model. It effectively acts as a biological decoder, guiding the synthesis of realistic child portraits that reflect inherited identity cues from both parents.

Table I: Sample regional statistics before and after augmentation.

Region	Before	After	% Before	% After
Europe	2400	2400	33.3%	23.5%
Asia	1100	2700	15.3%	26.4%
Africa	700	2111	9.7%	20.7%
North America	1800	1800	25.0%	17.6%
South America	1200	1200	16.7%	11.8%

IV. EXPERIMENTAL RESULTS

This section details the dataset construction, including the collection and preprocessing steps from both public benchmarks and curated online sources, to ensure input diversity and quality. We then present experimental settings and comparative evaluations, demonstrating the effectiveness of ZenID in predicting child facial features from parental images, which outperform existing methods in both accuracy and realism.

A. Dataset Collection and Augmentation

Each training sample must include a pair of parent images as input and a corresponding child image as output (Fig 1). However, collecting sufficient high-quality data across diverse ethnicities and regions remains a significant challenge. To address this, we integrate and extend multiple public datasets:

FF-DataBase. Introduced in [3], FF-DataBase contains 7,488 aligned triplets (father, mother, child) of 128×128 face images across diverse ethnicities and continents. It serves as the primary dataset for training.

KinFaceW. KinFaceW-I/II provides paired images (e.g., father-son, mother-daughter). Since full triplets are unavailable, we generate pseudo-pairs by duplicating the parent image, treating it as both father and mother.

Parent-Child in the Wild. PCW¹ includes 766 family triplets with wide ethnic variation. It complements FF-DataBase in capturing facial diversity.

Data Balancing and Augmentation. The combined dataset exhibited geographic imbalance, primarily favoring Europe and North America. To enhance generalization, we supplemented it with additional samples from Asia and Africa via open sources. For augmentation, we applied a series of transformations including horizontal flipping, random cropping, and color jittering to simulate natural variation in pose, lighting, and appearance. We also adopted mild geometric perturbations (e.g., rotation $\pm 10^\circ$ and scaling) to increase robustness without distorting facial identity. Then, after alignment and quality filtering, the final dataset includes 10,211 training samples. Table I illustrates the improved regional diversity after augmentation ².

¹<https://github.com/ghunkins/Child-Face-Generation>

²Despite these efforts, residual regional bias and unmodeled non-hereditary traits remain, which may affect realism. We leave these challenges for future dataset expansion and model design.

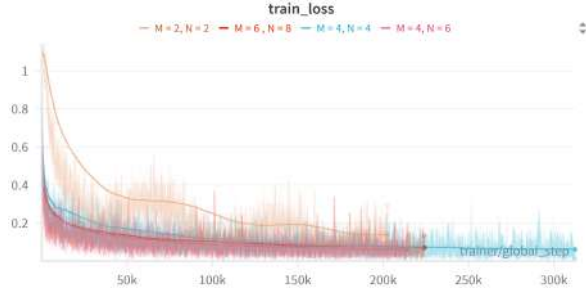


Figure 2: Grid search results over different combinations of M and N . The pink line highlights the optimal configuration $M = 4, N = 6$, which yields the lowest validation loss and best training stability.

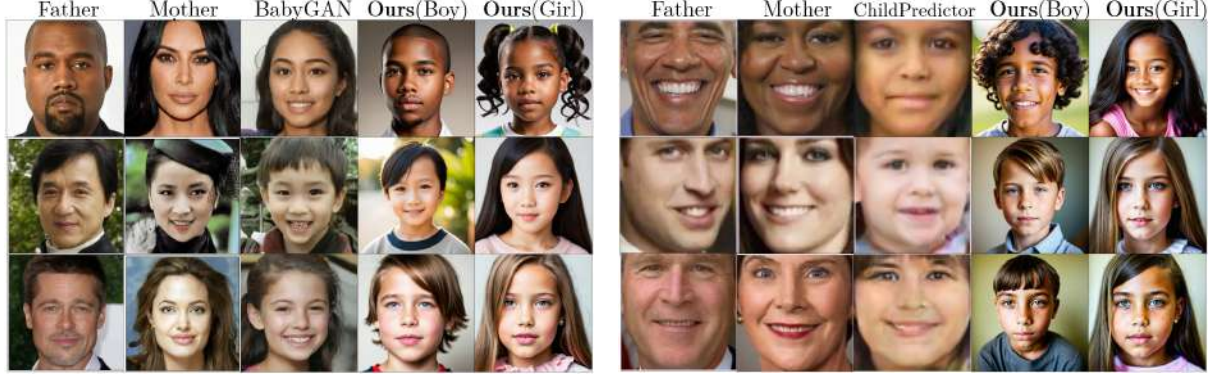
B. Model Training

Training FaceFusionNet. The first stage of the ZenID pipeline learns to encode a child’s facial representation from the input images of their parents. As discussed in Section III, the training objective is a composite loss combining MSE and Cosine similarity, weighted equally with $\lambda_1 = \lambda_2 = 0.5$, encouraging both numerical closeness and directional alignment between parent-child embeddings and ensuring a structurally coherent latent space. FaceFusionNet includes two key architectural hyperparameters: the number of encoder layers M , and the number of residual blocks N . A grid search was conducted to determine the optimal configuration. As shown in Fig 2, the setting $M = 4, N = 6$ yielded the lowest training loss and the fastest convergence, and was therefore selected for the final model.

Fine-tuning the IP Adapter. During this stage, the FaceFusionNet parameters are frozen to preserve the learned facial representation. The IP-Adapter is trained using the loss function defined in (2). The key consideration during this stage is to avoid overfitting to embedding-level loss, which could hinder the generative diversity of the diffusion model. Excessive optimization may lead to rigid, unnatural outputs that overfit training features rather than capturing plausible child facial characteristics. To mitigate this, we adopt a regularized fine-tuning strategy that maintains a balance between embedding accuracy and visual realism.

C. Comparison with Existing Methods

1) *Qualitative evaluation:* To evaluate the effectiveness of ZenID, we compare it with two prominent baseline models: ChildPredictor [3] and BaByGAN [4]. For fair comparison, we use the same parent image pairs from the original publications of BaByGAN and ChildPredictor, and apply ZenID to synthesize child faces. Unlike the baselines, ZenID allows for explicit control over the child’s gender via input conditioning, offering greater flexibility for practical applications. As illustrated in Fig 3, ZenID produces sharper, more realistic images with facial traits that clearly reflect those of both parents. Structural features such as



(a) Comparison between BaByGAN and ZenID. ZenID generates faces with clearer inheritance of bone structure, eyes, and mouth shape.

(b) Comparison between ChildPredictor and ZenID. ZenID produces more structurally consistent details such as eye spacing, brow curvature, and chin shape.

Figure 3: Visual comparison between ZenID and existing methods using the same parent inputs.

Table II: Cosine Similarity between synthesized child faces and real identities.

Method	Father	Mother	Real Child
BaByGAN [4]	0.31	0.29	0.37
ChildPredictor [3]	0.22	0.23	0.43
ZenID (Ours)	0.42	0.39	0.41

jawline, eye shape, and facial proportions are better preserved.

2) *Cosine similarity analysis*: To quantitatively measure similarity, we compute cosine similarity³. As shown in Table II, ZenID achieves the highest similarity to both parents and ranks second in resemblance to the real child, closely trailing ChildPredictor. BaByGAN exhibits lower similarity scores across all metrics.

3) *Key Advantages of ZenID (Fig 4)*: Beyond accuracy, ZenID introduces practical innovations that enhance flexibility, user control, and applicability in real-world creative tasks.

Gender and Age Control. ZenID allows users to specify the desired gender and age of the synthesized child, from infant to adult, using conditioning inputs. This capability enables realistic aging simulations and personalized predictions, which are not supported by prior models like BaByGAN and ChildPredictor.

Output Size and Format Flexibility. Leveraging the capabilities of Stable Diffusion, ZenID can generate images in various formats such as square, portrait, or landscape, meeting diverse requirements across social media, print, or visual content creation.

Style and Background Control via Text Prompts. ZenID supports dynamic control of background, clothing, facial expressions, and artistic styles through text prompts. This enables users to generate highly cus-

tomized outputs suited for storytelling, family visualization, or digital media.

D. Deployment and Community Adoption

ZenID is publicly released as an open-source module integrated into the ComfyUI framework, leveraging the intuitive node-based interface for visual pipeline construction. The system is built atop Stable Diffusion XL (SDXL) and requires a high-performance GPU (e.g., NVIDIA H100 SXM with 24GB+ VRAM) to support modules such as the fine-tuned IP-Adapter. Since its release, ZenID has gained notable traction in the AI community. Its GitHub repository⁴ has received nearly 200 stars, with a steady rate of daily clones and usage. The model has been featured by content creators across platforms such as YouTube⁵ and technical blogs, highlighting its strengths in identity preservation and user-controlled facial synthesis. Furthermore, ZenID has been adopted in several large-scale creative software platforms, demonstrating its practical value beyond academic research through robust image quality and flexible generative control.

V. CONCLUSION

This work introduces ZenID, a framework for child face synthesis conditioned on parental facial features. At its core, ZenID employs a dual-stream encoder architecture, FaceFusionNet, which effectively fuses latent identity representations from both parents into a unified child embedding. To enhance identity preservation during generation, we fine-tune the IP-Adapter module, enabling improved alignment between the synthesized identity and parental inputs. Quantitative evaluation shows that ZenID achieves strong identity consistency with both parents while maintaining plausible generative diversity. The entire pipeline is implemented within the ComfyUI environment to facilitate reproducibility and modular experimentation.

³Cosine similarity was adopted due to its strong alignment with identity embedding models. Other metrics (e.g., L2 distance, FID) could be explored in future extensions, but we found cosine sufficient and reliable for this setting.

⁴https://github.com/vuongminh1907/ComfyUI_ZenID

⁵<https://youtu.be/UnFK-SjkIS0?si=EqrhldfpMtrmhuub>

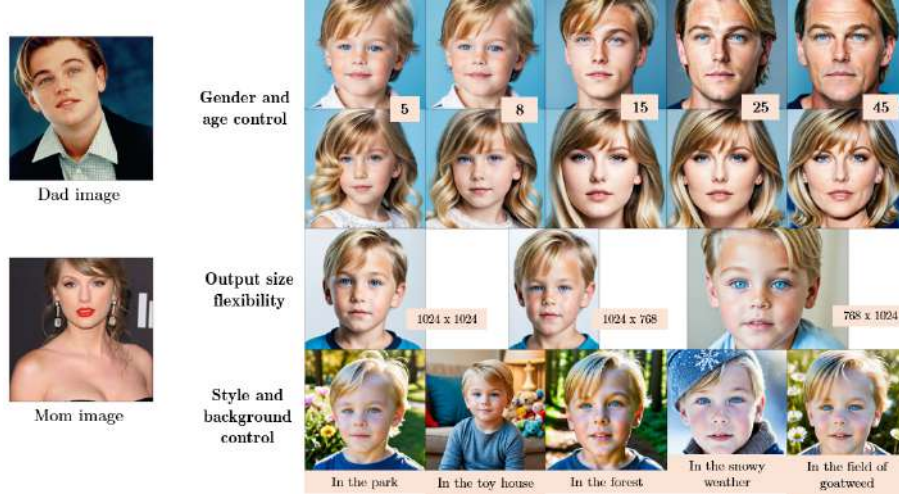


Figure 4: Face predictions by ZenID across genders, age ranges, resolutions, and backgrounds using prompt-based control.

Future direction will explore integration with higher-capacity diffusion backbones such as Flux [21], and further refine the conditioning mechanism through semantic and regional prompt controls. We also aim to investigate multi-modal extensions and architectural refinements to improve both fidelity and generalizability in hereditary face synthesis tasks.

REFERENCES

- [1] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. Vol. 27. 2014, pp. 2672–2680.
- [2] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *Proceedings of the IEEE/CVPR*. 2022, pp. 10684–10695.
- [3] Yuzhi Zhao et al. “ChildPredictor: A child face prediction framework with disentangled learning”. In: *IEEE Transactions on Multimedia* 25 (2022), pp. 3737–3752.
- [4] Huan Yao, Wanying Bao, and Hequn Wu. “BabyGAN for Facial Contour Reversion: AI Course Applications Using U-Net Architecture”. In: *Proceedings of the 2024 International Conference on Computer and Multimedia Technology*. 2024, pp. 324–329.
- [5] Hu Ye et al. “IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models”. In: *arXiv preprint arXiv:2308.06721* (2023).
- [6] Jiankang Deng et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *CVPR*. URL: <https://arxiv.org/abs/1801.07698>. 2019, pp. 4685–4694.
- [7] Yin Hu et al. “InsightFace: A 2D and 3D Face Analysis Toolkit”. In: *arXiv preprint arXiv:1905.01248* (2020).
- [8] Savas Özkan and Akin Özkan. “KinshipGAN: Synthesizing of Kinship Faces from Family Photos by Regularizing a Deep Face Network”. In: *2018 IEEE International Conference on Image Processing (ICIP)*. 2018, pp. 207–211.
- [9] Pengyu Gao et al. “What Will Your Child Look Like? DNANet: Age and Gender Aware Kin Face Synthesizer”. In: *arXiv preprint arXiv:1911.07014* (2019).
- [10] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVPR*. 2019, pp. 4401–4410.
- [11] Che-Hsien Lin et al. “StyleDNA: A High-Fidelity Age and Gender Aware Kinship Face Synthesizer”. In: *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 2021, pp. 1–8.
- [12] Hao Li et al. “StyleGene: Crossover and Mutation of Region-Level Facial Genes for Kinship Face Synthesis”. In: *Proceedings of the IEEE/CVPR*. 2023, pp. 20960–20969.
- [13] Pin-Yen Chiu et al. “StyleDiT: A Unified Framework for Diverse Child and Partner Faces Synthesis with Style Latent Diffusion Transformer”. In: *arXiv preprint arXiv:2412.10785* (2024).
- [14] Nataniel Ruiz et al. “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation”. In: *Proceedings of the IEEE/CVPR*. 2023, pp. 22500–22510.
- [15] Edward J. Hu et al. “LoRA: Low-rank adaptation of large language models”. In: *International Conference on Learning Representations (ICLR)*. 2022.
- [16] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021.
- [17] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations (ICLR)*. 2021.
- [18] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: [2112.10752](https://arxiv.org/abs/2112.10752) [cs.CV].
- [19] Qixun Wang et al. “Instantid: Zero-shot identity-preserving generation in seconds”. In: *arXiv preprint arXiv:2401.07519* (2024).
- [20] Junnan Li et al. “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation”. In: *International Conference on Machine Learning (ICML)*. 2022.
- [21] Black Forest Labs. “FLUX: A Pre-Trained Large-Language-and-Vision-Model for talking about and seeing the world”. In: *arXiv preprint arXiv:2405.20454* (2024).

Towards Real-ESRGAN Refinement for Realistic Image Restoration via Advanced Degradation and Knowledge Distillation

LE Hoai Nam¹, DOAN Nhat Quang¹, GIANG Anh Tuan¹,
TRAN Hoang Tung¹, LE Huu Ton², NGUYEN Hoang Ha^{*1}

Abstract—Deep learning has become a powerful paradigm for image quality enhancement, yet restoring images degraded by complex real-world factors such as noise, blur, and compression remains challenging. To address this, we propose a framework for realistic image restoration. Our contributions are twofold. First, we design an advanced degradation pipeline that generates diverse and realistic training data through randomized operation ordering, refined parameter ranges, and explicit camera noise simulation. Second, we introduce a lightweight Student model distilled from the original Real-ESRGAN via a Teacher–Student strategy, substantially reducing computational cost while preserving restoration quality. Experimental results demonstrate that our method produces outputs more consistent with human visual perception and generalizes effectively to real-world degradations. This framework offers a practical solution for applications such as photo enhancement, medical imaging, and surveillance analysis. Source code is available at https://github.com/namle24/RealESRGAN_Degradation_KD.git.

Index Terms—Image Quality Enhancement, Super-Resolution, Real-ESRGAN, Image Degradation, Knowledge Distillation

I. INTRODUCTION

Image quality enhancement with deep learning combines image processing and feature learning to transform low-quality images—affected by low resolution, blur, and noise—into high-quality counterparts with sharper details and improved realism. This task has broad applications across media, healthcare, security, and cultural heritage, where visually pleasing yet realistic outputs are critical. Unlike some AI-based reconstructions that yield oversmoothed or artificial results, our approach prioritizes natural and perceptually consistent enhancements.

Recent advances, such as the Video Restoration Transformer [1], achieve state-of-the-art restoration but incur high computational cost and long inference times. GAN-based methods like ESRGAN [2] and BSRGAN [3] improve efficiency, yet their reliance on bicubic downsampling or synthetic degradations limits robustness to real-world artifacts, including noise, blur, and compression.

To address these limitations, we propose a novel framework with two main contributions. First, we introduce an improved Realistic High-Order Degradation process that generates diverse and realistic synthetic datasets through randomized degradation order, refined parameter ranges, and explicit camera-noise simulation. Second, we employ a Teacher–Student knowledge distillation framework to develop a lighter, more efficient variant of Real-ESRGAN. The Student network learns from a pre-trained Teacher, reducing computational cost while maintaining high-quality restoration. Together, these contributions enable our model to handle complex real-world degradations and support deployment in real-time, resource-constrained environments.

II. RELATED WORK

Early work in this field began with the Super-Resolution Convolutional Neural Network (SRCNN) [4], which laid the foundation for deep learning in super-resolution. To improve perceptual quality, researchers later introduced generative adversarial networks (GANs), though most methods relied on bicubic interpolation to simulate low-quality images. More recently, advanced approaches have emerged, leveraging transformers [5], [6], reinforcement learning [7], or unsupervised learning [8]. Residual-based architectures also proved effective, with EDSR [9] improving performance by removing Batch Normalization layers, and RDN [10] further validating the role of residual dense connections.

Research then shifted toward blind SR [11]–[13], which aims to handle diverse and unknown degradations such as blur, noise, and compression. Another direction involves collecting paired high- and low-quality images captured by the same camera [14], [15], though this approach is limited to specific capture settings and cannot fully represent real-world degradations.

To overcome these challenges, Xintao et al. [16] proposed Real-ESRGAN, which introduces higher-order degradation models to generate more realistic training pairs. While not the highest-performing method today, Real-ESRGAN demonstrates strong adaptability to real-world images, making it well-suited for multi-task image enhancement and detail reconstruction. Building on this foundation, our study aims to

¹ University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, 18, Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam

² CMC University, Hanoi, Vietnam

*Corresponding author: nguyen-hoang.ha@usth.edu.vn

extend Real-ESRGAN by diversifying degradations, training on larger datasets, and further optimizing the model for practical applications.

III. METHODOLOGY

This study proposes an image restoration framework that combines an improved Realistic High-Order Degradation process and a Teacher–Student knowledge distillation architecture. The goal is to handle multiple real-world degradations while reducing computational cost. An overview of the proposed pipeline is shown in Fig. 1.

A. Data Augmentation by Realistic High-Order Degradation Method

Most existing image restoration methods, including Real-ESRGAN [16], generate synthetic low-quality (LQ) images from high-quality (HQ) data using a two-stage degradation pipeline. Each stage applies downsampling, blurring, noise injection, and compression in a fixed order, with an optional second stage. While effective, this strategy still falls short of capturing the variability and complexity of real-world degradations.

Our work retains the general two-stage process (resizing, blurring, noise addition, JPEG compression, and camera-specific noise; see Fig. 2) but introduces several improvements for more diverse and realistic degradations:

- **Randomized Degradation Order:** Unlike the fixed sequence in [16], we adopt a flexible order (with resizing prioritized first), better reflecting unpredictable imaging conditions.
- **Refined Parameter Ranges:** JPEG compression is applied within narrower ranges to avoid extreme artifacts and better match realistic imaging scenarios.
- **Explicit Camera Noise:** We incorporate chromatic aberration and sensor-specific pattern noise to simulate artifacts from smartphones, webcams, and low-quality lenses.
- **Pre-Generated Degradation:** Instead of on-the-fly degradation, we pre-generate LQ–HQ pairs, ensuring reproducibility, full control over severity, and faster training.

This pipeline produces a diverse, realistic dataset crucial for training robust models that generalize to real-world imperfections. Examples of generated LQ images are shown in Fig. 3.

Degraded LQ images are generated at one-quarter the resolution of their GT counterparts to enable $\times 4$ super-resolution training. To preserve degradation characteristics and improve convergence, GT images are cropped into 240×240 patches, while LQ images are cropped into 120×120 patches (Fig. 4). For training, the LQ images serve as the input for the network, while the GT images are used as a reference for the discriminator.

B. Teacher-Student Technique Architecture

After obtaining the training dataset from the Realistic High-Order Degradation process, we further enhance efficiency

using a Teacher–Student knowledge distillation framework. Beyond serving as data augmentation, the degradation system enables the model to learn from LQ cases that closely resemble real-world scenarios. Unlike the original model, which learns from on-the-fly generated data, our approach directly leverages the pre-generated Realistic High-Order Degradation dataset, ensuring more accurate and controllable outcomes.

To reduce complexity and computational cost, we design a lighter Student network by simplifying the Teacher architecture. The Student is trained on a pre-trained Teacher model, inheriting its knowledge while being optimized for efficiency. The objective is for the Student to achieve competitive or improved performance metrics compared to training independently, while remaining more suitable for deployment.

Teacher Network

The Teacher network, which is based upon the original Real-ESRGAN architecture [2], [16], is trained on 340,508 images generated by our degradation pipeline. This ensures that the Teacher is capable of handling real-world degraded samples, providing high-quality HR output for distillation. We use this training session as a Teacher for our Student model.

Student Network

To reduce computational cost while maintaining performance, we redesigned the Student model as a lighter, more efficient variant of the Teacher. It follows the Teacher’s core design but is optimized for deployment on resource-limited devices. Using knowledge distillation, it preserves reconstruction quality while reducing complexity—from 23 RRDBs (64 channels) to 16 RRDBs (48 channels). Built on the RRDBNet backbone, it also integrates a discriminator for perceptual enhancement.

The input layer maps an LQ RGB image to 48 channels via a 3×3 convolution. The core extractor consists of 16 RRDBs, each with three RDBs of five convolutions and 24 growth channels, using dense connections and residual scaling (0.2) for stability. Intermediate features from selected RRDBs (e.g., 2, 4, 6, 8) support feature-based distillation.

Upsampling employs two PixelShuffle layers, progressively enlarging the feature maps to $4 \times$ scale, followed by a convolution to refine details. The final 3×3 convolution outputs the RGB super-resolution image without activation.

The Student is trained in a GAN framework with combined Student–Teacher losses, balancing perceptual quality and reconstruction fidelity.

- 1) **Input:** Both Teacher and Student receive LQ images as input, ensuring uniformity during training.
- 2) **Output:** Generate Student High-Quality image (S_{HQ}) and intermediate features (S_{feat}), then compare with the output of Teacher Network (Teacher High-Quality (T_{HQ}) and T_{feat}).

Loss Calculation:

- **Pixel-wise Distillation Loss:** $\mathcal{L}_{distill\_pixel} = L1(S_{HQ} - T_{HQ})$ enables the student to learn directly from the teacher’s output.

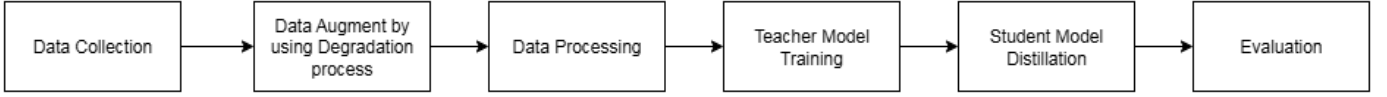


Fig. 1. Overview of the entire pipeline

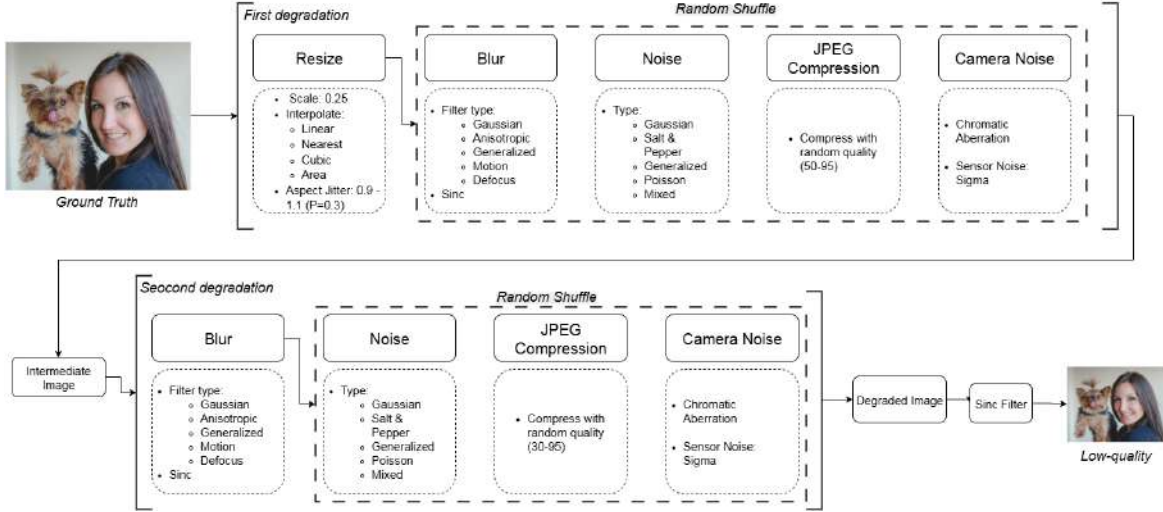


Fig. 2. Our realistic high-order degradation scheme.

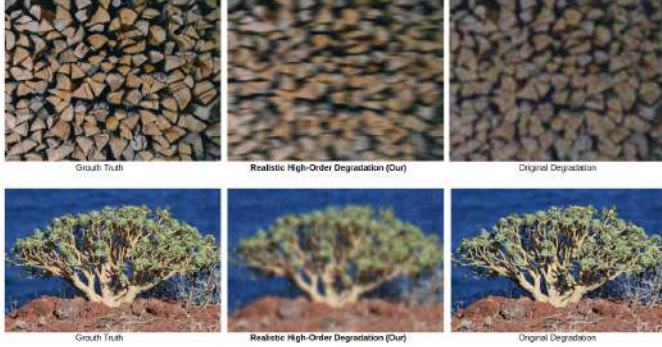


Fig. 3. Comparison between the image created by Realistic High-Order Degradation and the author's original degradation

- **Feature-based Distillation Loss:** $\mathcal{L}_{\text{distill\_feat}} = L1(S_{\text{feat}} - T_{\text{feat}})$ leverages knowledge from intermediate features.
- **Total Loss:**

$$\begin{aligned} \mathcal{L}_{\text{total}} = & w_{\text{pixel}} \cdot \mathcal{L}_{\text{pixel}} \\ & + w_{\text{perceptual}} \cdot \mathcal{L}_{\text{perceptual}} + w_{\text{distill\_pixel}} \cdot \mathcal{L}_{\text{distill\_pixel}} \\ & + w_{\text{distill\_feat}} \cdot \mathcal{L}_{\text{distill\_feat}} + w_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}} \end{aligned}$$

where the weights w are adjusted during training.

For example: $w_{\text{pixel}} = 0.7$, $w_{\text{distill\_pixel}} = 0.5$

The gradient from L_{total} is backpropagated through the Student Network using the Adam optimizer, while the Teacher remains in evaluation mode with `torch.no_grad()`. This setup optimizes resources, stabilizes knowledge distillation,

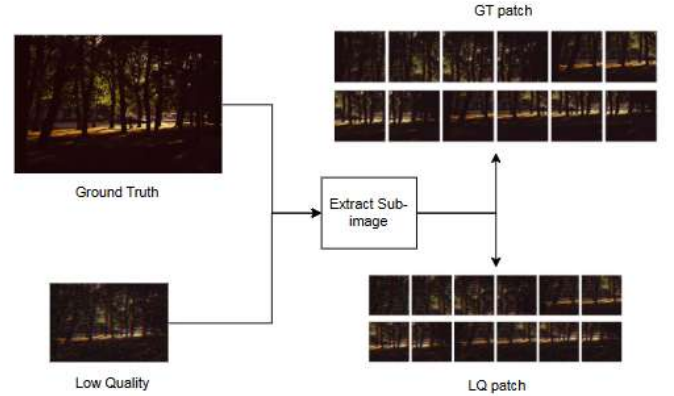


Fig. 4. Illustration of cropped image into small images in each patch to preserve the image properties

and ensures a clear role division: the Teacher provides guidance, while the Student adapts based on the losses. With its optimized RRDBNet backbone (16 RRDBs, 48 channels), the Student demonstrates the effectiveness of teacher-student architectures for real-world image enhancement. By combining knowledge distillation with GAN training, it inherits the Teacher's expertise while improving perceptual realism. Initial results show strong potential, and further refinements may yield an even more practical solution.

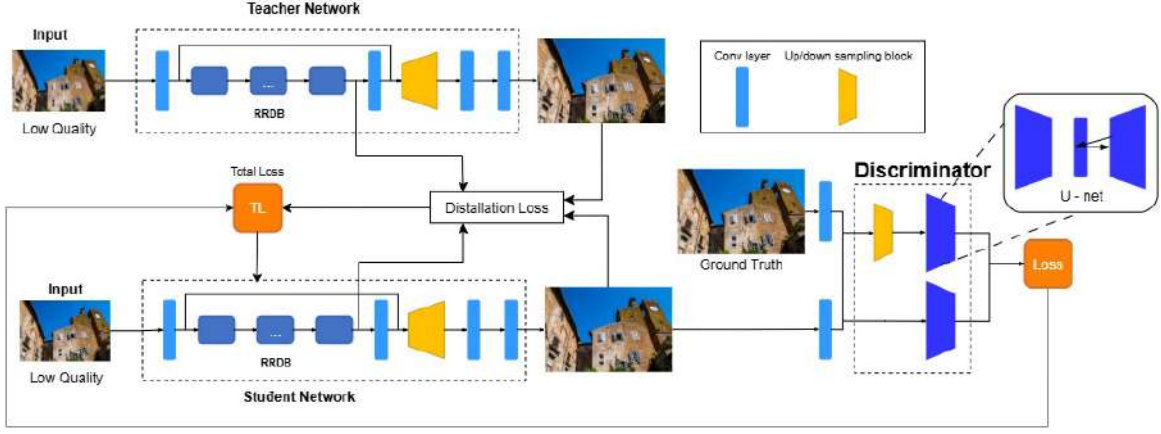


Fig. 5. Our Real-ESRGAN-based distillation architecture(Reference drawing from [17]).

IV. EXPERIMENTS

A. Data

This study sourced images from DF2K, Unsplash, and RealSR. DF2K comprises three popular NTIRE 2017 challenge datasets, DIV2K [18] (800 training, 100 validation, 100 test images), OutdoorScene [19] (8,984 training, 300 test images), and Flickr2K [20] (2,650 images), with resolutions around 2040×1356 and varied aspect ratios. We further included 1,254 images from Unsplash and 192 from RealSR [21]. While DF2K and Unsplash provide paired HQ–LQ images, their LQ data is generated via bicubic downsampling and thus lacks the complexity of real-world degradations. Therefore, we retained 14,038 HQ images, applied our degradation pipeline to create LQ counterparts, and used these for training. Model evaluation was conducted on four benchmarks: Set5 [22], Set14 [20], and RealSR [21].

B. Training Details

We follow the training configuration from the official Real-ESRGAN repository, with a patch size of We follow the training configuration from the official Real-ESRGAN repository, with a patch size of 256 and Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$). The Teacher model is trained for 800,000 iterations with a learning rate of 1×10^{-4} , using L1 loss, GAN loss, and perceptual loss based on the pre-trained VGG19 [23]. The GAN loss is weighted by 0.1, while pixel and perceptual losses are set to $\{1.5, 1.2\}$, balancing PSNR/SSIM with NIQE stability.

For the Student, we additionally apply distillation losses between Teacher and Student outputs, with pixel and feature distillation weights of $\{0.7, 0.5\}$. It is also trained for 800,000 iterations, enabling the Student to follow the Teacher closely, thus reducing NIQE without degrading PSNR or SSIM. Training is conducted on NVIDIA® GeForce RTX 2080 GPUs.

C. Qualitative Results

We employ three metrics to evaluate our framework. The Peak Signal-to-Noise Ratio (PSNR) measures pixel-level fi-

delity in decibels (dB), with higher values indicating better accuracy. The Structural Similarity Index (SSIM) assesses luminance, contrast, and structural similarity, ranging from -1 to 1 , and aligns more closely with human perception. The Natural Image Quality Evaluator (NIQE) is a no-reference metric; lower scores reflect greater perceptual realism.

Unlike many recent approaches that rely on bicubic inputs or focus solely on super-resolution, Real-ESRGAN handles multiple degradations simultaneously and has shown superior results [16]. Table I compares the baseline model and our fine-tuned Student model. As the authors mentioned in [16], metrics cannot accurately represent the performance of models and human visual perception in practice. Therefore, we also provide qualitative examples in Figure 6 for a more objective view.

While Real-ESRGAN achieves strong sharpness recovery, our enhanced model produces clearer images with more natural colors and improved visual comfort. For instance, in Nikon_001, EDSR and ESRGAN fail to remove blur, and SwinIR or BSRGAN introduce artifacts such as unrealistic details or color shifts (e.g., yellow cast, oversmoothing). In contrast, our method effectively removes degradations while preserving natural textures. Compared to RealSR and Transformer-based SwinIR, our model consistently balances perceptual quality and quantitative performance on real-world images.

Compared to the original 16.7M parameters, we reduced it to 6.5M parameters. Table II gives a more objective view of the student’s performance, we evaluated both Real-ESRGAN and Real-ESRGAN student on the same RTX 3090 GPU. The student model consistently achieved faster inference times on all benchmark datasets (Set5, Set14, RealSR_V3, and General100) while maintaining comparable performance. This demonstrates that the student model is more efficient for large-scale or real-time applications.

TABLE I
PERFORMANCE COMPARISON BETWEEN THE ORIGINAL REAL-ESRGAN AND THE DISTILLED STUDENT MODEL.

Model	Dataset	Set5		Set14		RealSR_V2		RealSR_V3	
		PSNR/SSIM	NIQE	PSNR/SSIM	NIQE	PSNR/SSIM	NIQE	PSNR/SSIM	NIQE
Real-ESRGAN	Original Degradation Dataset	26.4532/0.8082	10.8999	24.6748/0.6939	6.8784	25.9410/ 0.7922	6.7829	26.7976/0.8069	6.5549
Real-ESRGAN student	Our Degradation Generated Dataset	25.1896/0.7670	6.5374	23.8923/0.6541	5.8193	25.9630/0.7846	5.9161	27.3028/0.8068	6.0984

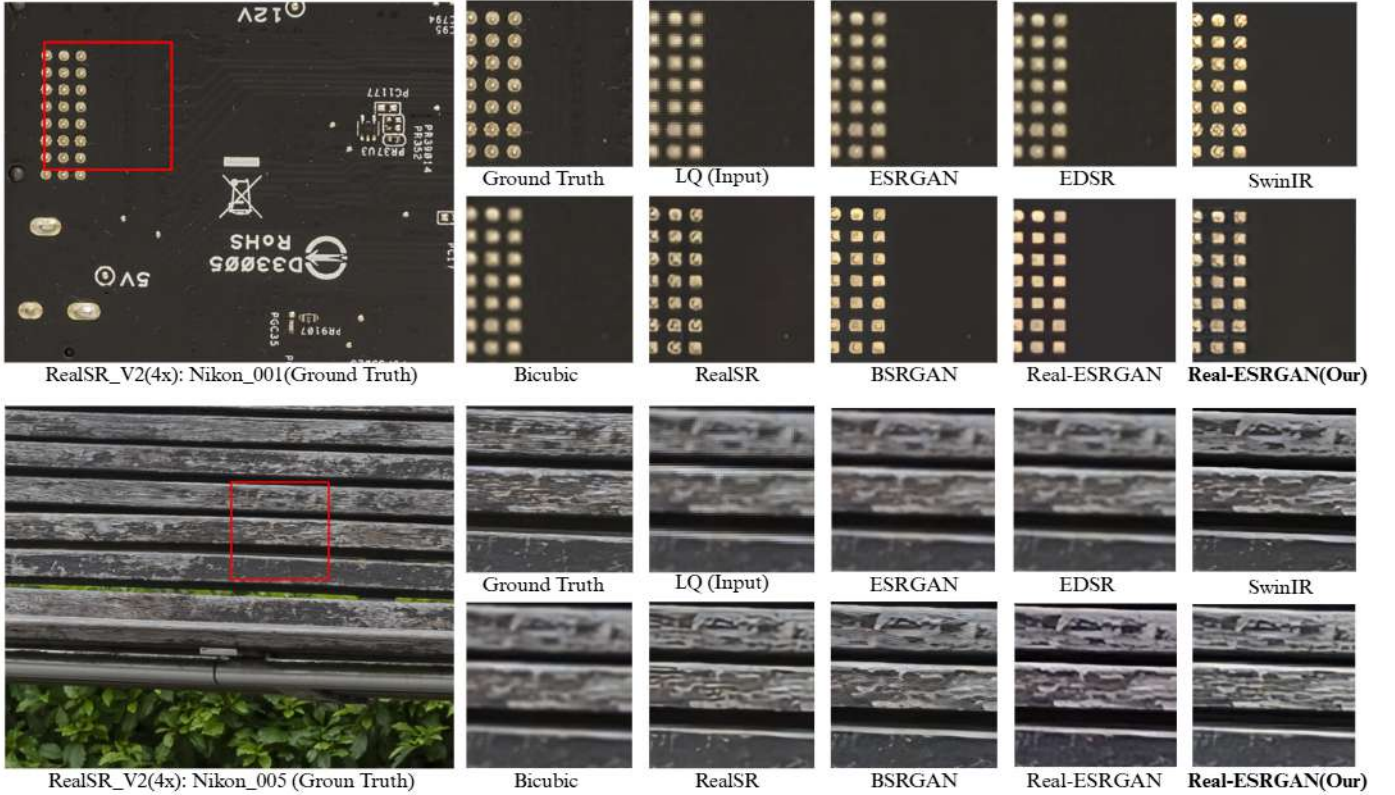


Fig. 6. Qualitative comparisons on real-world images with an upsampling scale factor of 4.

TABLE II
COMPARISON OF INFERENCE TIME AND MEMORY USAGE BETWEEN REAL-ESRGAN AND REAL-ESRGAN STUDENT ON RTX 3090.

Benchmark Datasets	Quantity	Real-ESRGAN		Real-ESRGAN student	
		Inference time (s)	Memory Usage (MB)	Inference time (s)	Memory Usage (MB)
Set5	5	5.35	556	5.18	460
Set14	14	6.18	684	5.76	566
RealSR_V3	50	18.83	6434	15.98	4880
General100	100	12.42	1180	11.09	878

D. Limitations

Overall, our method enables the model to handle most real-world low-quality images effectively. However, challenges remain with highly complex degradations (see Fig. 7). While the results still provide a pleasant viewing experience, complete clarity cannot always be achieved. Additional limitations are

illustrated in Fig. 8: (1) severely degraded details may appear fused together, preventing full recovery; (2) some restorations produce artifacts that look unrealistic; and (3) aliasing effects can distort fine structures after restoration. Addressing these limitations is essential for future work to achieve robustness against all forms of real-world degradations.



Fig. 7. Limited to super complex images. Zoom in for the best view.

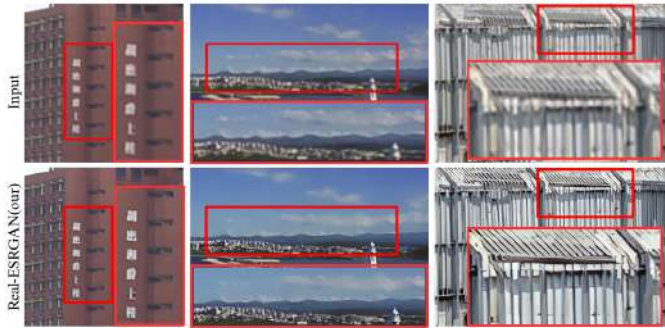


Fig. 8. Limited to super complex images. Zoom in for the best view.

V. CONCLUSION

This study advances Real-ESRGAN by coupling an improved high-order degradation pipeline with a Teacher-Student knowledge distillation framework for real-world image restoration. The proposed pipeline introduces randomized operation ordering, refined parameter ranges, explicit camera noise simulation, and pre-generated degradation pairs to create more diverse and realistic training data. A lightweight Student model distilled from a high-capacity Teacher achieves comparable reconstruction quality with substantially lower computational cost. Experimental results confirm improved perceptual realism, robust handling of diverse degradations, and strong generalization to real-world scenarios. Future research will focus on overcoming limitations in extreme degradations and further optimizing the architecture for real-time use on resource-constrained devices.

REFERENCES

- [1] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *IEEE Transactions on Image Processing*, vol. 33, pp. 2171–2182, 2024.
- [2] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "EsrGAN: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.
- [3] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4791–4800.
- [4] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.
- [5] D. Dutta, D. Chetia, N. Sonowal, and S. K. Kalita, "State-of-the-art transformer models for image super-resolution: Techniques, challenges, and applications," *arXiv preprint arXiv:2501.07855*, 2025.
- [6] C.-C. Hsu, C.-M. Lee, and Y.-S. Chou, "Drct: Saving image super-resolution away from information bottleneck," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6133–6142.

- [7] K. Yu, C. Dong, L. Lin, and C. C. Loy, "Crafting a toolchain for image restoration by deep reinforcement learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2443–2452.
- [8] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 701–710.
- [9] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.
- [10] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.
- [11] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3262–3271.
- [12] Y. Huang, S. Li, L. Wang, T. Tan *et al.*, "Unfolding the alternating optimization for blind super resolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5632–5643, 2020.
- [13] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1604–1613.
- [14] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *European conference on computer vision*. Springer, 2020, pp. 101–117.
- [15] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3086–3095.
- [16] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.
- [17] Z. Wei, Y. Huang, Y. Chen, C. Zheng, and J. Gao, "A-esrgan: Training real-world blind super-resolution with attention u-net discriminators," 12 2021.
- [18] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1122–1131.
- [19] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 606–615.
- [20] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, X. Wang, Y. Tian, K. Yu, Y. Zhang, S. Wu, C. Dong, L. Lin, Y. Qiao, C. C. Loy, and Q. Guo, "Ntire 2017 challenge on single image super-resolution: Methods and results," 07 2017, pp. 1110–1121.
- [21] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [22] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi-Morel, "Low-complexity single image super-resolution based on nonnegative neighbor embedding," 09 2012.
- [23] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

Data Generation Based on Multimodal Language Models for Vietnamese Visual Question Answering

1st Truong Trong-Dai Le

Faculty of Computer Science

University of Information Technology University of Information Technology

Ho Chi Minh City, Vietnam

22521576@gm.uit.edu.vn

2nd Quan Hoang Ngoc

Faculty of Computer Science

University of Information Technology

Ho Chi Minh City, Vietnam

22521178@gm.uit.edu.vn

3rd Nghia Hieu Nguyen

Faculty of Information Science and Engineering

University of Information Technology

Ho Chi Minh City, Vietnam

nghiangh@uit.edu.vn

4th Kiet Van Nguyen

Faculty of Information Science and Engineering

University of Information Technology

Ho Chi Minh City, Vietnam

kietnv@uit.edu.vn

Abstract—Visual Question Answering (VQA) has shown promising potential in education, particularly in supporting children through interactive learning. However, the lack of high-quality, large-scale Vietnamese VQA datasets, especially in primary education contexts, remains a major bottleneck for training accurate models. To address this critical gap, we propose an automated data labeling pipeline that leverages Multimodal Large Language Models (MLLMs) to generate high-quality question-answer pairs from Vietnam textbook images. Our method combines strong prompting techniques, such as Few-shot or Chain-of-Thought, with the reasoning power of models like Gemini 1.5 Pro, Gemini 2.0 Flash, and Vintern-1B to ensure both the accuracy and naturalness of the generated data. Based on this pipeline, we introduce ViEduVQA, a new dataset comprising over 18,000 Vietnamese question-answer (QA) pairs derived from elementary school textbooks in Vietnam. Moreover, we also built a human-refined benchmark (ViEduVQA-benchmark) to evaluate data quality and model performance. Experimental results show that models trained on ViEduVQA achieve strong performance, validating the effectiveness of our pipeline and dataset. This work lays the foundation for generating scalable and cost-effective VQA datasets across various application domains.¹

Index Terms—Visual Question Answering, Primary Education, Vietnamese Language, Multimodal AI.

I. INTRODUCTION

Visual Question Answering (VQA) [1] is a fundamental task in multimodal Artificial Intelligence (AI), where a system is asked to generate a natural language answer based on an input image and a related question. With its ability to combine visual perception and linguistic reasoning, VQA has shown strong potential in a wide range of real-world applications, including healthcare, retail, and especially education. In the context of elementary education, VQA systems can create engaging and interactive learning environments by allowing students to ask and answer questions about visual content, such as textbook images. This not only fosters active learning but also supports the development of observation, logical

reasoning, and language skills, particularly valuable for young learners who often struggle with attention when interacting with screen-based educational materials.

Despite the potential, VQA research in domain-specific contexts like education remains limited, especially for low-resource languages such as Vietnamese. Existing large-scale VQA datasets are primarily designed for general-purpose use and are predominantly available in English (e.g., VQAv2, GQA [2]). These datasets often fail to meet the linguistic, cognitive, and pedagogical requirements of primary school learners. Moreover, manually creating high-quality datasets tailored for Vietnamese elementary education is costly and time-consuming, as it requires domain knowledge, age-appropriate language, and consistent annotation across thousands of image-question-answer (I-Q-A) samples.

Recent advancements in Multimodal Large Language Models (MLLMs), such as Gemini [3] or GPT-4V [4], have opened new opportunities for automating the data labeling process. These models exhibit strong capabilities in image understanding and natural language generation, allowing them to serve as effective tools for scalable dataset creation. By leveraging these strengths, it becomes feasible to build automated labeling pipelines for VQA, which can drastically reduce the burden of manual annotation while maintaining reasonable quality and diversity.

In this work, we propose a novel data labeling pipeline that harnesses the power of MLLMs and Vision-Language Models (VLMs) to automatically generate VQA datasets in Vietnamese, targeting the domain of elementary education. The pipeline includes modules for question generation, answer generation, answer fusion, and quality evaluation, each module supported by state-of-the-art models such as Gemini 1.5 Pro [5], Gemini 2.0 Flash, and Vintern-1B [6]. Using this pipeline, we construct a new dataset called **ViEduVQA**, consisting of over 18,000 high-quality question-answer pairs generated from textbook images. We further fine-tune a PaLI-Gemma [7] model on this dataset and evaluate its performance on a

¹The dataset and source code are available at https://github.com/ShouyiLee/ViEduVQA-VietnameseVQA_DataGen

manually curated benchmark to validate the effectiveness of our approach.

Our main contributions are as follows:

- We propose a modular, scalable, and low-cost pipeline for automated VQA data labeling in Vietnamese using MLLMs.
- We introduce **ViEduVQA**, which is a new VQA dataset for Vietnamese primary education, consisting of question-answer pairs generated from textbook images.
- We fine-tune the PaLI-Gemma model on ViEduVQA and demonstrate its effectiveness on downstream Vietnamese VQA tasks through extensive experiments and human evaluations.

II. RELATED WORK

A. Data Labeling Approaches

High-quality data annotation is fundamental to building robust Visual Question Answering (VQA) [1] systems. Depending on the level of automation and human involvement, existing approaches can be broadly categorized into three main types: Manual annotation, Rule-based labeling, and Automatic generation via large language models.

Manual annotation is the most traditional and reliable approach, where human annotators carefully generate question-answer (QA) pairs based on visual content and prior knowledge. While this method produces linguistically rich and pedagogically appropriate data, it is inherently expensive, time-consuming, and difficult to scale, particularly for low-resource languages like Vietnamese or domains requiring expert knowledge, such as primary education.

To address scalability, some studies adopt **rule-based** or **template-based** methods for generating QA pairs. While faster and more consistent, these approaches often lack linguistic naturalness and diversity, and are generally limited to simple factual or object-level questions. They also struggle to support higher-level reasoning or adapt to the semantic and cultural nuances found in educational materials.

Recently, there has been increasing attention on **Automated data labeling** using Multimodal Large Language Models (MLLMs) such as Gemini [3] and GPT-4V [4], which possess strong reasoning abilities and cross-modal understanding. These models can effectively interpret visual inputs and generate coherent, contextually relevant question-answer pairs, making them well-suited for automated annotation in VQA. Their ability to combine visual perception with natural language generation offers a scalable and cost-effective alternative to manual labeling, particularly in low-resource and domain-specific settings like Vietnamese primary education.

Motivated by these benefits, this study adopts automatic annotation using MLLMs as the primary method for constructing Vietnamese VQA datasets. Our approach leverages recent progress in MLLMs to design a structured labeling pipeline capable of generating high-quality QA pairs aligned with the pedagogical and linguistic requirements of elementary education. The following section reviews key MLLMs that serve as the foundation of our proposed pipeline.

B. Multimodal Large Language Models (MLLMs) for VQA

The recent emergence of Multimodal Large Language Models (MLLMs), such as Gemini, GPT-4V, and Qwen-VL, has significantly improved the ability of machines to understand and generate natural language grounded in visual content. These models combine powerful language understanding with strong visual perception, enabling them to perform a variety of multimodal tasks, including question generation, answer inference, and even semantic evaluation.

For Vietnamese VQA, a few open-source MLLMs have been introduced, such as Vintern-1B [6], LaVy [14], and V-Vistral. These models are trained on Vietnamese multimodal corpora and show reasonable performance in downstream tasks. However, they are often used for inference, not for automated data generation or labeling. Moreover, their limited parameter scale and resource constraints make them less competitive in complex reasoning tasks compared to frontier models like Gemini.

In this work, we integrate both general-domain MLLMs (Gemini 1.5 Pro [5], Gemini 2.0 Flash) and Vietnamese-specialized models (Vintern-1B) into our labeling pipeline. This hybrid approach aims to balance naturalness, accuracy, and cultural appropriateness in the generated QA pairs, while also mitigating potential bias and hallucinations often encountered in single-model setups.

III. METHODOLOGY

A. Automatic VQA Labeling Pipeline

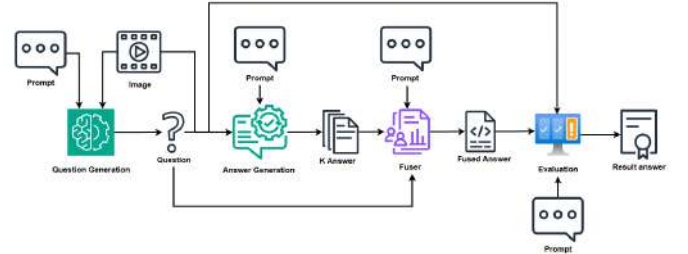


Fig. 1. Overall architecture of the automatic VQA labeling pipeline.

The proposed pipeline is composed of four main modules: (1) Question Generation, (2) Answer Generation, (3) Answer Fusion, and (4) Answer Evaluation. Each module leverages the generation capabilities of Multimodal Large Language Models (MLLMs) to automate the data labeling process from raw educational images.

1) **Question Generation:** Given an image input extracted from Vietnamese elementary school textbooks, the system initiates the labeling process by generating an age-appropriate, visually grounded question. We employ Gemini 1.5 Pro [5], a high-capacity MLLM developed by Google DeepMind, known for its strong reasoning capabilities, multilingual support, and long-context understanding.

To ensure pedagogical relevance, we adopt Structured Prompting, wherein prompts are designed with explicit constraints tailored to the Vietnamese educational context. For example, prompts include instructions such as: “Generate a short and clear question about the image, suitable for students aged 7–10”. This prompt design enforces simplicity, clarity, and alignment with primary-level curriculum standards. The output questions are designed to span a variety of types (e.g., object recognition, counting, spatial reasoning, inference) relevant to textbook learning objectives.

2) **Answer Generation:** For each generated question, we use a multi-source answer generation strategy to improve diversity and mitigate model bias. Specifically, two models are employed in parallel:

- **Gemini 2.0 Flash [3]:** Optimized for speed and cost-efficiency, this lightweight variant of Gemini delivers fast, fluent answers with strong multimodal understanding. It is particularly effective in generating concise, context-relevant answers for simple visual questions.
- **Vintern-1B [6]:** A Vietnamese-optimized MLLM pre-trained on over 3 million Vietnamese VQA samples, Vintern-1B incorporates a visual encoder (InternViT-300M), an MLP projector, and a Qwen-based decoder. It excels in generating culturally aligned, semantically accurate responses in Vietnamese, particularly for domain-specific or linguistically nuanced content.

This dual-model setting ensures that both general reasoning and Vietnamese-specific linguistic features are represented in the answer candidates.

3) **Answer Fusion:** To consolidate the outputs from the two answer generators, we introduce a dedicated LLMs Fuser module, which synthesizes a final answer by prompting Gemini 2.0 Flash. The fusion prompt incorporates Few-shot [11] examples and Chain-of-Thought (CoT) [15] reasoning to guide the model through multi-step inference and ensure that the resulting answer is more complete, semantically rich, and context-aware than any individual output.

This component is crucial for resolving minor discrepancies between answer candidates and for enriching short or underspecified responses with more natural and pedagogically appropriate language. Weights can be assigned to prioritize outputs from stronger models.

4) **Answer Evaluation:** Finally, the LLMs Evaluator module (using Gemini 1.5 Pro) is applied to assess the final answer based on four criteria: correctness, completeness, naturalness, and context relevance. If the answer does not meet a predefined threshold, it is discarded or flagged for manual inspection. This evaluation step helps reduce hallucination, grammatical errors, and irrelevant content in the final dataset.

B. Dataset Construction

To evaluate the effectiveness of the proposed labeling pipeline, we constructed ViEduVQA, a synthetic Vietnamese VQA dataset tailored for the elementary education domain. The dataset serves as both the primary training resource for

Vietnamese VQA models and a demonstration of the pipeline’s data generation capability.

1) **Image Collection:** We curated over 3,900 high-quality images from officially approved Vietnamese elementary school textbooks. Images were manually selected based on clarity, pedagogical value, and the potential to generate diverse question types. The collected images cover five thematic categories: *Education, Life, Nature, Object, and Others*.

2) **Automatic QA Generation:** Each image is processed through the proposed labeling pipeline to generate 3 to 5 QA pairs using MLLMs. The generation process incorporates structured and few-shot prompting to ensure age-appropriate, grammatically correct, and educationally relevant content. In total, the dataset contains over 18,000 QA pairs, each tagged with metadata such as question type (e.g., counting, location, inference), domain category, and the model used for generation.

3) **Benchmark Construction:** To assess data quality, we extracted a 250-sample benchmark subset from ViEduVQA. This subset was stratified across the five image domains and manually reviewed by educational experts. Each QA pair was corrected or confirmed to ensure accuracy, clarity, and alignment with primary school pedagogy. This benchmark serves as a ground-truth reference for evaluating VQA models trained on synthetic data.



Fig. 2. An Example for Vietnamese VQA in Education.

IV. EXPERIMENTS AND RESULT

A. Experimental Setup

To assess the effectiveness of the proposed automatic labeling pipeline and the quality of the generated ViEduVQA dataset, we conduct experiments by fine-tuning PaLI-Gemma [7], a multilingual encoder-decoder VLM, on the ViEduVQA dataset. To reduce cost, we adopt Parameter-Efficient Fine-Tuning (PEFT) [12] with 4-bit quantization and LoRA [13] adapters.

For evaluation, we use two benchmarks: (i) **ViEduVQA-Benchmark**, a 250-sample expert-verified subset, and (ii) **EVJVQA** [8], a public dataset to extract a subset of 8,320 Vietnamese QA pairs. These allow us to test both the generalization of models fine-tuned on ViEduVQA and the quality of our synthetic data.

To further assess our pipeline, we create two EVJVQA variants: **EVJVQA_1**, which keeps the original images and questions but replaces answers with synthetic ones, and **EVJVQA_2**, which generates both questions and answers

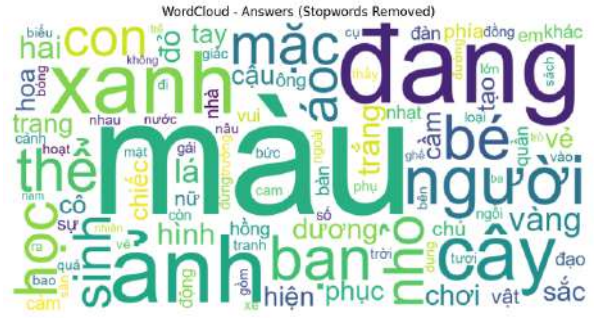


Fig. 3. ViEduVQA WordCloud.

from images. These enable us to evaluate the quality of answers and the pipeline’s ability to generate complete synthetic datasets. Table I summarizes the configurations.

TABLE I
OVERVIEW OF DATA VERSIONS FOR EVALUATION

Dataset	Image	Question	Answer
EVJVQA	Human	Human	Human
EVJVQA_1	Human	Human	Synthetic
EVJVQA_2	Human	Synthetic	Synthetic

We assess answer quality using both human and automatic metrics: **Accuracy** (expert judgment), **BLEU-4** (lexical overlap), **BERTScore** [9] (semantic similarity via multilingual embeddings), and **GeminiScore**: uses Gemini 1.5 Pro as an LLM judge [10] to assess correctness, completeness, relevance, and naturalness, while considering the image context. It complements traditional metrics with grounded, context-aware evaluation.

TABLE II
STATISTICAL COMPARISON ACROSS VQA DATASETS

Metric	ViEduVQA	EVJVQA	EVJVQA_1	EVJVQA_2
Images	3915	3763	3763	3763
QA Pairs	18838	8320	8320	18306
Q. Avg. Length	11.82	8.62	8.62	13.27
A. Avg. Length	26.15	7.23	8.86	6.64
Vocab Size	2847	1751	1925	3287

Table II shows that ViEduVQA is larger and linguistically richer than EVJVQA. EVJVQA_1, with regenerated answers, slightly increases length and vocabulary, while EVJVQA_2 yields the longest questions and largest vocabulary, reflecting higher diversity in fully synthetic data.

B. Results

TABLE III
PERFORMANCE OF EVJVQA_1 ON 500 EVJVQA SAMPLES

Accuracy	BLEU	BERT-P	BERT-R	BERT-F1
0.9040	0.4552	0.8822	0.8740	0.8780

Table III shows that EVJVQA_1 reaches 90.4% accuracy and high BERTScores, confirming strong alignment with human annotations.

Table IV summarizes the performance of fine-tuned PaLI-Gemma and baselines on two benchmarks. On the in-domain ViEduVQA-Benchmark, PaLI-Gemma, despite its poor **zero-shot** performance in Vietnamese answer generation, outperformed all baselines across all metrics after training, demonstrating the effectiveness of the proposed pipeline in producing high-quality, model-trainable data. On the out-of-domain EVJVQA dataset, PaLI-Gemma shows a slight performance drop but still achieves the highest Gemini Score and competitive results in Accuracy and BERT-F1. Although BLEU-4 is lower due to longer and more descriptive answers, the results confirm PaLI-Gemma’s ability to generalize and perform reliably even on unseen, human-annotated datasets.

C. Ablation Studies

To investigate the impact of the Fuser and Evaluator modules within the proposed pipeline, we conduct an ablation study by evaluating four configurations: (i) using neither Fuser nor Evaluator, (ii) using Fuser only, (iii) using Evaluator only, and (iv) using both modules. In the absence of the Fuser module, the answer selected is directly generated by the Gemini 2.0 Flash model without any aggregation or voting mechanism.

The results, presented in Table V, demonstrate that both modules contribute significantly to the quality of the labeled data:

- The Evaluator plays a critical role in improving Accuracy and BLEU-4 by filtering out low-quality or hallucinated answers that fail to match the image context or linguistic standards.
- The Fuser serves a complementary role by combining multiple answer candidates into a more complete and natural response, thereby improving semantic richness and coherence.
- The configuration with both modules enabled achieves the best overall performance across all three evaluation metrics, confirming that these components provide mutual benefits and are essential for high-quality data generation.

TABLE IV
PERFORMANCE OF VQA MODELS ON ViEduVQA-BENCHMARK (250 SAMPLES) AND EVJVQA (500 SAMPLES)

Dataset	Model	Accuracy ↑	BLEU-4 ↑	BERT-P ↑	BERT-R ↑	BERT-F1 ↑	Gemini Score ↑
ViEduVQA	PaLI-Gemma_ViEduVQA	0.6760	0.4052	0.7782	0.7864	0.7824	73.00/100
	Vintern-1B	0.6480	0.3774	0.7232	0.7158	0.7194	64.04/100
	V-Vistral 7B	0.6360	0.3507	0.7584	0.6992	0.7276	64.40/100
	LaVy	0.6520	0.3659	0.7421	0.7297	0.7358	69.72/100
EVJVQA	PaLI-Gemma_ViEduVQA	0.6380	0.3817	0.7534	0.7479	0.7506	70.09/100
	Vintern-1B	0.6380	0.3911	0.7715	0.7887	0.7800	65.10/100
	V-Vistral 7B	0.6220	0.3683	0.7397	0.7592	0.7492	69.83/100
	LaVy	0.6420	0.3898	0.7721	0.7538	0.7628	69.86/100

TABLE V
ABLATION STUDY RESULTS ON EVJVQA (500 SAMPLES)

Configuration	Accuracy ↑	BLEU-4 ↑	BERT-F1 ↑
No Fuser, No Evaluator	0.8760	0.3882	0.8361
No Fuser	0.8740	0.3759	0.8395
No Evaluator	0.8800	0.4011	0.8407
Full (Fuser + Evaluator)	0.9040	0.4552	0.8740

Prompting Ablation: Prompt design plays a critical role in steering the generative behavior of MLLMs for Vietnamese VQA. We compared three prompting strategies *Structured Prompting*, *Few-shot Prompting*, and *Chain-of-Thought (CoT) Prompting* on a 200-sample subset to analyze their impact on accuracy, completeness, and fluency.

- **Structured Prompting:** Provided explicit task instructions tailored to the educational domain. This approach yielded the highest factual correctness, as the model was constrained to focus on visually grounded details.
- **Chain-of-Thought Prompting:** Encouraged step-by-step reasoning and produced more complete and contextually rich answers. However, responses were sometimes longer than necessary.
- **Few-shot Prompting:** Improved fluency by providing high-quality in-context examples. While effective in producing natural and coherent answers, it occasionally led to verbose outputs or over-generalization.

Example Prompts:

- *Structured Prompt (Question Generation):*
“You are an expert in educational data labeling. Generate a short and clear question in English about the image, focusing only on visible objects, actions, or context. The question must be suitable for primary school students aged 7–10. If no relevant question can be asked, respond with ‘No valid question can be generated’.”
- *Few-shot Prompt (Answer Fusion):*
Q: What is the cat doing in the picture?
A1: The cat is sleeping.
A2: The cat is lying on the chair.
Task: Synthesize a single, natural English answer that is correct, concise, and age-appropriate.

- *CoT Prompt (Answer Evaluation):*

“Step 1: Analyze the question’s meaning. Step 2: Extract relevant visual details. Step 3: Compare the generated answer with the image context. Step 4: Decide if the answer is correct, complete, and natural. If incorrect or incomplete, provide a corrected version in Vietnamese.”

The comparative results demonstrate that no single prompting strategy is sufficient on its own. Instead, their combination ensures both factual correctness and linguistic naturalness. Structured prompts anchor the generation process, CoT enhances reasoning depth, and Few-shot improves fluency. This validates the integration of multiple prompting strategies into our pipeline, enabling the generation of high-quality Vietnamese VQA data.

D. Error Analysis

To better understand the limitations of the proposed labeling pipeline, we conducted a detailed error analysis on the ViEduVQA-Benchmark (250 manually verified samples). Table VI summarizes the distribution of common error types.

TABLE VI
ERROR STATISTICS OF THE ViEduVQA-BENCHMARK (250 SAMPLES)

Error Type	Description	Rate (%)
No Error	The question and answer are reasonable and grammatically correct	90.80
Factual Error	Answer contains incorrect visual information (e.g., miscounting, misidentification)	3.60
Incomplete Answer	Answer is too short or unclear to convey sufficient meaning	2.00
Over-Extended Answer	Answer includes redundant content not present in the image	1.60
Ambiguity	The question or answer is vague or ambiguous	0.80
Grammatical/Spelling Error	Incorrect word usage or ungrammatical Vietnamese sentence structure	0.80
No Answer	Model skipped the question or failed to generate a valid answer	0.40

1) *Error Breakdown*: As shown in Table VI, the vast majority of question–answer pairs (90.80%) were error-free, indicating that the proposed pipeline produces reliable outputs in most cases. However, approximately 9.2% of the samples contained some form of error. Among them, *factual errors* (3.60%) were the most frequent, typically caused by incorrect visual grounding such as miscounting objects or misidentifying items in the image. For example, in a question asking “How many apples are on the table?”, the system answered “5” instead of the correct “7”. *Incomplete answers* (2.00%) were also observed, where the system produced overly short or vague outputs (e.g., answering “Yes” to “What are the children doing?”). Conversely, *over-extended answers* (1.60%) added redundant details not present in the image, such as describing clothing colors that were not visible. Less frequent errors include *ambiguous expressions* (0.80%) and *grammatical or spelling mistakes* (0.80%), which slightly reduced naturalness and readability. In rare cases (0.40%), the model failed to produce any valid answer.

2) *Discussion*: Overall, factual grounding and counting remain the primary sources of errors, aligning with known limitations of current MLLMs. The error distribution suggests that integrating structured visual perception (e.g., object detection or segmentation) and refining the evaluation module could substantially reduce factual errors. Additionally, targeted prompt design and linguistic post-processing would help mitigate incomplete, verbose, or grammatically inconsistent responses.

V. CONCLUSION AND FUTURE WORKS

In this study, we proposed an automated labeling pipeline for Vietnamese Visual Question Answering (VQA) and constructed the ViEduVQA dataset using images from Vietnamese primary school textbooks. The pipeline leverages the generation and reasoning capabilities of Multimodal Large Language Models (MLLMs), including Gemini 1.5 Pro, Gemini 2.0 Flash, and Vintern-1B, to generate high-quality question–answer pairs that are semantically accurate and pedagogically appropriate. Our benchmark results show that the data generated by the pipeline is reliable enough to train VQA models effectively. The PaLI-Gemma model, when fine-tuned on ViEduVQA, achieves competitive performance across multiple evaluation metrics, demonstrating the quality and usefulness of the dataset.

For future work, we aim to extend the dataset and pipeline to other domains such as secondary education and healthcare, where labeled data remains scarce. We also plan to integrate the pipeline into interactive educational tools, enabling real-time VQA-based learning support for students. Additionally, we intend to enhance the evaluation and hallucination filtering mechanisms through dynamic prompting and multi-agent feedback to further improve the reliability and generalization of the labeling process.

Dataset Release

To facilitate future research, we will publicly release ViEduVQA under the **CC-BY-NC-4.0** license. The dataset will be accessible via Hugging Face and the official GitHub repository https://github.com/ShouyiLee/ViEduVQA-VietnameseVQA_DataGen, together with the benchmark and pipeline code. Since the dataset is constructed from images extracted from Vietnamese primary school textbooks, its use is strictly limited to non-commercial research purposes. Any commercial usage is not permitted under this release policy.

ACKNOWLEDGMENT

This research is funded by University of Information Technology-Vietnam National University Ho Chi Minh City under grant number D4-2025-05.

REFERENCES

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2425–2433.
- [2] D. A. Hudson and C. D. Manning, “GQA: A new dataset for real-world visual reasoning and compositional question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 6700–6709.
- [3] R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, *et al.*, “Gemini: A family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [4] OpenAI, “GPT-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [5] M. Reid *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arXiv preprint arXiv:2403.05530*, 2024.
- [6] K. T. Doan, B. G. Huynh, D. T. Hoang, T. D. Pham, N. H. Pham, Q. T. M. Nguyen, B. Q. Vo, and S. N. Hoang, “Vintern-1B: An efficient multimodal large language model for Vietnamese,” *arXiv preprint arXiv:2408.12480*, 2024.
- [7] L. Beyer *et al.*, “PaliGemma: A versatile 3B VLM for transfer,” *arXiv preprint arXiv:2407.07726*, 2024.
- [8] N. Luu-Thuy Nguyen, Nghia Hieu Nguyen, D. T.D. Vo, K. Q. Tran, and K. V. Nguyen, “EVJVQA challenge: Multilingual visual question answering,” *J. Comput. Sci. Cybern.*, vol. 39, no. 3, p. 237–259, Sep. 2023.
- [9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *Proc. 8th Int. Conf. Learn. Representations (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.
- [10] O. Mañas, B. Krojer, and A. Agrawal, “Improving automatic VQA evaluation using large language models,” in *Proc. 38th AAAI Conf. Artif. Intell. (AAAI)*, Vancouver, Canada, Feb. 2024, pp. 4171–4179.
- [11] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020.
- [12] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, “Parameter-efficient fine-tuning for large models: A comprehensive survey,” *Trans. Mach. Learn. Res.*, vol. 2024, 2024.
- [13] E. J. Hu *et al.*, “LoRA: Low-rank adaptation of large language models,” in *Proc. 10th Int. Conf. Learn. Representations (ICLR)*, Virtual Event, Apr. 2022.
- [14] C. Tran and H. L. Thanh, “LaVy: Vietnamese multimodal large language model,” *CoRR*, vol. abs/2404.07922, 2024.
- [15] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems*, vol. 35, New Orleans, LA, USA, Nov.–Dec. 2022.

Sketch2Comic: A Diffusion-Based Framework for Consistent Multi-Character Comic Synthesis

Nguyen Hoang Lam[†], Nguyen Trong Tuan[†], Vuong Dinh Minh[†],
Nguyen Vo Ngoc Khue, Nguyen Quang Tri, Dinh Viet Sang^{*}
School of Information and Communication Technology
Hanoi University of Science and Technology, Hanoi, Vietnam

Abstract—*Sketch2Comic* rethinks how structural and subject cues are integrated into large diffusion transformers for comic creation. Existing text- or sketch-guided models often struggle to preserve character identity across comic panels and require separate networks for each control task, limiting their flexibility. To address these challenges, we propose a unified three-stage pipeline built upon a 12-billion-parameter Flux Diffusion Transformer. In the first stage, a ControlNet-based module transforms a hand-drawn sketch and accompanying caption into a colored single-character illustration. The second stage introduces OminiControl Comic LoRA, a lightweight low-rank adapter trained on our Comic Character Pairs dataset of 3,000 identity-consistent sketch-image pairs, which injects reference-image tokens into cross-attention to enable one-shot, identity-preserving pose transfer. In the third stage, compact trigger-token LoRAs are learned per character, and multiple adapters can be merged at inference to generate consistent multi-character scenes, optionally guided by a layout sketch. Evaluated on hundreds of unseen pose prompts, Sketch2Comic outperforms Stable Diffusion XL with IP-Adapter and SSR-Encoder on various scores, demonstrating that character-consistent, multi-character comics can be produced end-to-end from casual sketches without architectural overhead.

Index Terms—Character consistency, Comic synthesis, Diffusion models, Image generation, Low-Rank Adaptation (LoRA)

I. INTRODUCTION

Generative AI has made text-to-image synthesis ubiquitous, with models such as DALL-E2 [1], Stable Diffusion [2] and Midjourney [3]. While these latent diffusion systems excel at generating single images from natural language prompts, creating sequential art, such as comics, introduces distinct requirements. Comics rely not only on visual quality but also on narrative coherence and stylistic consistency across multiple panels. Each panel must faithfully preserve characters’ appearances, poses, and spatial relationships in service of the story.

When applied to this domain, general-purpose diffusion models reveal two persistent limitations. First,

while sketch-conditioned methods such as ControlNet [4] can enforce geometric structure, they offer no mechanism to preserve subject identity. Second, subject-personalisation methods such as DreamBooth [5] or Textual Inversion [6] require model fine-tuning and struggle to generalise across poses, lighting conditions, and especially in multi-subject settings. Prior methods cannot simultaneously guarantee sketch fidelity, identity preservation, and visual coherence across character interactions, which are the requirements central to comic generation.

To address these challenges, we propose a lightweight, semi-automatic workflow that integrates multiple conditioning strategies into a unified pipeline, built upon the FLUX diffusion transformer [7]. A ControlNet branch ensures that each panel conforms to the user’s sketch. To preserve subject identity, we introduce OminiControl Comic LoRA, a low-rank adapter that enables one-shot identity transfer from a single reference image. This module is trained on a novel dataset, namely **Comic Character Pairs**, consisting of 3,000 identity-consistent sketch-image pairs, curated and filtered using DINO-v2 [8]. To support multi-character scenes, we adopt a modular strategy where each character is represented by a separate LoRA, activated via unique trigger tokens. This approach enables coherent multi-character composition without retraining or architectural changes. Our experiments demonstrate that the proposed system outperforms strong baselines in CLIP-based similarity metrics. Our contributions are threefold:

- 1) **Comic Character Pairs**, a publicly released dataset for identity-preserving comic research;
- 2) **OminiControl Comic LoRA**, a 0.1%-parameter adapter enabling efficient one-shot identity transfer in diffusion transformers;
- 3) **An end-to-end sketch-to-comic workflow** that unifies geometric control, subject conditioning and multi-character consistency.

The remainder of this article is organised as follows: Section II surveys related work; Section III details the dataset, architecture and training strategy; Section IV

[†] These authors contributed equally to this work.

^{*} Corresponding author.

presents experimental results; Section V concludes and discusses future extensions.

II. RELATED WORK

This section reviews recent literature on sketch-based conditioning, identity-preserving diffusion models, and modular control strategies relevant to multi-character comic generation.

A. Diffusion models

Diffusion probabilistic models have become the backbone for modern image synthesis, surpassing GAN-based translators [9] in both fidelity and mode coverage. Early pixel-space DDPMs [10] generate convincing pictures but require large memory and long chains of noise steps. Latent diffusion [2] mitigates this by performing denoising in a VAE-compressed space, reducing computational cost while retaining visual quality. Recent work adds transformer capacity to the diffusion core. Diffusion Transformer [11, 12, 13] views the latent feature map as a sequence, allowing deep self-attention layers to capture global dependencies and scale generation to megapixel resolutions. FLUX [7] extends DiT with a flow-matching objective [14] that aligns the forward and reverse processes, yielding faster convergence and state-of-the-art quality on large text-to-image benchmarks. Nevertheless, these models still lack fine-grained controllability, motivating the conditioning techniques reviewed in the next subsection.

B. Image conditioning

Conditioning mechanisms inject external signals into pretrained backbones without retraining the full model. ControlNet [4] attaches a zero-initialised branch that reuses encoder features to impose edges, depth, or pose maps during the early diffusion steps. Lighter variants, i.e. T2I-Adapter and MCM, insert small modulating layers that accept sketches or segmentation masks and keep base weights frozen [15, 16]. Subject-driven methods fine-tune a few million parameters (DreamBooth, Textual Inversion) or add low-rank updates (LoRA) to bind a character embedding to the model [17, 5, 6]. IP-Adapter improves efficiency by injecting a reference-image token through cross-attention, while AnyDoor fuses discrete condition tokens on-the-fly to relocate objects without retraining [18, 19]. UniControl mixes layout, sketch, and identity signals with a Mixture-of-Experts adapter, showing that heterogeneous conditions can coexist in one network [20]. However, reliable composition of multiple personalised adapters remains open: current pipelines struggle when more than two LoRA modules are merged, and identity drift appears in crowded scenes. FLUX’s transformer backbone is well suited to modular adapters because new LoRA blocks can be interleaved between attention layers with negligible speed cost [7]. Building on this insight, *OminiControl* [21] introduces a single adapter that supports both

spatial (sketch/edge) and subject tokens, but it has not been explored for comics where long-term panel-to-panel consistency and multi-character layouts are required. Our **Sketch2Comic** framework closes this gap by combining sketch-level ControlNet guidance, character-specific LoRA modules triggered by prompt tokens, and an OminiControl-style adapter, enabling pose-accurate and identity-consistent generation for several characters across sequential comic panels.

III. METHODOLOGY

Our proposed method consists of three main stages, illustrated in Fig. 1a. First, each character sketch is converted into a colored comic-style image using a diffusion model with ControlNet (Stage 1). Next, we generate additional poses for each character using an OminiControl diffusion model to augment the character’s images (Stage 2). Finally, we train a LoRA for each character, and use the LoRAs to compose multi-character scenes (Stage 3). This section describes each stage in detail, along with the data preparation and model architectures involved.

A. Stage 1: Sketch-to-Image generation with Flux and ControlNet

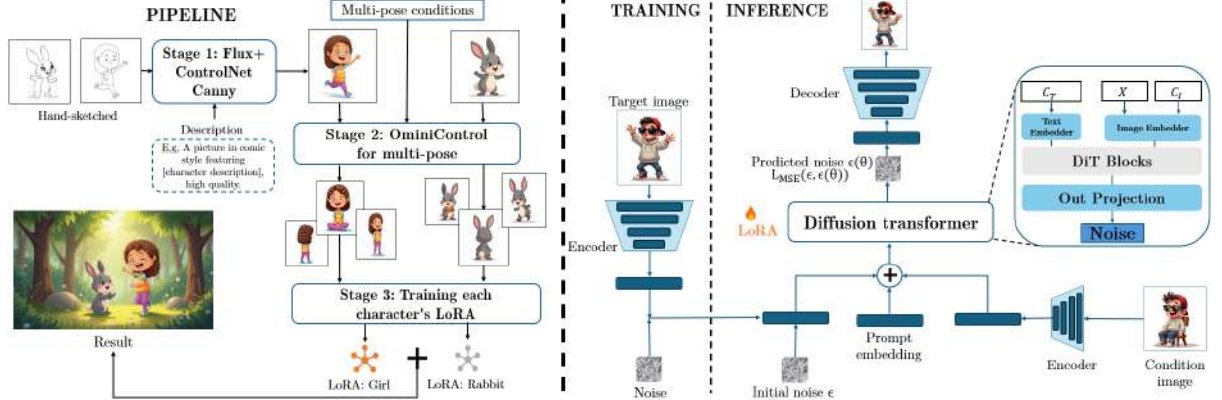
The first stage converts a single hand-drawn sketch into a full-colour comic-style illustration while strictly respecting the original geometry. As shown in Fig. 1, the pipeline combines a Flux diffusion backbone with a Canny-edge ControlNet branch. The sketch is passed directly to ControlNet and treated as an edge map that conditions the denoising trajectory, whereas a text prompt supplies semantic attributes such as colour, material, and expression:

“A picture in comic style featuring [character description], high quality.”

Here, [character description] specifies hair colour, clothing, action, and facial expression. During sampling, ControlNet is applied with `weight = 0.7` from `start_step = 0` to `end_step = 0.7`, i.e. the first 70 % of the diffusion chain. This scheduling anchors the global layout early when structural tokens dominate, while leaving later steps free to refine texture and lighting under prompt guidance. The model therefore operates under dual conditioning: the edge map enforces spatial fidelity, and the prompt conveys high-level style. The resulting images supply high-quality data for subsequent pose augmentation and character-specific LoRA training.

B. Stage 2: Multi-pose training with OminiControl Comic LoRA

After obtaining complete character images from hand-drawn sketches, the next step is to enhance the expressiveness of characters across various poses, viewpoints, and facial expressions. This is achieved by training the OminiControl Comic LoRA model, which enables image generation based on a single reference image combined with a desired pose.



(a) Overview of the process of creating illustrations from sketches. (b) Workflow of training the OminiControl Comic LoRA model (Stage 2).

Figure 1: The overall workflow for comic illustration generation and model training

1) *Comic character pairs dataset construction:* To train the OminiControl Comic LoRA model, a specialized dataset called Comic Character Pairs was manually constructed. Each pair in this dataset consists of images of the same character rendered in two different poses or viewpoints, preserving visual consistency in character identity while introducing diversity in pose and context. We define 40 common cartoon character types (e.g., man, woman, girl, rabbit, panda, shark, etc.). For each type, we generate 80 diverse character descriptions covering visual traits such as clothing, hairstyle, and color. Then, for each description, two distinct pose prompts (`condition[1]` and `condition[2]`) are created. These triplets serve as paired prompts for generating the same character in two different poses. Using Flux.dev [7], we adopt an IC-LoRA [22] inspired prompt structure to synthesize 1536×768 resolution images containing two halves, each representing a different pose. To minimize pose blending issues known as cross-conditioning, we apply regional prompting [23]. Each image half is conditioned on its own prompt via an attention mask M :

$$M_{ij} = \begin{cases} 0, & \text{if tokens } i, j \text{ in same region} \\ -\infty, & \text{if tokens } i, j \text{ in different regions} \end{cases} \quad (1)$$

This attention mask is incorporated into the dot-product attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} + M \right) V, \quad (2)$$

To balance coherence, we combine both base and regional attention. Here, $\gamma(\cdot)$ denotes a normalization function applied to the attention outputs before weighted summation:

$$z = \frac{1}{3} \cdot \gamma(\text{Attention}_{\text{base}}) + \frac{2}{3} \cdot \gamma(\text{Attention}_{\text{regional}}) \quad (3)$$

This hybrid scheme is applied in the first 70% denoising steps to build consistent yet expressive character frames. Generated dual-pose images are split into single-pose images using Grounding DINO [24] and SAM [25]. Masks are used to extract each character, bounding boxes are expanded and squared, then resized to 512×512 . To ensure consistency, DINO-Score [8] is computed for each pair, filtering out pairs with scores below 0.75. The final dataset approximately includes 3,000 character pairs (6,000 individual images).

2) *OminiControl Comic LoRA architecture:* OminiControl extends the DiT framework by incorporating image condition tokens C_I alongside image tokens X and text condition tokens C_T :

$$Z = [X; C_T; C_I] \quad (4)$$

This allows multi-modal attention to be applied, enabling control over both identity (via C_I) and pose/context (via C_T), illustrated in Fig. 1b. Given a clean image x_0 , noise is added to create x_t according to timestep t sampled uniformly. The model then predicts the noise based on x_t , timestep t , and conditions $c = C_I, C_T$. The MSE loss between predicted and real noise is minimized:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \epsilon \sim \mathcal{N}(0, 1), c} \left[\|\epsilon - \hat{\epsilon}_\theta(x_t, t, c)\|_2^2 \right] \quad (5)$$

LoRA is used for parameter-efficient fine-tuning. After training, the model is used to generate different images for each character by combining the reference image with varied pose prompts. This enables the production of visually consistent yet pose-diverse character sets with minimal manual effort.

C. *Stage 3: Per-Character LoRA training with trigger words*

For each character in the system, we train a dedicated LoRA model to learn independent identity traits

such as hairstyle, clothing, color palette, and facial expression across various poses and contexts. Each LoRA model is fine-tuned using the image set generated in Stage 2, where all images in a set depict the same character in different situations. To support character identification during image generation, we introduce the trigger word technique. Each character is assigned a unique token added to the model’s tokenizer. This trigger word is appended to all prompts corresponding to the character and serves as a linguistic identifier in the embedding space. It helps the model associate the token with the visual appearance of that character, regardless of pose, background, or action. Training prompts are rebuilt by modifying the original captions from Stage 2 to include the corresponding trigger word. For example, given the caption:

"A picture in comic style featuring a young girl with curly brown hair, wearing a striped orange shirt and purple pants, running happily with a big smile, high quality, no blur."

The version with the trigger word would be:

*"A picture in comic style featuring **OHW** girl, running happily with a big smile, high quality, no blur."*

Embedding the trigger word in a natural context allows the model to associate the token with that character’s specific visual features. Once trained, a character-specific LoRA model can be loaded at runtime and activated via a prompt containing the trigger word. This allows rendering that character in any scene or pose simply by prompting with the correct token. A key advantage of this method is the flexible loading/unloading of LoRA modules during inference. Depending on the content, the system may load one or multiple LoRA models corresponding to different characters. Combined with prompts that include their respective trigger words, this enables the generation of multi-character scenes, where each character is portrayed with high fidelity to their trained identity. This modularity makes it easy to expand and reuse character models, which is particularly beneficial for multi-character comic storytelling where consistent identity across panels is critical, yet variety in action and interaction must be preserved.

IV. EXPERIMENTAL RESULTS

We conduct a series of experiments to evaluate our system, focusing on character consistency and faithfulness to inputs.

A. Experimental Setup

Evaluation Metrics. We evaluate our models using three vision–language metrics. **CLIP Text Score** [26] measures how well the image matches the input prompt; **CLIP Image Score** evaluates identity similarity between the generated and reference images using CLIP embeddings; and **DINO Score** [8] assesses structural consistency using DINOv2 features, offering

robustness to style variations. All reported scores are averaged over 5 runs; variance was found to be within 1–2%

Baselines. We compare our full pipeline with three baselines: **SSR-Encoder** [27], **SDXL + IP-Adapter** [28], and **Flux.dev + IP-Adapter** [7], all using the same prompts and inference settings. These models serve as state-of-the-art subject-driven generation methods without character-specific LoRA training or regional prompting.

Regional Prompting Setup. To assess regional prompting [23], we generated 500 multi-pose character prompts (3 seeds each) using Flux.dev with 25 denoising steps, applying regional guidance in the first 7 steps. This setup helps disentangle pose and identity.

OminiControl Comic Training. OminiControl Comic LoRA is trained on 3,000 character pairs using Flux.dev as the base model. We use CLIP-L and T5-XXL for text encoding, LoRA rank 4, learning rate **1e-4**, batch size **1** (with gradient accumulation of 4), AdamW optimizer, weight decay **0.01**, and 3 training epochs on a single A100 (80GB) GPU.

Data Scale Ablation. To evaluate the effect of data scale, we compare models trained on the **full dataset** (3,000 pairs), a **50% subset**, and a **target-only** setup where the same image is used as both input and target.

Character-Specific LoRA. For character-specific personalization, we fine-tune LoRA modules from OminiControl Comic using 15 pose-varied images per character. Each module is trained for **900 steps** with learning rate **7e-4**, LoRA rank **16**, and batch size **1** (accum. 4), injecting weights into attention and linear layers of the diffusion backbone. A unique trigger word is used per character to ensure identity preservation during inference. We also test composability by combining two LoRA modules in one pipeline. For 10 character pairs, each evaluated on 20 prompts with 5 seeds.

B. Performance evaluation

Table I: Comparison of evaluation metrics with and without regional prompting.

	With RP	Without RP
CLIP Text Score	0.33	0.29
CLIP Image Score	0.91	0.93
DINO Score	0.83	0.85

Effect of Regional Prompting [23] in dataset generation. Table I shows that regional prompting significantly improves text-image alignment, as reflected in higher CLIP Text Scores. Although CLIP Image and DINO Scores slightly decrease, this indicates more diverse poses and viewpoints rather than degraded identity fidelity. Thus, regional prompting effectively balances pose diversity and character consistency, enhancing training data quality for pose-conditioned generation.



Figure 2: Qualitative comparison. Left: Spatially aligned generation across multiple baselines. Right: Subject-driven generation using character-specific LoRA models. Our method consistently achieves better controllability and visual quality.

Table II: Comparison of performance across different methods.

Method	DINO $\uparrow$	CLIP-Image $\uparrow$	CLIP-Text $\uparrow$
SSR-Encoder	0.768	0.863	0.267
SDXL IP-Adapter	0.755	0.865	0.312
Flux.dev IP-Adapter	0.732	0.842	0.309
OminiControl Base	0.798	0.907	0.315
OminiControl Comic	0.814	0.921	0.322

Training and evaluation of OminiControl Comic LoRA. As presented in Table II, OminiControl Comic achieves the highest scores across all metrics. These results stem from two factors: (1) the stronger base model (Flux vs. SDXL/SD1.5), which enables better prompt understanding, and (2) the specialized training dataset with controlled comic character variation. This is strongly reinforced by the illustration in Figure 2a, confirming the model’s superiority in multi-pose character generation and its relevance for digital comic creation tasks.

Table III: Evaluation metrics across dataset configurations.

Configuration	DINO $\uparrow$	CLIP-Image $\uparrow$	CLIP-Text $\uparrow$
Full (100%)	0.814	0.921	0.322
Partial (50%)	0.772	0.843	0.318
Target-only	0.940	0.984	0.204

Impact of the Comic Character Pairs dataset. Results in Table III show that while the Target-only setup achieves the lowest training loss and highest visual similarity (DINO and CLIP-Image), it performs poorly on CLIP-Text, indicating overfitting and loss of text alignment. In contrast, using the full dataset balances visual fidelity and prompt adherence, confirming its value for training pose-controllable models with consistent character identity.

Training character-specific LoRA modules. Results shown in Table IV show LoRA-generated images

Table IV: Performance after LoRA training per character

Method	DINO $\uparrow$	CLIP-I $\uparrow$	CLIP-T $\uparrow$
Char. Pairs	0.83	0.91	0.33
LoRA-generated	0.78	0.87	0.34
+ White BG	0.81	0.91	0.26

achieve higher CLIP-Text scores, indicating better prompt alignment. However, lower DINO and CLIP-Image scores suggest background variation affects visual similarity. When backgrounds are removed, scores improve, confirming character features are preserved.

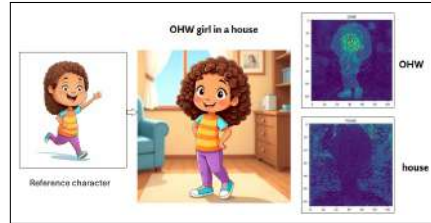


Figure 3: Attention map visualization of the trigger word’s influence on the generated images.

To validate trigger word influence, we compute its attention map by averaging cross-attention weights across layers and heads during generation. The resulting Figure 3 highlights spatial focus on the keyword, confirming its role in character control.

Table V: Prompt alignment in dual-character scenarios.

Method	CLIP Score $\uparrow$
LoRA (2-char)	0.33
Flux.dev (no-LoRA)	0.34

We test the composability of our approach by combining two independently trained LoRA modules within a single pipeline, as illustrated in Figure 2b. The LoRA parameters are merged through weighted interpolation, with the mixing coefficient α tuned in the range 0.3–0.7 depending on dataset balance. We

then compare the average CLIP score against the base Flux model (without LoRA or trigger words). Results in Table V show that the LoRA-based method performs comparably to the baseline, indicating that two character-specific modules can be composed to generate coherent images while preserving each identity.

V. CONCLUSION

This paper presents a pipeline for generating consistent multi-frame comic scenes from hand-drawn character inputs. By leveraging regional prompting, we construct a high-diversity dataset Comic Character Pairs to train a conditional generation model named OminiControl Comic LoRA. The approach enables flexible pose synthesis from a single reference image while preserving character identity. Individual LoRA modules are then fine-tuned for each character, allowing multiple characters to be composed in a single frame with visual consistency. Experimental results show that the system can maintain key identity traits, including color, style, and expression, across varied prompts and compositions. Currently, we rely primarily on CLIP-based scores, which do not fully capture multi-character identity preservation or cross-frame consistency. Moreover, limitations arise when more than two character-specific LoRA modules are combined in a single inference, often causing feature interference or visual distortion. This highlights the current challenge of multi-LoRA integration in diffusion models. Future directions include optimizing LoRA composition strategies (e.g., using lower-rank modules or inpainting-based synthesis), incorporating identity-specific metrics and extending the system to support automatic dialogue insertion for complete comic panels.

REFERENCES

- [1] Aditya Ramesh et al. “Hierarchical Text-Conditional Image Generation with CLIP Latents”. In: *arXiv preprint arXiv:2204.06125* (2022).
- [2] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684–10695.
- [3] Midjourney, Inc. *Midjourney*. 2022. URL: <https://www.midjourney.com>.
- [4] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Adding Conditional Control to Text-to-Image Diffusion Models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 3636–3647.
- [5] Nataniel Ruiz et al. “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation”. In: *Proceedings of the IEEE/CVPR*. 2023, pp. 22500–22510.
- [6] Rinon Gal et al. “An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion”. In: *arXiv preprint arXiv:2208.01618* (2022).
- [7] Black Forest Labs. “FLUX: A Pre-Trained Large-Language-and-Vision-Model for talking about and seeing the world”. In: *arXiv preprint arXiv:2405.20454* (2024).
- [8] Maxime Oquab et al. “DINOv2: Learning Robust Visual Features without Supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).
- [9] Ian J. Goodfellow et al. “Generative Adversarial Networks”. In: *arXiv preprint arXiv:1406.2661* (2014).
- [10] Alex Nichol and Pratul Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. 2021. arXiv: [2102.09672](https://arxiv.org/abs/2102.09672) [cs.LG].
- [11] William Peebles and Saining Xie. *Scalable Diffusion Models with Transformers*. 2023. arXiv: [2212.09748](https://arxiv.org/abs/2212.09748) [cs.CV].
- [12] Junsong Chen et al. *PixArt- α : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis*. 2023. arXiv: [2310.00426](https://arxiv.org/abs/2310.00426) [cs.CV].
- [13] Junsong Chen et al. *PIXART- δ : Fast and Controllable Image Generation with Latent Consistency Models*. 2024. arXiv: [2401.05252](https://arxiv.org/abs/2401.05252) [cs.CV].
- [14] Yaron Lipman et al. *Flow Matching for Generative Modeling*. 2023. arXiv: [2210.02747](https://arxiv.org/abs/2210.02747) [cs.LG].
- [15] Cusuh Ham et al. “Modulating Pretrained Diffusion Models for Multimodal Image Synthesis”. In: *ACM SIGGRAPH 2023 Conference Proceedings* (2023).
- [16] Chong Mou et al. “T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models”. In: *AAAI’24/IAAI’24/EAAI’24*. AAAI Press, 2024.
- [17] Edward J. Hu et al. “LoRA: Low-rank adaptation of large language models”. In: *International Conference on Learning Representations (ICLR)*. 2022.
- [18] Hu Ye et al. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. 2023. arXiv: [2308.06721](https://arxiv.org/abs/2308.06721) [cs.CV].
- [19] Xi Chen et al. *AnyDoor: Zero-shot Object-level Image Customization*. 2024. arXiv: [2307.09481](https://arxiv.org/abs/2307.09481) [cs.CV].
- [20] Can Qin et al. *UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild*. 2023. arXiv: [2305.11147](https://arxiv.org/abs/2305.11147) [cs.CV].
- [21] Zhenxiong Tan et al. “Ominicontrol: Minimal and universal control for diffusion transformer”. In: *arXiv preprint arXiv:2411.15098* (2024).
- [22] Lianghua Huang et al. *In-Context LoRA for Diffusion Transformers*. 2024. arXiv: [2410.23775](https://arxiv.org/abs/2410.23775) [cs.CV].
- [23] Anthony Chen et al. “Training-free regional prompting for diffusion transformers”. In: *arXiv preprint arXiv:2411.02395* (2024).
- [24] Shilong Liu et al. “Grounding dino: Marrying dino with grounded pre-training for open-set object detection”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 38–55.
- [25] Alexander Kirillov et al. “Segment anything”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 4015–4026.
- [26] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.
- [27] Yuxuan Zhang et al. “SSR-Encoder: Encoding selective subject representation for subject-driven generation”. In: *Proceedings of the IEEE/CVPR*. 2024, pp. 8069–8078.
- [28] Hu Ye et al. “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models”. In: *arXiv preprint arXiv:2308.06721* (2023).

3D Vietnamese Woodblocks Repair with Normal-Enhanced Fused Shape Representation

1st Viet Nam Le
Fac. of Information Technology
VNU University of Engineering
and Technology
Hanoi, Vietnam
levietnam@hdu.edu.vn

2nd Thanh Tung Hoang
Fac. of Information Technology
VNU University of Engineering
and Technology
Hanoi, Vietnam
21021661@vnu.edu.vn

3rd Van Canh Vu
Fac. of Information Technology
VNU University of Engineering
and Technology
Hanoi, Vietnam
23025054@vnu.edu.vn

4th Cong Thuong Le
Fac. of Information Technology
VNU University of Engineering
and Technology
Hanoi, Vietnam
thuonglc@vnu.edu.vn

5th Thi Duyen Ngo
Fac. of Information Technology
VNU University of Engineering
and Technology
Hanoi, Vietnam
duyennnt@vnu.edu.vn

6th Thanh Ha Le*
Fac. of Information Technology
VNU University of Engineering
and Technology
Hanoi, Vietnam
ltha@vnu.edu.vn

Abstract—Vietnamese woodblocks, recognized by UNESCO as a World Documentary Heritage, are carved with Han-Nom characters for mass printing during the feudal period. Over time, many woodblocks have deteriorated, losing fine surface details. We propose a learning-based 3D surface restoration method using a fused shape representation with normal vector supervision to safely repair damaged 3D regions. Experimental results on the ShapeNet dataset show a 3.8% improvement in Normal Consistency (NC) and a slight reduction in Non-Fractured Region Error (NFRE) compared to the DeepJoin baseline. On the Nguyen Dynasty 3D Woodblock dataset, our method achieves a 25.2% NC increase and lower NFRE, effectively reconstructing fine carvings and sharp relief edges. Qualitative evaluation demonstrates that the restored meshes more accurately preserve geometric details compared to baseline methods.

Index Terms—3D reconstruction, surface restoration, implicit functions, cultural heritage preservation, normal vector supervision.

I. INTRODUCTION

Vietnamese woodblocks, most notably those of the Nguyen Dynasty, which include 34,619 engraved plates made in the 19th and early 20th centuries, are recognized by UNESCO as Vietnam’s first legacy of the World Memory [1]. These woodblocks, inscribed with Han-Nom characters, were utilized for the printing of official historical records and legal documents, fulfilling both archival and public dissemination functions.

Due to weather factors, biological agents, and the possibility of fire or conflict, many woodblocks have experienced significant physical deterioration over time. Despite the fact that more than 50,000 blocks were initially manufactured, only a proportion of them remain in their original form today. Many are now fragmented or damaged, posing substantial challenges for preservation and restoration.

*Corresponding author is Thanh Ha Le



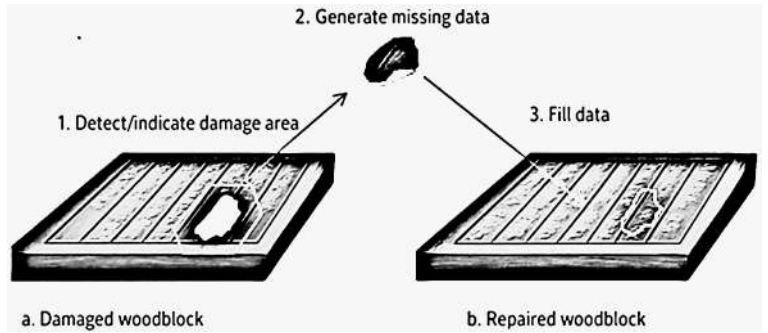
Fig. 1: Printing the cover page of “Dai Viet Su Ky Toan Thu” (Complete Annals of Dai Viet) from a Nguyen Dynasty woodblock (source: <https://mocban.vn/en/document/complete-annals-of-dai-viet/>).

Physical restoration by archivists and artisans remains essential for preserving cultural artifacts. Alongside this, digital preservation and restoration techniques are rapidly developing, enabling the recovery of lost geometry while maintaining the original characteristics. Figure 2(a) shows the typical physical condition of damaged Nguyen Dynasty woodblocks, highlighting cracks, missing fragments, and surface erosion. In parallel, digital methods, leveraging recent advances in 3D imaging and machine learning, offer a noninvasive alternative that does not interfere with the original objects. Figure 2(b) illustrates a general digital restoration workflow, where 3D scans of damaged woodblocks are processed to extract geometric features, predict implicit fields, and reconstruct surfaces.

3D reconstruction of severely damaged cultural artifacts presents significant challenges due to complex geometry and



(a) Physically damaged Nguyen Dynasty woodblock (source: disanso.vn)



(b) Digital repairation workflow.

Fig. 2: Illustration of physically damaged Nguyen Dynasty woodblocks and the basic conceptual workflow for 3D repairation.

limited training data. These issues are particularly evident in historic woodblocks, where fine details are often lost or degraded. This study focuses on Nguyen Dynasty woodblocks, a collection of exceptional historical value and the most widely digitized woodblock archive in Vietnam, due to the availability of 3D scanned data and their national significance.

To address these challenges, we propose a deep learning method for restoring 3d mesh from degraded woodblock scans, incorporating additional geometric cues as learnable signals to enhance reconstruction in damaged regions.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed methodology. Section IV outlines the experimental setup, and Section V presents the results and discussion. Finally, section VI concludes the paper and suggests future directions.

II. RELATED WORKS

The restoration of 3D models has been extensively studied in computer vision and graphics. Existing techniques can be broadly categorized into traditional geometry-based approaches and deep learning-based methods.

A. Traditional 3D Reconstruction Techniques

Classical approaches such as Structure from Motion [2] reconstruct 3D geometry from multiple 2D views by detecting and matching keypoints (e.g., SIFT [3]), followed by point cloud generation. Other methods, including volumetric, point-based, and implicit surface reconstructions [4], [5], rely on depth information or multi-view inputs to recover 3D structure.

However, these approaches face limitations when only a single color image is available, as is often the case with historical woodblock prints [6]. Without reliable depth cues, traditional geometry-based methods are often inapplicable or yield low-fidelity reconstructions.

B. Deep Learning Approaches for 3D Reconstruction

Recent advances in deep learning have enabled the inference of 3D shapes from limited 2D inputs. Convolutional Neural Networks (CNNs) can extract spatial features, while autoencoders and variational autoencoders (VAEs) [7] have been used to learn latent 3D representations. Generative Adversarial

Networks (GANs) [8] have further improved realism in 3D generation, with enhancements in training stability and data efficiency [9], [10].

For 3D point clouds, PointNet and PointNet++ [11], [12] are foundational models capable of learning from unordered point sets, making them suitable for complex geometric inference. DeepSDF [13] and hierarchical implicit models [14] are other influential techniques for shape representation learning.

C. Mesh Repair and Inpainting

The task of completing missing regions in images and 3D models is addressed by inpainting techniques. VAE-based models [7] learn latent shape spaces, while GAN-based inpainting [15]–[18] improves realism and coherence.

In the context of 3D restoration, DeepMend [19] uses occupancy-based fracture repair, while MendNet [20] predicts restorations from partial point clouds. Pix2Repair [21] introduces the “Fantastic Breaks” [22] dataset to support 3D repair from a single image.

For the restoration of 3D representations of Vietnamese historical woodblocks, gated convolution methods have been applied to depth map inpainting [23], through the adaptation of 2D image inpainting models to operate on 3D depth data.

III. METHOD

A. Overview

This work investigates the reconstruction of damaged 3D woodblock surfaces using neural implicit representations enhanced with surface normals. The approach jointly models occupancy, signed distance, and normal fields to refine incomplete inputs and improve surface fidelity across varying degradation and complexity levels. The framework, illustrated in Figure 3, extracts features from a raw 3D scan and its normals, fuses them into an implicit surface representation, and trains the model with multiple loss functions to produce a detailed 3D surface output. Within the latent fusion stage, two networks are used: f_C learns from complete shapes to provide global structural guidance, while f_B learns from damaged shapes to focus the reconstruction on missing regions. Their outputs are combined before being passed to the implicit decoder.

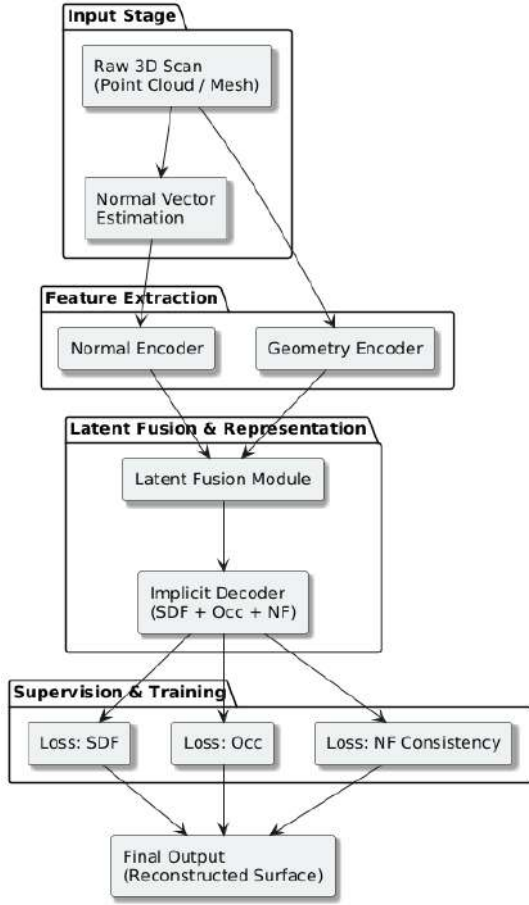


Fig. 3: Normal-Enhanced Deep Learning Framework for 3D Woodblock Surface Reconstruction.

B. Network Architecture

Our model extends the baseline DeepJoin [24], which itself is inspired by DeepSDF [13]. It follows a dual branch auto-decoder design, where the branches process complete (C for Complete) and damaged (B for Break) shapes separately. Specifically, we define two implicit networks:

- $f_C(\mathbf{x}, \mathbf{z}_C)$ models complete shapes, providing global shape guidance.
- $f_B(\mathbf{x}, \mathbf{z}_B)$ handles the damaged input, marking missing regions for repair.

Each network takes a 3D point $\mathbf{x} \in \mathbb{R}^3$ and a learnable latent vector $\mathbf{z} \in \mathbb{R}^d$ representing global shape information:

$$f(\mathbf{x}, \mathbf{z}) = o(\mathbf{x}), s(\mathbf{x}), \mathbf{n}(\mathbf{x}). \quad (1)$$

Here, $o(\mathbf{x})$ predicts inside/outside occupancy, $s(\mathbf{x})$ gives the signed distance to the surface, and $\mathbf{n}(\mathbf{x})$ is the local normal. This extends the implicit function formulation with normal supervision, improving fine-detail preservation.

The network is trained to minimize reconstruction losses on all outputs. Latent codes $\mathbf{z}_C$ and $\mathbf{z}_B$ are optimized during training, and a restoration code $\mathbf{z}_R$ is interpolated from them to guide surface restoration.

a) *Normal-aware Feature Encoding.*: Surface normals are added as input channels and processed with a residual block, enabling the network to learn local surface variations independently from global shape features.

C. Surface Normal Integration

Normal vectors are extracted using face, vertex, or PCA-based methods to provide robust local shape cues. They are incorporated in two ways:

- *Input Fusion*: Normals estimated from partial surfaces are concatenated with point coordinates and passed through the encoder. This improves the model’s awareness of surface orientation and curvature.
- *Loss Supervision*: A cosine similarity loss enforces consistency between predicted and ground-truth normals:

$$\mathcal{L}_{\text{norm}} = \frac{1}{N} \sum_{i=1}^N \left(1 - \langle \mathbf{n}_i^{\text{pred}}, \mathbf{n}_i^{\text{gt}} \rangle \right) \quad (2)$$

This constraint improves local surface smoothness and preserves sharp features, which is critical for fine details such as carved text and motifs.

D. Loss Function

The model is trained with a combined loss that jointly supervises occupancy, signed distance, and normal predictions to improve overall shape and surface fidelity:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{occ}} \mathcal{L}_{\text{occ}} + \lambda_{\text{sdf}} \mathcal{L}_{\text{sdf}} + \lambda_{\text{norm}} \mathcal{L}_{\text{norm}} \quad (3)$$

Here, $\mathcal{L}_{\text{occ}}$ measures the occupancy prediction error, $\mathcal{L}_{\text{sdf}}$ measures the error in signed distance estimation, and $\mathcal{L}_{\text{norm}}$ evaluates the alignment between predicted and ground-truth surface normals. The weights λ_{occ} , λ_{sdf} , and λ_{norm} balance the contribution of each component.

E. Surface Extraction.

The fused signed distance field $s_{\text{fused}}(\mathbf{x})$ is converted to a mesh via Marching Cubes. Additional refinements include smoothing, decimation, and normal re-orientation to ensure high-quality output for visualization or downstream use.

IV. EXPERIMENTS

A. Datasets

We evaluate the proposed method on two datasets with distinct geometric characteristics.

ShapeNet. ShapeNet [25] is a large-scale repository of synthetic 3D CAD models covering diverse object categories. We use watertight meshes from ShapeNetCore and apply the damage simulation from [22] to introduce fractures of varying severity. ShapeNet mainly contains smooth, low-frequency surfaces, serving as a standard benchmark for 3D reconstruction methods.

Woodblock Character Dataset. This dataset consists of 3D character meshes extracted from scanned Vietnamese woodblocks. From the original collection of 90,259 character meshes, we select 5,000 high-quality characters for this study.

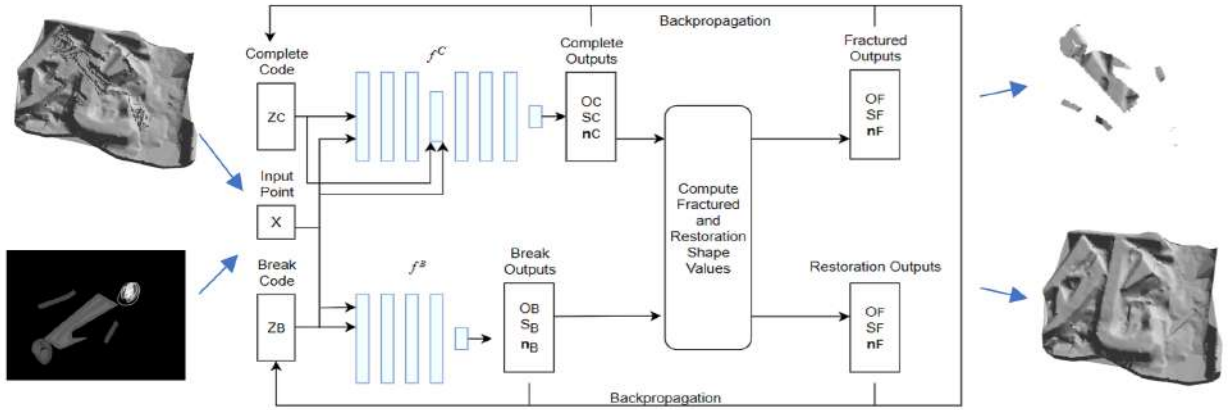


Fig. 4: Dual-branch normal-aware auto-decoder architecture (extending DeepJoin). Surface normals are injected into feature encoding to enhance local geometric detail.

Each mesh was preprocessed to ensure geometric completeness by making it watertight, adding side and bottom faces and estimating normal vectors to facilitate surface reconstruction. Figure 5 illustrates (a) an initial 3D character mesh and (b) the corresponding watertight mesh with estimated normals. The curated dataset is split into training, validation, and testing sets (70%/20%/10%). Compared to ShapeNet, these meshes exhibit high-frequency geometric details such as engraved strokes, sharp reliefs, and stylistic ornamentation.

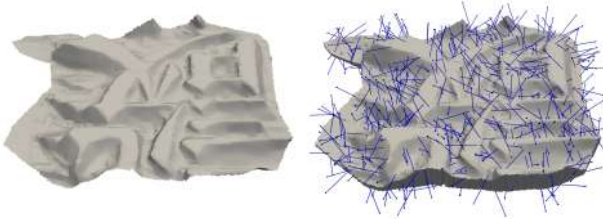


Fig. 5: 3D woodblock character meshes: (a) initial mesh, (b) watertight mesh and normal vectors.

B. Implementation Details

We compare our method with DeepJoin [24], a state-of-the-art learning-based approach for implicit surface reconstruction and shape joining. Both methods use the same backbone architecture and are implemented in PyTorch. Training employs the Adam optimizer with an initial learning rate of 1×10^{-4} , batch size 4, and early stopping based on validation loss. For all experiments, we set the loss weights in the ratio $\lambda_{occ} : \lambda_{sdf} : \lambda_{norm} = 4.5 : 4.5 : 1$. For ShapeNet, we follow the training protocol from DeepJoin. For the woodblock dataset, we train from scratch, increasing the latent vector size from 128 to 256 to capture more surface details, fractures, and normal vectors.

C. Evaluation Metrics

The results are reported using three complementary quantitative metrics, as follows:

- *Chamfer Distance (CD)* [26]: Measures bidirectional point-to-point distances between predicted and ground-truth surfaces, capturing geometric accuracy.
- *Normal Consistency (NC)* [27]: Computes the average cosine similarity between corresponding surface normals, assessing preservation of fine geometric details.
- *Non-Fractured Region Error (NFRE)* [19]: Quantifies reconstruction errors in undamaged regions, penalizing artificial discontinuities and ensuring topological integrity.

These metrics jointly evaluate both geometric fidelity and structural coherence. Quantitative comparisons and qualitative visualizations are presented in Section V.

V. RESULTS

A. Quantitative Results

We evaluate our approach on two datasets: the synthetic ShapeNet benchmark and the real world Nguyen Dynasty woodblock dataset, using the learning-based DeepJoin method as the primary baseline. Tables I and II summarize the results, where lower CD/NFRE and higher NC indicate better performance.

TABLE I: Quantitative results on ShapeNet.

Method	CD ↓	NC ↑	NFRE ↓
DSDf	0.072	0.583	0.181
ONet	0.141	0.423	0.634
ESSC	0.115	0.289	0.152
DeepJoin	0.062	0.631	0.038
Ours	0.067	0.655	0.037

On ShapeNet, our method achieves the highest NC (3.8% improvement over DeepJoin) and the lowest NFRE, while DeepJoin attains the best CD. Given that ShapeNet consists mainly of smooth synthetic shapes with low-frequency geometric variation, the improvements from our method are most pronounced in surface orientation accuracy rather than in point-to-surface distance.

On the Nguyen Dynasty woodblock dataset, our method delivers a large NC improvement (25.2%) and slightly better

TABLE II: Quantitative results on woodblock dataset.

Method	CD ↓	NC ↑	NFRE ↓
DeepJoin	0.081	0.572	0.052
Ours	0.093	0.716	0.049

NFRE, though DeepJoin achieves a lower CD. The pronounced NC gain demonstrates the advantage of our normal aware approach in capturing high frequency carving strokes and sharp relief edges, features critical for accurately restoring intricate cultural heritage artifacts.

Overall, while ShapeNet results show only moderate gains in some metrics, the consistently higher NC on both datasets confirms that incorporating normal vector information strengthens surface detail preservation. The larger improvements on the woodblock dataset suggest that our method is particularly well-suited for real-world cases with complex geometry and fine structural details.

B. Qualitative Results

To complement the quantitative analysis, we developed a customized visualization tool for side-by-side mesh comparison using PyQt5, PyVista, and Open3D, allowing multiple 3D view windows to be displayed simultaneously for easier observation and synchronizing viewpoint, zoom, tilt, and pan for consistent evaluation.

Figure 7 shows representative results from the Nguyen Dynasty woodblock dataset. The degraded inputs exhibit missing regions and worn surface details. Our reconstructions restore these areas with sharper edges and more faithful carving details, closely matching the original ground-truth meshes. In contrast, DeepJoin often produces over smoothed surfaces, losing subtle strokes and ornamental patterns.

C. Discussion

The results across datasets highlight the effectiveness of our approach. On synthetic, smooth, low-frequency models like ShapeNet, improvements are moderate, mainly enhancing normal consistency and preserving surface details. Visual inspection and weight tuning confirm that normal supervision is the key factor for NC improvement, while λ_{sdf} has a stronger impact on CD.

On real-world woodblock scans with high-frequency carvings, sharp edges, and complex topologies, the method achieves significant gains in geometric fidelity and structural coherence.

The slightly higher CD observed in our results reflects a trade-off between preserving local details and maintaining global smoothness. Generalization is challenging due to varying object geometries, but many heritage artifacts share similar high-frequency features are suitable for this approach. These findings demonstrate the advantage of normal-aware reconstruction for detailed cultural artifacts.

VI. CONCLUSION

In summary, we present a normal-enhanced fused shape representation framework for 3D Vietnamese woodblocks,

improving the recovery of fine-scale geometric features in degraded scans. Incorporating normal vector information enhances surface orientation consistency and preserves high-frequency details. Quantitative and qualitative results demonstrate improved normal consistency and faithful reproduction of intricate carvings and surface textures. Challenges such as training data quality, damage variability, and handling large missing regions remain open problems. Building on this work, we aim to develop a more general reconstruction strategy and extend the approach to other artifacts, including stone carvings and metal engravings, to support broader cultural heritage preservation.

REFERENCES

- [1] Regional Memory of the World Committee for Asia and the Pacific, *Memory of the World: 20 years of MOWCAP; Documentary Heritage on the Asia-Pacific Register 2018*. Gwangju: MOWCAP, 2018.
- [2] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.
- [3] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [4] W. Lorensen and H. Cline, “Marching cubes: A high resolution 3d surface construction algorithm,” *ACM SIGGRAPH Computer Graphics*, vol. 21, pp. 163–, 08 1987.
- [5] H. Hoppe, “Poisson surface reconstruction and its applications,” 06 2008, p. 10.
- [6] N. Duyen, V. Ta, T. Ma, M. Nguyen, X. Nguyen, and H. Le, “Quality assessment criteria and methods for 3d digital replica of historical printing woodblocks,” *Journal on Computing and Cultural Heritage*, vol. 16, 07 2022.
- [7] Y. Chen, J. Liu, L. Peng, Y. Wu, Y. Xu, and Z. Zhang, “Auto-encoding variational bayes,” *Cambridge Explorations in Arts and Sciences*, vol. 2, 02 2024.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Advances in Neural Information Processing Systems*, vol. 3, 06 2014.
- [9] A. Sauer, K. Schwarz, and A. Geiger, “Stylegan-xl: Scaling stylegan to large diverse datasets,” in *ACM SIGGRAPH 2022 Conference Proceedings*, ser. SIGGRAPH ’22. ACM, 2022, pp. 1–10.
- [10] Z. Shi, S. Peng, Y. Xu, A. Geiger, Y. Liao, and Y. Shen, “Deep generative models on 3d representations: A survey,” Oct. 2022, arXiv preprint arXiv:2210.14267, last revised 28 Aug 2023.
- [11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” 2017. [Online]. Available: <https://arxiv.org/abs/1612.00593>
- [12] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5099–5108.
- [13] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, “DeepSDF: Learning continuous signed distance functions for shape representation,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 165–174.
- [14] D. Tran, R. Ranganath, and D. Blei, “Deep and hierarchical implicit models,” 02 2017.
- [15] U. Demir and G. Unal, “Patch-based image inpainting with generative adversarial networks,” Mar. 2018, arXiv preprint arXiv:1803.07422.
- [16] J. Yu, Z. Lin, J. Yang, X. Shen, and X. Lu, “Free-form image inpainting with gated convolution,” 06 2018.
- [17] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Gool, “Repaint: Inpainting using denoising diffusion probabilistic models,” 01 2022.
- [18] W. Li, Z. Lin, Z. Kun, L. Qi, Y. Wang, and J. Jia, “Mat: Mask-aware transformer for large hole image inpainting,” 06 2022, pp. 10748–10758.
- [19] N. Lamb, S. Banerjee, and N. Banerjee, *DeepMend: Learning Occupancy Functions to Represent Shape for Repair*, 11 2022, pp. 433–450.

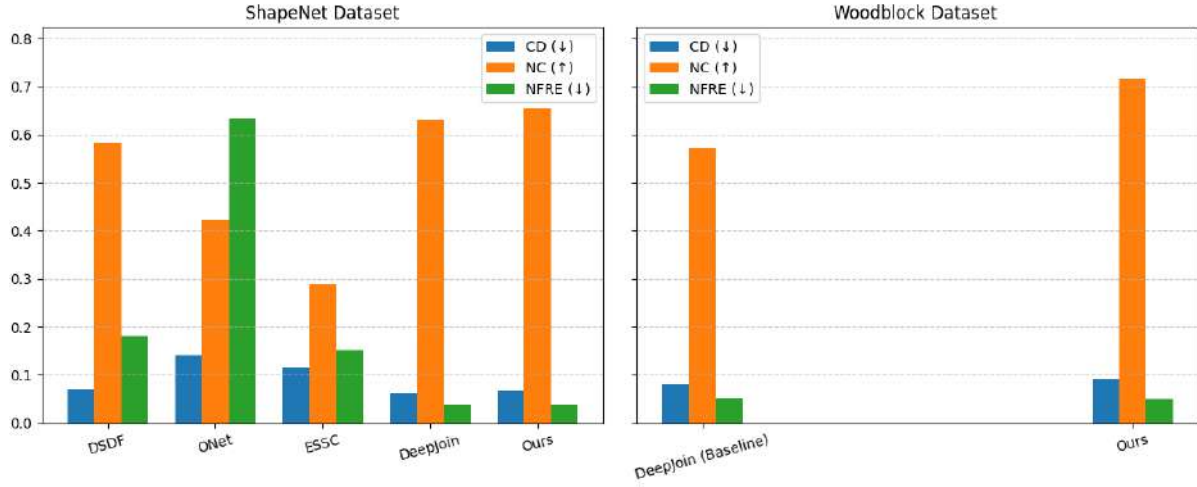


Fig. 6: Quantitative comparison on the ShapeNet dataset (left) and the Nguyen Dynasty woodblock dataset (right).

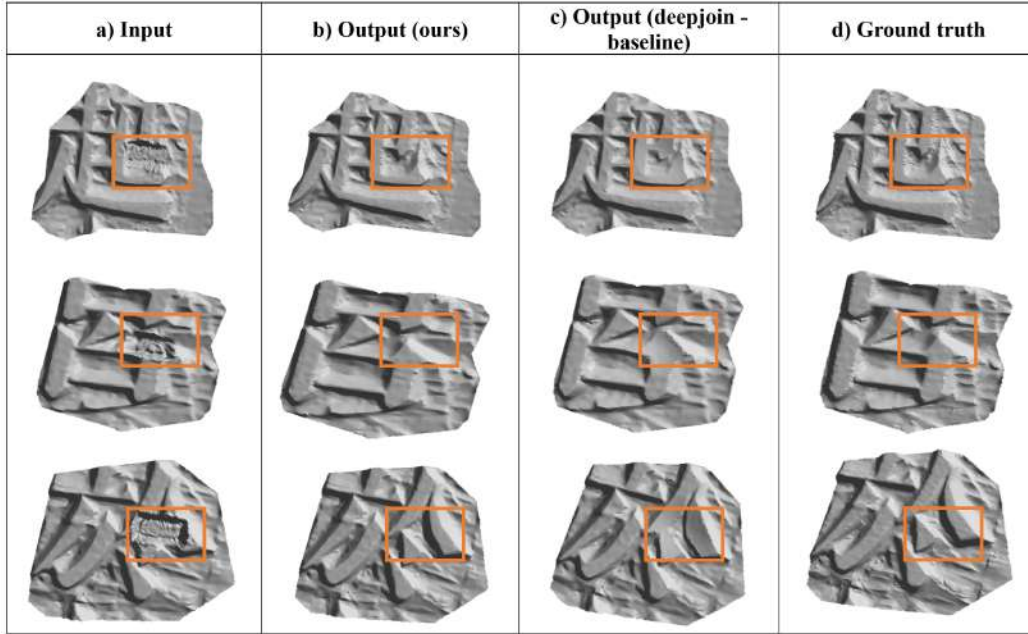


Fig. 7: Qualitative comparison on the Nguyen Dynasty woodblock dataset. From left to right: (a) degraded input mesh, (b) reconstruction by our method, (c) reconstruction by the baseline, and (d) ground-truth mesh prior to degradation.

- [20] —, “Mendnet: Restoration of fractured shapes using learned occupancy functions,” *Computer Graphics Forum*, vol. 41, pp. 65–78, 10 2022.
- [21] X. Song, N. Lamb, S. Banerjee, and N. K. Banerjee, “Pix2repair: Implicit shape restoration from images,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.18273>
- [22] N. Lamb, C. Palmer, B. Molloy, S. Banerjee, and N. K. Banerjee, “Fantastic breaks: A dataset of paired 3d scans of real-world broken objects and their complete counterparts,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4681–4691.
- [23] N. le Viet, V. Cuong Ta, T. Duyen Ngo, A. Nguyen Thai, S.-W. Jung, and T. Ha Le, “3d depth map inpainting for vietnamese historical printing woodblocks: A gated convolution approach,” in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6.
- [24] N. Lamb, S. Banerjee, and N. K. Banerjee, “Deepjoin: Learning a joint occupancy, signed distance, and normal field function for shape repair,” *ACM Transactions on Graphics*, vol. 41, no. 6, p. 1–10, Nov. 2022. [Online]. Available: <http://dx.doi.org/10.1145/3550454.3555470>
- [25] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, “ShapeNet: An Information-Rich 3D Model Repository,” Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep. arXiv:1512.03012 [cs.GR], 2015.
- [26] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, “Parametric correspondence and chamfer matching: two new techniques for image matching,” in *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI’77. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1977, pp. 659–663.
- [27] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy networks: Learning 3d reconstruction in function space,” 2019. [Online]. Available: <https://arxiv.org/abs/1812.03828>

Enhancing Urban Traffic Control with Priority-Aware Multi-Objective Reinforcement Learning

LY Nhat Huy¹, DOAN Nhat Quang^{*1}, GIANG Anh Tuan¹,

LE Huu Ton², NGUYEN Hoang Ha¹, TRAN Hoang Tung¹

¹ University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology,
18, Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam

² CMC University, Hanoi, Vietnam

Abstract—Efficient urban traffic control is crucial for enhancing public safety, reducing travel times, and ensuring timely responses from emergency vehicles, including ambulances and fire engines. While reinforcement learning (RL) has shown promise in adaptive traffic management, existing approaches often lack the flexibility to dynamically prioritize emergency vehicles. In this paper, we propose MOP-MADDPG, a Multi-Objective Prioritized extension of the Multi-Agent Deep Deterministic Policy Gradient framework, which enables decentralized traffic signal agents to optimize for both regular and emergency scenarios adaptively. Our method introduces context-aware learning and dynamically reweighted reward functions based on vehicle priority, ensuring emergency passage while maintaining overall traffic efficiency. We further perform hyperparameter tuning to optimize reward coefficients for robust performance. Experiments conducted on a real-world traffic network from Hanoi using the SUMO simulator demonstrate that MOP-MADDPG reduces waiting time by 75% and ensures a travel speed for emergency vehicles, while maintaining competitive performance for regular traffic.

Index Terms—reinforcement learning, Q-learning, traffic management, multi-level agent, multi-objective reward

I. INTRODUCTION

Optimizing urban traffic networks while accounting for emergency vehicles, such as ambulances and fire engines, presents a significant challenge. These scenarios introduce stringent time constraints and dynamic priorities, transforming the traffic signal control problem into a combinatorially complex task. In fact, balancing the efficient movement of regular traffic alongside the timely clearance of emergency vehicles is widely recognized as an NP-hard problem [1], [2]. Traditional rule-based traffic systems often fail to adapt quickly enough to such rapidly changing conditions, resulting in suboptimal traffic flow and delayed emergency responses.

RL has emerged as a promising solution for adaptive traffic signal control [3], enabling agents to learn optimal strategies directly from interaction with the environment. However, existing approaches often treat all vehicles equally or rely on fixed

preemption rules, lacking the flexibility to adjust priorities in real-time when emergency scenarios arise dynamically.

To address these limitations, we propose a novel approach based on an enhanced Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework [4], augmented with context awareness and dynamic reward shaping. Our method, called MOP-MADDPG (Multi-Objective Prioritized MADDPG), is designed to adapt its behavior based on the presence and location of emergency vehicles while maintaining overall traffic efficiency [5]. Specifically, we introduce a dynamic reward function that is reweighted depending on traffic conditions and vehicle priority, ensuring that emergency vehicles receive immediate passage through intersections without severely penalizing the performance of regular traffic flow [6].

Our main contributions to this paper are as follows: Firstly, we incorporate contextual awareness into the learning framework to distinguish between normal and emergency traffic scenarios, allowing agents to adapt their strategies accordingly. Secondly, we develop a multi-objective reward formulation that adjusts weights based on the presence and urgency of emergency vehicles, integrating multiple traffic performance metrics, such as queue length, waiting time, and vehicle speed. Finally, we conduct extensive hyperparameter tuning to identify optimal reward coefficient configurations, ensuring balanced and resilient performance under diverse traffic densities and emergency conditions.

We evaluate the proposed framework on a real-world traffic network extracted from a Hanoi map section using the SUMO simulator. Experimental results demonstrate that our MOP-MADDPG method significantly reduces 30-50% waiting times and improves average travel speeds for both regular and emergency vehicles compared to baseline methods.

The rest of our paper is as follows: Section II summarizes existing solutions in the literature. Section III provides background knowledge on reinforcement learning. The research problem is modeled in Section IV. Our proposal solution is given in Section V. Section VI provides experiment results to

^{*}Corresponding author: doan-nhat.quang@usth.edu.vn

validate our proposal. Finally, we conclude in Section VII.

II. RELATED WORKS

The literature implies that optimizing traffic flow problems for emergency vehicles requires intelligent control strategies that minimize waiting times while prioritizing EVs. Reinforcement learning has been widely applied to this problem, yielding diverse approaches. We survey a series of scientific papers that apply RL to traffic management with EV priority, comparing their RL techniques, unique features, and pros/cons. One may classify them into Deep Q-learning methods, Actor-Critic (policy gradient) methods, and Multi-Agent RL coordination.

Deep Q-learning techniques [3], [7], [8] use neural networks to approximate the Q-value function, enabling an agent to learn optimal actions for traffic signals. These approaches treat traffic signal timing as a discrete action selection problem, where the goal is to minimize waiting times (or queue lengths) with heavy penalties for delaying emergency vehicles. Despite huge improvements ([7] reports a 40% decrease in emergency vehicle delay under multiple-EV scenarios), the pure Q-learning approach still suffers from some limitations due to its discretization nature.

Actor-critic RL methods (including policy gradient algorithms like Advantage Actor-Critic and Proximal Policy Optimization) have been applied to traffic problems to handle continuous or large action spaces and provide more stable convergence [9]–[11]. In the context of emergency vehicle priority, actor-critic approaches often shine when the action involves timing durations or continuous maneuvers (like vehicle acceleration/lane-changing) or when combining decisions in a parameterized way. In those works, actor-critic methods are chosen for scenarios requiring either continuous actions (timing or vehicle control) or stable multi-agent learning. Besides making decisions, coordinating multiple agents in optimizing traffic flow problems is also crucial due to the correlation between multiple traffic intersections.

Multi-agent reinforcement learning (MARL) is central to managing traffic networks where multiple intersections or vehicles must work in tandem. Several of the above papers already incorporate MARL elements (EMVLight [9], RL-CMAS [10], and EVPRT [12] all involve multiple agents). Multi-agent RL in traffic EV priority is about balancing cooperation and complexity. Methods like EVPRT and EMVLight introduced ways for infrastructure agents to cooperate (via shared policies or communication networks). At the same time, [6], [13] focus on vehicle agents cooperating or even multiple EVs cooperating. A common thread is the use of centralized training, giving the MARL system a bird's-eye view during learning, to achieve coordination that decentralized methods (each agent selfishly optimizing) would miss.

Techniques such as value decomposition (Q-LSTM) [13] and graph neural networks (GAT in EVPRT) [12] are employed to handle the combinatorial complexity of multi-agent interactions by imposing structure (factorizing the value or exploiting road network topology). These innovations make

the RL more sample-efficient and the learned policies more effective in a multi-agent context.

III. PRELIMINARY

A. Preliminary

In an RL problem, $D = \{(s_t, a_t, r_t, s_{t+1})\}$ is a set of data tuple consisting of a current state s_t , a taken action a_t , a reward function r_t , and a transition must be defined s_{t+1} . At each time step t , the agent is in a state $s_t \in S$ and selects an action $a_t \in A$. The environment evolves by transitioning the agent to a new state s_{t+1} according to the transition function: $T(s_t, a_t)$. Simultaneously, a reward r_t is given to the agent that is expressed by the reward function: $R(s_t, a_t, s_{t+1})$.

The agent needs to respect a policy $\pi(s)$, which defines the probability of selecting one of the available actions. The objective is to determine the optimal policy π that optimizes the accumulated reward R . Learning an optimal policy is to estimate the state-action value function $Q^\pi(s, a)$. This function is defined as: $Q^\pi(s, a) = \mathbb{E}[R \mid s_t = s, a_t = a]$

B. Multi-Agent Deep Deterministic Policy Gradient - MADDPG

MADDPG is a multi-agent reinforcement learning algorithm designed to address the challenges of coordination and stability in environments involving multiple interacting agents. During training, each agent, a Q-network, maintains its own actor-critic architecture. The actor learns a deterministic policy based on local observations. At the same time, the critic is trained with access to global information, including the observations and actions of all agents in the environment.

The reward for agent ag_i at time step t is given by: $r_i^t = R_i(s^t, a_1^t, a_2^t, \dots, a_n^t)$ where a_i^t is the action taken by the agent ag_i at time t . Each agent in MADDPG uses a centralized critic that considers the global state and actions of all agents $Q_i(s, a_1^t, a_2^t, \dots, a_n^t)$. The critic is updated using the Bellman equation: $y_i = r_i + \gamma Q_i(s^{t+1}, a_1^{t+1}, a_2^{t+1}, \dots, a_n^{t+1})$. The actor policy μ_i for agent ag_i is updated using the deterministic policy gradient:

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s, a \sim D} [\nabla_{\theta_i} \mu_i(o_i) \nabla_{a_i} Q_i(s, a_1, \dots, a_n)]$$

IV. PROBLEM STATEMENTS

In this context, we address the RL problem of traffic management in a realistic urban scenario. A traffic network is modeled using a real-world map comprising n signalized intersections. Each intersection is managed by a corresponding agent ag_i where $i = 1, 2, \dots, n$. These agents interact with the environment by receiving observations in the form of feature vectors that capture the dynamic aspects of nearby vehicles and traffic signals.

A. States

Each agent ag_i works on the data tuple $D_i = \{(s_t, a_t, r_t, s_{t+1})\}$. At each time step t , the environment provides a state s_t , which includes the following components:

Each intersection may support up to several-way, in each way, there are some possible lanes. So, each intersection is

totally described by p lanes. Fig. 1 captured from SUMO ¹ shows 20 different lanes in an intersection.

For each lane in the way, we define traffic data as $d_i = \{v_m^i, v_h^i\}$ where $i = 1, \dots, p$ and v_m^i, v_h^i denote the number of moving and halting vehicles, respectively. Fig. 2, captured from SUMO, illustrates an example of these vehicle types in an intersection. Typically, moving vehicles are observed on green-light lanes; otherwise, halting vehicles accumulate before red lights. A state s_t encapsulates the full distribution of both moving and stopping vehicles across all lanes in the road network.

Additionally, the state at time t should also contain the position and waiting time of any emergency vehicle present in the intersection. The position is represented by a one-hot vector of length p , where value 1 indicates the exact location of the emergency vehicle.



Fig. 1: Moving direction samples in an intersection.



Fig. 2: Sample for moving and halting vehicles in an intersection.

B. Actions

At each step, an agent ag_i selects an action a_t to transition the system from state s_t to the next state s_{t+1} . In this setup, an

action corresponds to toggling the traffic light phase—either switching from red to green or vice versa. Each phase is held for a vary duration to allow vehicles to proceed. For simplicity in implementation, traffic light phases are encoded in binary form, with green = 1 and red = 0.

Each intersection is equipped with q traffic lights, represented as a binary vector of length q , ordered as [North, East, South, West]. For example, in a standard four-way intersection, the vector [1, 0, 1, 0] indicates that the north–south direction has green lights, while the east–west direction has red lights.

C. Regular Rewards

The reward function in our framework is carefully formulated to enable each agent ag_i to make effective traffic signal switches for both regular and emergency traffic conditions. Decentralized agents are expected to prioritize the passage of these vehicles over others to ensure unobstructed movement.

Under regular traffic conditions, rather than relying on a single objective, the reward integrates multiple traffic performance metrics [14], [15], enabling the agent to make decisions that contribute to system-wide optimization. The following criteria are selected to evaluate the reward function:

- Waiting time f_{wt} : The total cumulative waiting time of vehicles approaching or halting at the intersection is minimized. By reducing waiting time, the system aims to enhance overall traffic flow efficiency.
- Queue length f_{ql} : The agent attempts to minimize the number of vehicles queued at each lane. A shorter queue length generally reflects smoother traffic movement and reduced congestion. The change δf_{ql} in f_{ql} between time steps t and $t + 1$ should be considered in the reward function; however, only the remaining queue length is currently incorporated.
- Vehicle speed f_{sp} : The system encourages higher average speeds within safe operational limits. Low speeds may indicate congestion, while consistently high speeds (when appropriate) reflect effective signal timing decisions.

Thus, under normal conditions, the reward function is affected by the set of performance factors $f_{gi} = \{f_{wt}, f_{ql}, \delta f_{ql}, f_{sp}\}$ where $i = 1, \dots, 4$.

D. Emergency Rewards

In emergency scenarios, the reward function must be adaptively adjusted to account for the presence and priority of emergency vehicles with the following factors:

- Emergency queue length f_{ql} : Similar to the queue length in regular traffic, we propose a reward mechanism that incentivizes changes in queue length δf_{ql} while imposing penalties when queues remain persistent.
- Waiting time f_{wt} : The reward function penalizes delays encountered by emergency vehicles. Minimizing their waiting time becomes a primary objective.
- Emergency vehicle speed f_{sp} : The high speed of emergency vehicles is favored in the reward. The agent should emphasize maintaining or increasing this speed by adjusting signal phases.

¹<https://sumo.dlr.de/docs/index.html>

Similarly, in emergency conditions, the reward function is affected by the set of performance factors $\{f_{wt}, f_{ql}, \delta f_{ql}, f_{sp}\}$ where $i = 1, \dots, 4$. This emergency-aware reward design ensures that agents can prioritize critical traffic in real time while still retaining the ability to return to standard optimization goals once the emergency scenario subsides.

V. PROPOSED METHOD

Our approach, MOP-MADDPG, builds upon the MADDPG algorithm. However, our method introduces a priority-aware extension specifically designed to handle emergency traffic scenarios. The key point is to adjust the reward function based on the context of involving emergency vehicles.

A. Agents

An agent ag_i is described by: a local policy $\pi(s)$ to choose optimal actions at a given state; and a centralized critic $Q^\pi(s, a_1^t, a_2^t, \dots, a_n^t)$ determined by the actions of all n agents. During training, the centralized critic uses global information to stabilize learning. However, the agents operate independently at test time while still benefiting from global knowledge during learning. A feed-forward neural network of two 64-unit layers is implemented to estimate the Q-value for each agent's action.

B. Reward Design

In normal conditions, the reward function focuses on general traffic optimization metrics, such as waiting time, queue length, and average speed. However, in emergency scenarios, the reward is reweighted to prioritize emergency vehicle throughput. These factors are emphasized during emergencies, allowing the agent to adapt its behavior in real time. Each agent ag_i receives a reward defined by the following function:

$$r_i = \sum_{j=1}^4 \alpha_j \cdot fg_j + \sum_{j=1}^4 \gamma_j \cdot fe_j \quad (1)$$

where α_j and γ_j correspond to the weighting coefficients for each component of the reward factors fg_j and fe_j , respectively. These weights are treated as hyperparameters and are carefully tuned to optimize the performance of the reward function in our experiments. This ensures that emergency vehicles are given preferential treatment in signal phase decisions, which can significantly reduce their travel time and improve overall response effectiveness.

C. Bellman Equation Update

In the proposed method, each agent maintains an individual actor and critic network. The Bellman equation is used to compute the target Q-value during training, which forms the basis for updating the critic network through the minimization of the temporal-difference error. This allows the agent to learn the expected return of its actions in the presence of other learning agents.

For each agent ag_i , the Bellman target is defined as:

$$y_i = r_i + \gamma(1 - d) \cdot Q_i^{\text{target}}(s^{t+1}, a_1^{t+1}, a_2^{t+1}, \dots, a_n^{t+1})$$

where:

- r_i is the immediate reward received by agent ag_i ,
- $\gamma \in [0, 1)$ is the discount factor,
- $d \in \{0, 1\}$ is the terminal flag indicating the end of an episode,
- Q_i^{target} is the target critic network for agent ag_i ,
- s^{t+1} is the next global state after taking action at time t ,
- a_j^{t+1} is the next action for agent ag_j , computed by the corresponding target actor network.

The current Q-value is obtained from the critic network as $Q_i(s^t, a_1^t, a_2^t, \dots, a_n^t)$, where s^t is the current global state and a_j^t denotes the action of agent j at time step t . The critic loss function is defined as the mean squared error between the predicted Q-value and the Bellman target:

$$\mathcal{L}_i^{\text{critic}} = \frac{1}{B} \sum_{k=1}^B (Q_i(s_k^t, a_k^t) - y_i^k)^2$$

where B is the batch size, and the summation is taken over a minibatch of transitions sampled from the replay buffer.

The model employs a centralized training framework with decentralized execution by conditioning each agent's critic on the actions of all agents. This allows the learning process to remain stable in the presence of multiple agents while supporting independent action selection during test time.

D. Exploration Strategy

We employ a ϵ -greedy strategy, a common and effective approach in reinforcement learning for balancing the trade-off between exploration and exploitation. Given a state s_t , an agent ag_i selects an action based on a probabilistic decision given by a Q-network as following

- **Exploration:** With probability ϵ , the agent selects a random action from the action set. This phase is crucial during the early stages of training, where the policy is underdeveloped and broad exploration is needed to discover informative state-action transitions.
- **Exploitation:** Otherwise, the agent selects the action that maximizes its own critic's estimate of the Q-value, conditioned on the joint state and actions of all agents.

$$a_i^t = \arg \max_{a_i} Q_i(s^t, a_1^t, \dots, a_i, \dots, a_N^t)$$

VI. EXPERIMENTS

We conduct a variety of experimental scenarios to evaluate and compare the effectiveness of the proposed method against baseline approaches. A specific area of Hanoi is chosen for detailed analysis, using accurate urban road information. The corresponding map is shown in Fig. 3, where five key intersections are highlighted, each representing an individual agent.

A. Models for Comparisons

In Table I, the State column represents the length of the state vector used during training. A state with full-lanes contains all moving and halting vehicles of every lane in the network. For MADDPG 2, local lane refers to training without

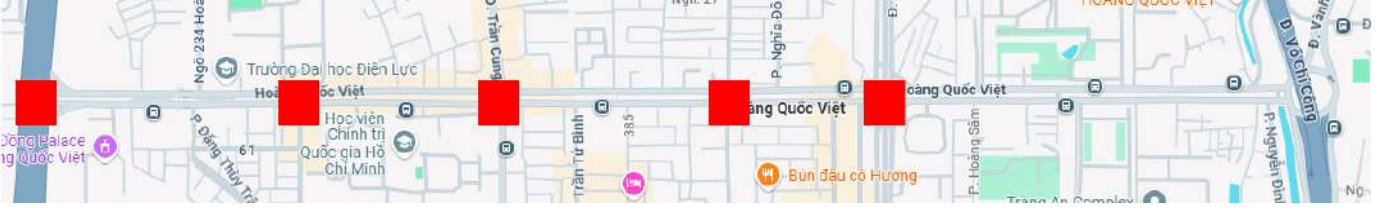


Fig. 3: Actual map of Hoang Quoc Viet street in Hanoi with 5 selected light traffic (extracted from Google Map).

TABLE I: Comparison of Reinforcement Learning Models

Model	State	Reward	Emergency
MOP-MADDPG	Full lanes	Multi-objective	Yes
MADDPG 1	Full lanes	Multi-objective	No
MADDPG 2	Local lanes	Multi-objective	Yes
MADDPG 3	2nd-order lanes	Multi-objective	Yes

incorporating the state from neighboring lanes. In contrast, second-order lanes of MADDPG 3 indicate that the agent considers two levels of neighboring lanes in its input. The Emergency column describes whether the reward function is integrated with emergency factors.

B. Environment Settings

We need to select several environment parameters for SUMO as follows:

- Experiments are trained a total number of 1,720 vehicles, including 4 emergency vehicles. The random route algorithm of SUMO handles their appearance. We exploit the exploration step for the scenario of 2,580 vehicles, including 6 emergency vehicles.
- Each vehicle is characterized by its type (car or motorbike) and speed. The speeds are randomly assigned within the range [0, 50] km/h for cars, [0, 40] km/h for motorbikes, and [0, 60] km/h for emergency vehicles.
- The duration of each light phase is set to 10 seconds for regular traffic conditions. In contrast, for emergency scenarios, it is evaluated and potentially switched every 1 second.
- Each model is trained and evaluated over 100 epochs to ensure performance consistency and convergence.

C. Hypertuning

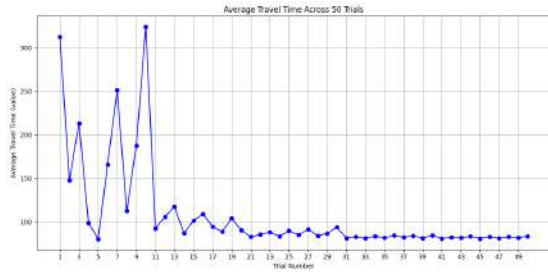


Fig. 4: Average waiting time of 1720 regular vehicles excluding emergency vehicles in 50 combination trials.

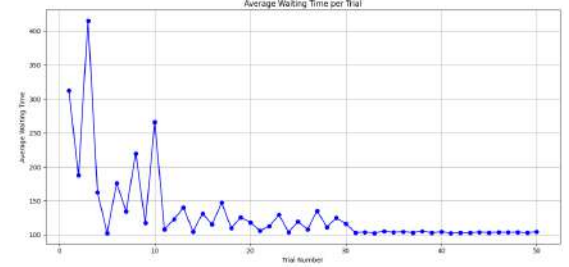


Fig. 5: Average waiting time of 1716 regular vehicles and 4 emergency vehicles in 50 combination trials.

With the previous setting, we examine the coefficient factors α and γ in Equation 1. Since vehicles require a certain amount of time to pass through intersections, the impact of queue length on traffic performance is more significant compared to factors such as waiting time and speed. Consequently, the corresponding reward weights for queue length and its variation should be assigned higher values, whereas lower weights are chosen for the remaining factors. The objective in this step is to identify the optimal combination (trial) of coefficient factors within a specified range that minimizes the average travel time, thereby achieving the best overall traffic performance.

- Regular factors: $\alpha = \{\alpha_{f_{wt}} = [0.0005, 0.15], \alpha_{f_{ql}} = [-0.6, -0.05], \alpha_{\delta f_{ql}} = [0.05, 1.5], \alpha_{f_{sp}} = [0.01, 0.6]\}$
- Emergency factors: $\gamma = \{\gamma_{f_{wt}} = [-15.0, -1.0], \gamma_{f_{ql}} = [-0.15, -0.01], \gamma_{\delta f_{ql}} = [1.0, 15.0], \gamma_{f_{sp}} = [-6.0, 0.0]\}$

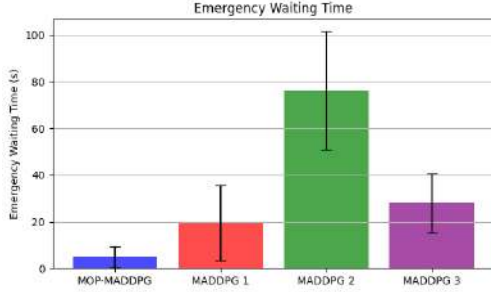
We execute 50 uniformly distributed trials within the defined range to find the best trials, as shown in Fig. 4 and 5. In Fig. 4, Trial 5 achieves the best performance with the coefficient configuration $\{\alpha_{f_{wt}} = 0.01, \alpha_{f_{ql}} = -0.2, \alpha_{\delta f_{ql}} = 0.5, \alpha_{f_{sp}} = 0.2\}$. In Fig. 5, the optimal result is obtained with the coefficient set $= \{\gamma_{f_{wt}} = -5.0, \gamma_{f_{ql}} = -0.05, \gamma_{\delta f_{ql}} = 5, \gamma_{f_{sp}} = -2.0\}$. These optimized combinations of coefficients are adopted in our experiments to configure the reward function for the learning framework.

D. Evaluation

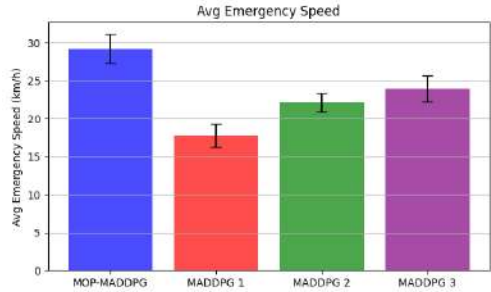
Figure 6 presents a comparison of the average waiting times for emergency scenarios across the models listed in Table I. The average results in this table are based on 20 simulation runs involving 2,580 vehicle. In Figure 6a, the bars represent the mean waiting time in seconds, while the error bars indicate the standard deviation for priority vehicles. The

TABLE II: Performance comparison of different models.

Model (Mean $\pm$ std)	Emergency Vehicle			All Vehicles	
	Waiting Time (s)	Travel Time (s)	Speed (km/h)	Travel Time (s)	Speed (km/h)
MOP-MADDPG	4.94 $\pm$ 4.38	113.37 $\pm$ 5.78	29.20 $\pm$ 1.94	94.54 $\pm$ 1.54	29.20 $\pm$ 0.14
MADDPG 1	19.57 $\pm$ 16.13	129.78 $\pm$ 18.36	17.78 $\pm$ 1.51	98.77 $\pm$ 2.88	28.66 $\pm$ 0.36
MADDPG 2	76.13 $\pm$ 25.45	188.01 $\pm$ 26.04	22.14 $\pm$ 1.19	95.41 $\pm$ 1.96	30.00 $\pm$ 0.22
MADDPG 3	27.92 $\pm$ 12.58	140.54 $\pm$ 12.75	23.94 $\pm$ 1.70	92.45 $\pm$ 0.91	29.27 $\pm$ 0.11



(a) Performance on average waiting time of 6 emergency vehicles.



(b) Performance on average speed time of 6 emergency vehicles.

Fig. 6: Comparison in different performance measures.

MOP-MADDPG model achieves the lowest waiting time at approximately 4.94 seconds, a 75% reduction in delay for emergency vehicles, but keeps good speed around 29.2km/h for six emergency vehicles as in Fig. 6b. MOP-MADDPG is still able to maintain comparable travel times and speeds for all vehicles within 3–5% of those achieved by the other models..

Moreover, Table. II demonstrates the good performance of the proposed MOP-MADDPG method over the baseline MADDPG variants according to multiple metrics. Most notably, MOP-MADDPG achieves the best scores for the emergency scenario. This indicates that MOP-MADDPG responds more promptly to critical scenarios.

VII. CONCLUSION

In this work, we proposed MOP-MADDPG, a novel Multi-Objective Prioritized Multi-Agent Deep Deterministic Policy Gradient framework for managing traffic environments. Our approach integrates multi-objective optimization with the MADDPG algorithm, enabling agents to balance general traffic flow while prioritizing emergency vehicles efficiently. Through comprehensive experiments and comparisons with multiple MADDPG variants, we demonstrated that MOP-MADDPG significantly outperforms baseline methods

in terms of emergency response efficiency, achieving the lowest emergency waiting and travel times. Furthermore, it maintains competitive average travel and waiting times for all vehicles. Future work will explore scalability to larger networks with real-time data.

REFERENCES

- [1] W. Genders and S. Razavi, "Using a deep reinforcement learning agent for traffic signal control," 2016. [Online]. Available: <https://arxiv.org/abs/1611.01142>
- [2] M. T. Rafique, A. Mustafa, and H. Sajid, "Reinforcement learning for adaptive traffic signal control: Turn-based and time-based approaches to reduce congestion," 2024. [Online]. Available: <https://arxiv.org/abs/2408.15751>
- [3] F. Kiani and Ö. F. Saraç, "A novel intelligent traffic recovery model for emergency vehicles based on context-aware reinforcement learning," *Information Sciences*, vol. 619, pp. 288–309, 2023.
- [4] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [5] C. Imane, M. El Khaïli, and M. Mestari, "Multi-agent traffic management system for emergency vehicle," in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2022, pp. 1–8.
- [6] P. Lin, Z. Chen, M. Pei, Y. Ding, X. Qu, and L. Zhong, "Multiple emergency vehicle priority in a connected vehicle environment: A cooperative method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 173–188, 2023.
- [7] M. Shamsi, A. Rasouli Kenari, and R. Aghamohammadi, "Reinforcement learning for traffic light control with emphasis on emergency vehicles," *The Journal of Supercomputing*, vol. 78, no. 4, pp. 4911–4937, 2022.
- [8] B. Sadek, J. D. Godier, M. J. Cassidy, and C. F. Daganzo, "Traffic signal plans to decongest street grids," *Transportation research part B: methodological*, vol. 162, pp. 195–208, 2022.
- [9] H. Su, Y. D. Zhong, J. Y. Chow, B. Dey, and L. Jin, "Emvlight: A multi-agent reinforcement learning framework for an emergency vehicle decentralized routing and traffic signal control system," *Transportation Research Part C: Emerging Technologies*, vol. 146, p. 103955, 2023.
- [10] P. Rosayyan, J. Paul, S. Subramaniam, and S. I. Ganesan, "Reinforcement learning based control scheme for emergency vehicle preemption with edge computing," *International journal of intelligent transportation systems research*, vol. 21, no. 1, pp. 48–62, 2023.
- [11] C. Ding, I. W.-H. Ho, E. Chung, and T. Fan, "V2x and deep reinforcement learning-aided mobility-aware lane changing for emergency vehicle preemption in connected autonomous transport systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 7281–7293, 2024.
- [12] B. Liu, W. Peng, J. Liu, W. Han, and E. Wang, "Evprpt: A marl-based approach for efficient passage of emergency vehicles in urban vehicular networks," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 4357–4362.
- [13] Z. Li, Q. Wang, J. Wang, and Z. He, "A flexible cooperative marl method for efficient passage of an emergency cav in mixed traffic," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [14] H. Duan, Z. Li, and Y. Zhang, "Multiobjective reinforcement learning for traffic signal control using vehicular ad hoc network," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, 12 2010.
- [15] T. Saiki and S. Arai, "Flexible traffic signal control via multi-objective reinforcement learning," *IEEE Access*, vol. 11, pp. 75 875–75 883, 2023.

CABENCH: Benchmarking Composable AI for Solving Complex Tasks through Composing Ready-to-Use Models

Tung-Thuy Pham, Duy-Quan Luong, Minh-Quan Duong, Trung-Hieu Nguyen,
Thu-Trang Nguyen*, Son Nguyen, and Hieu Dinh Vo

Faculty of Information Technology

University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
{25025009, 23020140, 23020138, 21020017, trang.nguyen, sonnguyen, hieuvd}@vnu.edu.vn

Abstract—Composable AI offers a scalable and effective paradigm for solving complex AI tasks by decomposing them into sub-tasks, each handled by ready-to-use models. However, systematically evaluating methods under this setting remains largely unexplored. This paper introduces CABENCH, the first public benchmark for composable AI, comprising 70 realistic tasks, along with a curated pool of 700 models across multiple modalities and domains. We also propose an evaluation framework to enable end-to-end assessment of composable AI solutions. To establish initial baselines, we provide human-designed reference solutions and compare their performance with two LLM-based approaches. Our results demonstrate the promise of composable AI in addressing real-world challenges while underscoring the need for methods capable of automatically generating effective execution pipelines. This work lays the foundation for future research toward building scalable, efficient AI systems through principled reuse and orchestration of existing models.

Index Terms—Composable AI, Problem formulation, Benchmark, Task decomposition, Model selection, Pipeline composition

I. INTRODUCTION

In recent years, the rapid progress of Artificial Intelligence (AI) has profoundly influenced a wide range of domains, including computer vision [1], autonomous driving system (ADS) [2], healthcare [3], finance [4], etc. AI models have been integrated into modern systems to perform diverse tasks with remarkable performance. These advancements have unlocked the potential of applying AI to tackle increasingly complex problems in real-world scenarios.

However, many real-world tasks are inherently complicated, requiring multiple steps, modalities, and processing components. For instance, a medical diagnostic system may involve the integration of various components such as image analysis, clinical note understanding, and risk assessment. Building a single monolithic AI model capable of handling such diverse and interdependent functionalities is often infeasible [5]. *This challenge highlights the need for a modular and flexible approach for solving complex AI tasks in practical settings.*

Furthermore, even when a complex task is decomposed into well-defined sub-modules, the process of developing

and training specialized models for each component remains highly resource-intensive. It often requires substantial amounts of data, computational resources, and domain expertise. Meanwhile, the AI community has released a large collection of well-trained ready-to-use models on public model hubs such as *Hugging Face*¹. These models cover a wide range of modalities and tasks, providing a valuable resource for accelerating AI development. This presents a promising opportunity: *instead of training new models from scratch for every complex task, it is more efficient to decompose the problem into manageable sub-tasks and compose appropriate ready-to-use models to solve them.* Leveraging these existing resources facilitates the construction of AI systems which capable of tackling real-world challenges more effectively. Despite this potential, *the problem of automatically decomposing complex AI tasks and composing ready-to-use well-trained models into executable pipelines remains largely underexplored.*

Addressing this problem poses several challenges. First, effective task decomposition must align with the high-level objectives of the original problem while remaining grounded in the practical capabilities of available well-trained models. Defining sub-tasks without considering the functionality and availability of the existing models can lead to decompositions that are impractical to implement. Second, selecting and orchestrating appropriate models for each sub-task requires a comprehensive understanding of both the functionalities of these models and the inter-dependencies among sub-tasks to ensure effective collaboration. Finally, composing the outputs from different models into a coherent final solution entails addressing format mismatches, managing error propagation across stages, and integrating heterogeneous outputs into a unified and reliable result.

In this paper, we take an initial step toward addressing these challenges by *formulating the problem of Composable AI (CA)*, which seeks to automatically decompose complex tasks into sub-tasks and compose suitable well-trained models to solve them. To support systematic investigation under this setting, we introduce, CABENCH, a *carefully designed*

*Corresponding author.

¹<https://huggingface.co>

benchmark comprising **70 realistic tasks** where CA is essential, along with a curated model pool of approximately **700 ready-to-use well-trained models**. The tasks and models in CABENCH span a wide range of topics and modalities, ensuring diversity and practical relevance. In addition, we propose a *full-pipeline evaluation framework* that enable end-to-end assessment of CA methods, from input tasks to final outputs. This work aims to establish a foundation for future research in building scalable and efficient AI systems through the principled reuse of existing resources.

To evaluate the practical feasibility of CA, we conduct several experiments comparing two Large Language Model (LLM)-based strategies against *Human-designed Reference Solution* across 70 tasks in CABENCH. The two LLM-based strategies are: (1) *Prompt-to-Solve*, where the LLM is prompted to directly solve queries of a given task, and (2) *Prompt-to-Pipeline*, where the LLM is prompted to construct a solution pipeline by leveraging the provided model pool. *Prompt-to-Solve*, which can exploit the LLM’s knowledge to produce direct answers, consistently outperforms *Prompt-to-Pipeline*. For instance, *Prompt-to-Solve* achieves an average accuracy of 0.56, which is double the performance of *Prompt-to-Pipeline*. Nonetheless, *Prompt-to-Solve* still falls short of *Human-designed Reference Solution* with performance gaps reaching up to 90%. *These results show that while LLMs can address many tasks to some extent, specialized models orchestrated through carefully designed pipelines remain significantly more effective, particularly for tasks requiring compositional reasoning and structured execution.*

The contributions of this paper are as follows:

- 1) *CA Problem Formulation*: We formally define the problem of Composable AI (CA).
- 2) *CABENCH*: We introduce CABENCH, the first public benchmark comprising 70 CA tasks, along with a curated model pool of 700 ready-to-use well-trained models.
- 3) *Evaluation Framework*: We propose a full-pipeline evaluation framework for systematical end-to-end assessment of CA methods.
- 4) *Human-Designed Reference Solutions*: We design and evaluate a manual CA workflow to illustrate practical strategies for task decomposition and model composition.
- 5) *Baselines*: We conduct several experiments which automatically solve the tasks in the benchmark and provide the results as baselines for future research.

Our benchmark and evaluation framework can be found at:

<https://github.com/iSE-UET-VNU/CABench>

II. COMPOSABLE AI PROBLEM FORMULATION

Composable AI (CA) offers a scalable and efficient paradigm for solving complex, real-world AI tasks by decomposing them into sub-tasks and leveraging suitable available models to addresses each sub-task. To enable systematic study and principled evaluation under this setting, this paper formally defines the problem of CA as follows:

Definition 1: Composable AI Problem. Given:

- An AI task $\mathcal{T}$, characterized by its input space $\mathcal{X}$, output space $\mathcal{Y}$, and objective function $\mathcal{O}$; and
- A pool of ready-to-use models $\mathcal{M} = \{M_1, \dots, M_n\}$, where each model is specialized by its input/output modalities and functionalities

The goal of composable AI is to:

- (i) **Decompose** $\mathcal{T}$ into a set of sub-tasks $\{t_1, \dots, t_k\}$, where each sub-task t_j ($1 \leq j \leq k$) is associated with a well-defined objective o_j that contributes to the overall objective $\mathcal{O}$;
- (ii) **Select** appropriate models from the pool $\mathcal{M}$ to solve each sub-task; and
- (iii) **Compose** the selected models into a *coherent executable pipeline* that, given an input $x \in \mathcal{X}$, produces an output $y \in \mathcal{Y}$ that effectively solves the original task $\mathcal{T}$.

Following the above formulation, CA solves a complex task by constructing an executable pipeline that connects selected models, each responsible for a decomposed sub-task. However, in practice, the models may exhibit incompatibilities in their input/output modalities or data formats, making direct integration difficult. To address these interoperability challenges and enable seamless orchestration, *glue code* is often necessary in the pipeline. Glue code refers to (lightweight) modules that perform operations such as data pre/post-processing, format transformation, and the integration of heterogeneous outputs. It ensures consistent data modality and interface compatibility across models, allowing independently developed models to function cohesively within the same pipeline. Formally, we define a CA solution as follows:

Definition 2: Composable AI Solution. A composable AI solution is an executable pipeline defined as a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where:

- Each node $v \in \mathcal{V}$ represents a computational unit which can be either a selected model invocation or a glue code module; and
- Each edge $e \in \mathcal{E}$ ($\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$) represents the data flow between computational units, ensuring the output of one node is correctly routed as input to another.

Fig. 1 illustrates an example of a composable AI task and its corresponding solution. Given an audio file and an image file, the task is to determine whether the claim made in the audio is *supported*, *refuted*, or *inconclusive* based on textual evidence in the image. To solve this task using the available models from the pool, the proposed solution leverages three models: (1) a *Speech2Text* model (openai/whisper-large-v3) to transcribe the spoken claim into text, (2) a *TextExtraction* model (microsoft/trocr-base-printed) to extract text from the image using OCR, and (3) a *SimilarityMeasurement* model (BAII/bge-m3) to compute the semantic similarity between the transcribed claim and the extracted evidence. To enable accurate similarity measurement, both the transcribed audio and the OCR-extracted text should be cleaned, e.g., by standardizing formats and removing noise. These preprocessing operations are performed by dedicated glue code modules. Additionally, the similarity measurement

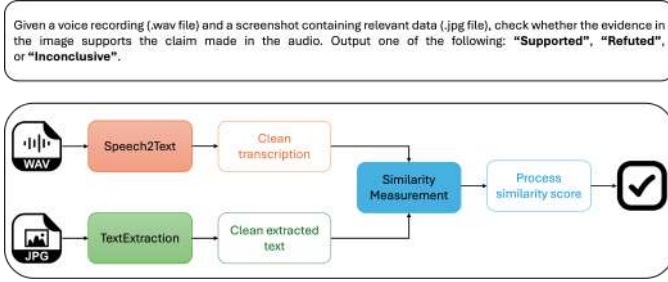


Fig. 1. A CA task and its corresponding solution

model BAAI/bge-m3 outputs a numerical similarity score, which must be interpreted and mapped to one of the final verdicts: *supported*, *refuted*, or *inconclusive*. This interpretation step is also handled by a glue code snippet. In summary, the solution consists of six computational units: three model invocations and three glue code modules, collectively forming an executable pipeline to solve the given complex task.

III. CABENCH: A COMPOSABLE AI BENCHMARK

To enable systematic investigation and principled evaluation of composable AI, we introduce CABENCH, a benchmark specifically designed to reflect the challenges and opportunities of building end-to-end compositional AI solutions. Unlike prior benchmarks such as HuggingGPT [6], which primarily focus on evaluating the correctness of task decomposition and model selection, CABENCH goes further by assessing whether the selected models can be correctly composed and executed as a *coherent pipeline* to ultimately fulfill the original objective of the given task. To the best of our knowledge, CABENCH is the first public benchmark that supports the evaluation of fully executable CA solutions.

The construction of CABENCH is guided by several key principles to ensure its utility and relevance for advancing CA. First, it prioritizes *realism* by designing tasks that reflect practical problems grounded in real-world application scenarios. Second, it ensures *decomposability* by formulating tasks that can be naturally broken down into meaningful sub-tasks aligned with the overall objective. Third, it guarantees *solvability* by ensuring each task can be solved using appropriate models from a defined model pool, with glue code introduced when necessary to enable seamless integration. Fourth, it promotes *diversity* by covering a wide range of domains and modalities, ensuring generalization across application areas. Finally, it supports *evaluability* by equipping each task with well-defined input-output specifications, ground-truth outputs, and standardized evaluation metrics, enabling objective, reproducible, and execution-based assessment of CA methods.

A. Benchmark Construction

1) *Task Construction*: Following the above mentioned principles, we construct a benchmark of composable AI tasks designed to reflect the complexity and practical demands of real-world applications. To guarantee *task realism*, each task $\mathcal{T}$ is

carefully derived from a popular dataset on Hugging Face² and Kaggle³. Dataset popularity is quantified based on download counts with a threshold of at least 10,000 downloads to ensure broad relevance and practical applicability. For each selected dataset, we prompt GPT-4o to generate a plausible user story or real-world application scenario that aligns with the dataset’s content and structure. These generated descriptions are then manually reviewed, refined, and validated to ensure clarity, feasibility, and alignment with realistic use cases. Through this process, we construct a total of 70 *realistic composable AI tasks*, each grounded a widely-used dataset and framed within meaningful application contexts.

Formally, each CA task $\mathcal{T}$ in CABENCH is represented as $\mathcal{T} = (\mathcal{D}, \mathcal{Q})$ where $\mathcal{D}$ is the *task description*, including an application scenario/user-story, a specification of input space $\mathcal{X}$, output space $\mathcal{Y}$, and an objective function $\mathcal{O}$ that defines the intended transformation from inputs to outputs. Meanwhile, $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^N$ ($x_i \in \mathcal{X}, y_i \in \mathcal{Y}$) denotes the *query set* consisting of N test with their corresponding ground-truth outputs. This set is partitioned into a validation set $\mathcal{Q}_V$ and a test set $\mathcal{Q}_T$, used respectively for validating and evaluating CA solutions. The queries in $\mathcal{Q}$ are randomly selected from the original dataset from which the task is derived.

2) *Model Pool Construction*: To solve the tasks in CABENCH, one would typically decompose each task into sub-tasks and select appropriate models from open ML model repositories such as Hugging Face, Kaggle, or the ONNX Model Zoo. However, these repositories are large, continuously evolving, and lack standardized configurations, making them unsuitable for consistent and reproducible evaluation in a benchmarking setting. For ease of comparison, we construct a fixed model pool tailored for CABENCH.

The model pool is curated according to several constraints to ensure quality, diversity, and practical usability. Specifically, we prioritize models that are widely adopted and well-documented on Hugging Face model hub, using these factors as signals of quality. To maintain diversity, we select models with different architectures, trained on varied datasets across multiple domains. Moreover, to accommodate hardware limitations and realistic deployment settings, we restrict the model size to a maximum of 500M parameters.

Overall, our model pool consists of about 700 ready-to-use models covering a wide spectrum of modalities and task types. Fig. 2 presents the distribution of models across modalities and types within the model pool of CABENCH. This curated pool ensures that every task in CABENCH is solvable using the available models while allowing meaningful evaluation of composition strategies under practical constraints.

B. Human-designed Reference Solutions

To demonstrate that each task in CABENCH is decomposable and solvable using the curated model pool, we construct reference solutions for all benchmark tasks. Each task is

²<https://huggingface.co/datasets>

³<https://www.kaggle.com/datasets>

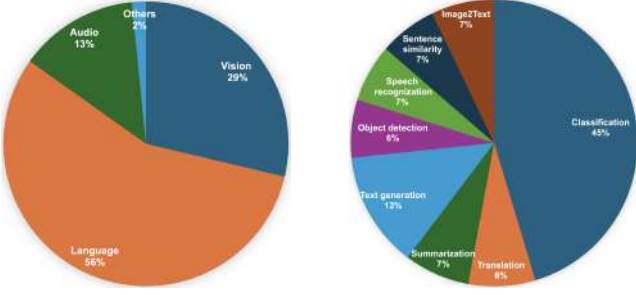


Fig. 2. ML models in CABENCH’s model pool by modalities and types

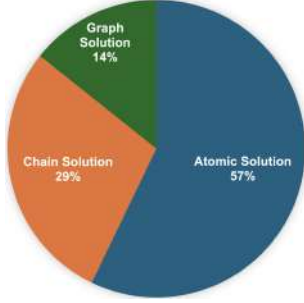


Fig. 3. Categorical distribution of human-designed reference solutions

independently addressed by the first four authors, who manually design decomposition strategies, select appropriate models from the pool, and implement necessary glue code. The initial solutions are then collaboratively validated and refined to ensure that the decomposition aligns with the task’s objective and that the composed pipeline yields correct outputs. Each reference solution is implemented as an execution pipeline, as defined in Definition 2.

To facilitate systematic analysis, we categorize the solutions based on their structural complexity. Specifically, solutions are grouped into three categories: *Atomic Solution*, *Chain Solution*, and *Graph Solution*.

Definition 3: Atomic Solution. A solution is an atomic solution if its execution pipeline, represented by a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consist of a *single node* with no edges, $|\mathcal{V}| = 1$ and $\mathcal{E} = 0$.

Definition 4: Chain Solution. A solution is a chain solution if its execution pipeline forms a *linear chain*, satisfying $|\mathcal{V}| = k \geq 2$, $|\mathcal{E}| = k - 1$, and there exists an ordering of the nodes $\mathcal{V} = \{v_1, \dots, v_k\}$ such that: $\mathcal{E} = \{(v_i, v_{i+1}) | 1 \leq i < k\}$

Definition 5: Graph Solution. A solution is a graph solution if its execution pipeline forms a *general DAG* with at least one node having multiple incoming/outgoing edges; formally, $|\mathcal{V}| \geq 2$ and $\exists v \in \mathcal{V}$ such that $out(v) > 1$ or $in(v) > 1$, where:

- $out(v) = |\{u \in \mathcal{V} | (v, u) \in \mathcal{E}\}|$ denotes the out-degree of node v ,
- $in(v) = |\{u \in \mathcal{V} | (u, v) \in \mathcal{E}\}|$ denotes the in-degree of node v .

Fig. 3 shows the distribution of human-designed reference solutions across the three structural categories. It is important

to note that a task may admit multiple valid solutions with different pipeline structures depending on the chosen decomposition strategies and model selections.

C. Evaluation Framework

In addition to the benchmark, we introduce an evaluation framework for automatically assessing the quality of CA solutions within CABENCH. The framework takes as input a task $\mathcal{T} = (\mathcal{D}, \mathcal{Q})$ and a corresponding solution $\mathcal{G}$, which may be either manually designed or automatically generated by a CA technique. The framework then executes $\mathcal{G}$ to perform inference on the provided query set $\mathcal{Q}$, producing predicted outputs for each query instance.

To evaluate the performance of $\mathcal{G}$ in solving $\mathcal{T}$, the predicted outputs are compared against the ground-truth outputs in $\mathcal{Q}$ using task-appropriate standard metrics. For example, Accuracy and F1-Score are used for classification tasks, R^2 -Score for regression tasks, and BLEU and ROUGE-L for generation tasks. Moreover, to enable consistent comparisons across tasks and methods, all metric scores are linearly normalized to the range $[0, 1]$, with higher scores indicating better performance. This normalization allows for cross-task aggregation while preserving relative performance rankings. Overall, this evaluation framework supports objective and automated assessment of CA solutions in CABENCH, providing a solid foundation for benchmarking future methods.

IV. EXPERIMENTAL RESULTS

A. Baselines

In this work, we leverage the capabilities of LLMs to automatically solve the benchmarking tasks in CABENCH, providing initial baselines alongside *Human-designed Reference Solution*. We evaluate the LLM under two strategies: *Prompt-to-Solve* and *Prompt-to-Pipeline*. Given a task $\mathcal{T}$ with description $\mathcal{D}$ and test set $\mathcal{Q}_T$:

- **Prompt-to-Solve** directly queries the LLM with the task description $\mathcal{D}$ and each query instance in $\mathcal{Q}_T$. The LLM is expected to *directly generate final output* for each query.
- **Prompt-to-Pipeline** instructs the LLM to approach the task in a *composable AI fashion* by explicitly prompting it to (i) decompose the task into sub-tasks, (ii) select suitable models from the provided model pool, and (iii) compose these models and glue code. The LLM is expected to *generate a complete executable pipeline*. This pipeline is then executed to produce the final outputs for the queries in $\mathcal{Q}_T$.

For both strategies, we experiment with three prompting patterns, including *Zero-shot*, *Few-shot*, and *Chain-of-Thought (CoT)* [7]. The prompt templates used for each strategy and prompting pattern are available on our website [8].

B. Experimental Procedure and Setup

In this work, all our experiments are conducted using GPT-4o-mini with fixed decoding parameters. To reduce the randomness, the temperature is set to 0. Experiments are conducted on machines equipped with 8-core CPUs, 32GB

RAM, and NVIDIA T4 GPUs. Due to the prompt length limitation (130,000 tokens), it is infeasible to include the full descriptions of all models in the entire model pool as context for the *Prompt-to-Pipeline* approach. To address this, we truncate the context appropriately. In these cases, we randomly shuffle the order of the model pool, repeat the experiments five times, and report the averaged results to reduce potential bias introduced by model ordering.

C. Experimental Results

1) *Performance of Baselines Across Task Types*: The performance of two automated approaches, i.e., *Prompt-to-Solve* and *Prompt-to-Pipeline*, and the manual approach, *Human-designed Reference Solution* across different task types is shown in Table I. For each task, the performance of the approaches is measured using the appropriate standard metrics and then linearly normalized to the range $[0, 1]$, where higher values indicates better performance.

For almost all task types, *Prompt-to-Solve* demonstrates superior performance over *Prompt-to-Pipeline*. For example, in translation tasks, *Prompt-to-Solve* achieves an average score of 0.74, which is more than three times higher than that of *Prompt-to-Pipeline*. Notably, *Prompt-to-Pipeline* fails to solve several task types such as regression and question answering, i.e., yielding scores of zero, whereas *Prompt-to-Solve* successfully handles these tasks, achieving normalized scores ranging from 0.15 to 0.50.

These results suggest that LLMs are more effective at generating direct answers than orchestrating multi-model pipelines from scratch. The advantage arises from *Prompt-to-Solve*'s ability to leverage the LLM's extensive pre-trained knowledge to directly produce answers for given input queries, allowing it to effectively handle many tasks without explicitly decomposition. Meanwhile, *Prompt-to-Pipeline* requires the LLM to not only understand the task structure but also select appropriate models and compose them into a valid execution pipeline, which is an inherently challenging process. Moreover, the prompt length limitation also prevents the LLM from accessing the full context of the model pool during pipeline generation, further hindering the effectiveness of *Prompt-to-Pipeline*.

In addition, *Prompt-to-Solve* shows strong performance in typical natural language processing tasks such as summarization and translation, yet it completely fails on structured perception tasks, like speech recognition and sentence similarity. Particularly, it achieves an average score of up to 0.74 in translation tasks, but yields a score of 0 across all settings for speech recognition tasks. Although the employed LLM, GPT-4o-mini, is a general-purpose LLM with multi-modal capabilities, it remains limited in specialized tasks. In particular, it is not explicitly fine-tuned for automatic speech recognition. Successfully solving these tasks often requires integrating multiple models and incorporating pre/post-processing steps, which are not equipped in *Prompt-to-Solve* to handle under a direct prompting strategy. As a result, LLM-based approaches' performance on such tasks remains poor.

The manually constructed *Human-designed Reference Solution* outperforms both LLM-based strategies across all task types, with average performance that is 90% higher than *Prompt-to-Solve* and 6.7 times higher than *Prompt-to-Pipeline*. The gap is particularly striking in tasks like speech recognition and sentence similarity, where *Human-designed Reference Solution* achieves an average score of 0.95, while both *Prompt-to-Solve* and *Prompt-to-Pipeline* performed extremely poorly on these tasks. *These findings highlight the significant performance gap between human-designed and automated CA solutions and emphasize the need for more capable and robust automatic composition methods.*

2) *Performance of Baselines across Solution Complexities*: Table II reports the performance of the approaches on tasks categorized by the structural complexity of their corresponding *Human-designed Reference Solutions*. Overall, *Prompt-to-Solve* consistently outperforms *Prompt-to-Pipeline* by a significant margin across all levels of complexity. Specifically, *Prompt-to-Solve* achieves an average score of 0.53, which more than three times the performance of *Prompt-to-Pipeline*. This demonstrates the effectiveness of directly leveraging the LLM's knowledge for end-to-end task resolution.

However, despite its strong performance, *Prompt-to-Solve* still lags behind the *Human-designed Reference Solution*. On average, *Human-designed Reference Solution* outperforms *Prompt-to-Solve* by 51%, with the performance gap is up to 85% on highly complex tasks such as those requiring graph-structured pipelines. This highlights that while LLMs can address these tasks to some extent, specialized models orchestrated through carefully designed pipelines remain substantially more effective, particularly for tasks requiring compositional reasoning and structured execution. *These results illustrate the promise of composable AI for tackling complex real-world problems and emphasize the need to develop methods for unlocking the full potential of composable AI systems by automatically generating effective pipelines.*

V. RELATED WORK

The rapid advancement of AI in recent years has been largely driven by the emergence of LLMs [9]–[11]. These models, such as GPT-3, GPT-4, PaLM, and LLaMA, have demonstrated remarkable generalization capabilities across a wide range of tasks, from text generation [12] and summarization [13] to mathematical reasoning [14], commonsense inference [15], and code generation [16]. However, they still often struggle with specialized tasks where domain expertise is required. To bridge this gap, Liang *et al.* [5] introduce TaskMatrix.AI which is an AI ecosystem that connects LLMs with millions of APIs. Given a user request, TaskMatrix.AI analyzes the task and invokes suitable APIs to fulfill it. Similarly, Toolformer [17] teaches LLMs to autonomously decide when and how to call external tools via APIs to enhance their performance on complex tasks. Different from these approaches, CABENCH not only focuses on selecting and invoking correct sequences of APIs, but also emphasizes the composition of these components into coherent and executable

TABLE I
PERFORMANCE OF BASELINES ACROSS TASK TYPES

Approach		Classification	Regression	Question Answering	Summarization	Translation	Object Detection	Generation	Image2Text	Speech Recognition	Sentence Similarity
Prompt-to-Solve	Zero-shot	0.54	0.15	0.51	0.52	0.74	0.48	0.57	0.45	0.00	0.00
	Few-shot	0.55	0.15	0.51	0.53	0.75	0.49	0.55	0.45	0.00	0.00
	CoT	0.54	0.15	0.47	0.51	0.74	0.44	0.50	0.45	0.00	0.00
Prompt-to-Pipeline	Zero-shot	0.09	0.00	0.00	0.06	0.03	0.16	0.23	0.05	0.03	0.05
	Few-shot	0.17	0.00	0.00	0.06	0.40	0.32	0.25	0.05	0.07	0.10
	CoT	0.23	0.00	0.00	0.05	0.23	0.38	0.25	0.06	0.02	0.03
Human-designed reference		0.80	0.85	0.85	0.49	0.76	0.83	0.29	0.67	0.95	0.95

TABLE II
PERFORMANCE OF BASELINES ACROSS SOLUTION COMPLEXITIES

Approach		Atomic solution	Chain solution	Graph solution	Average
Prompt-to-Solve	Zero-shot	0.51	0.64	0.38	0.53
	Few-shot	0.52	0.65	0.40	0.54
	CoT	0.49	0.66	0.39	0.52
Prompt-to-Pipeline	Zero-shot	0.09	0.09	0.07	0.09
	Few-shot	0.19	0.28	0.10	0.20
	CoT	0.20	0.36	0.09	0.23
Human-designed reference		0.82	0.79	0.72	0.80

pipelines. This includes managing interoperability, handling data transformations through glue code, and ensuring that the composed system can solve the original task end-to-end.

Our work is also closely related to HuggingGPT [6], which leverages LLMs as a controller to interpret user instructions, plan solutions, and delegate sub-tasks to pre-trained models hosted on the Hugging Face platform. While HuggingGPT explores task decomposition and model selection, it does not fully address the challenge of integrating selected models into executable pipelines. In contrast, CABENCH provides a benchmark that explicitly evaluates both the correctness of task decomposition and the feasibility of executing composed solutions, highlighting practical constraints such as model compatibility and execution consistency.

VI. CONCLUSION

In this work, we take an initial step toward advancing the study of composable AI, a promising paradigm for solving complex tasks by decomposing them into manageable sub-tasks and leveraging existing well-trained models. To enable rigorous and reproducible research in this domain, we introduce CABENCH, the first public benchmark designed specifically for composable AI. We also develop a full-pipeline evaluation framework that supports end-to-end performance assessment. Our experiments with LLM-based baselines highlight both the potential and current limitations of automatic methods in tackling composable AI tasks, especially for tasks requiring compositional reasoning and structured execution.

ACKNOWLEDGEMENT

This research is supported by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.03-2023.14.

REFERENCES

- [1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, no. 1, p. 7068349, 2018.
- [2] T.-H. Nguyen, T.-G. Vuong, H.-N. Duong, S. Nguyen, H. D. Vo, T. Aoki, and T.-T. Nguyen, "Generating critical scenarios for testing automated driving systems," *arXiv preprint arXiv:2412.02574*, 2024.
- [3] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.
- [4] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications: A survey," *Applied soft computing*, vol. 93, p. 106384, 2020.
- [5] Y. Liang, C. Wu, T. Song, W. Wu, Y. Xia, Y. Liu, Y. Ou, S. Lu, L. Ji, S. Mao *et al.*, "Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis," *Intelligent Computing*, vol. 3, p. 0063, 2024.
- [6] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *Advances in Neural Information Processing Systems*, pp. 38 154–38 180, 2023.
- [7] K. Stechly, K. Valmeekam, and S. Kambhampati, "Chain of thought-lessness? an analysis of cot in planning," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [8] T.-T. Pham, T.-H. Nguyen, D.-Q. Luong, M.-Q. Duong, T.-T. Nguyen, S. Nguyen, and D. H. Vo, "CABENCH: Benchmarking Composable AI for Solving Complex Tasks through Composing Ready-to-Use Models," 2025. [Online]. Available: <https://github.com/iSE-UET-VNU/CABench>
- [9] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang, and W. Zhang, "History, development, and principles of large language models: an introductory survey," *AI and Ethics*, vol. 5, no. 3, pp. 1955–1971, 2025.
- [10] Y. Annapaka and P. Pakray, "Large language models: a survey of their development, capabilities, and applications," *Knowledge and Information Systems*, vol. 67, no. 3, pp. 2967–3022, 2025.
- [11] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.
- [12] L. Qu, S. Wu, H. Fei, L. Nie, and T.-S. Chua, "Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 643–654.
- [13] H. Zhang, P. S. Yu, and J. Zhang, "A systematic survey of text summarization: From statistical methods to large language models," *ACM Computing Surveys*, vol. 57, no. 11, pp. 1–41, 2025.
- [14] Y. Sheng, L. Li, and D. D. Zeng, "Learning theorem rationale for improving the mathematical reasoning capability of large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 15 151–15 159.
- [15] J. Dong, J. Zhu, D. Liu, X. Qu, C. Bao, Z. Han, J. Zhu, and X. Wang, "Open-world fine-grained fashion retrieval with llm-based commonsense knowledge infusion," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025, pp. 223–232.
- [16] T.-D. Bui, D.-T. Luu-Van, T.-P. Nguyen, T.-T. Nguyen, S. Nguyen, and H. D. Vo, "Rambo: Enhancing rag-based repository-level method body completion," *arXiv preprint arXiv:2409.15204*, 2024.
- [17] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68 539–68 551, 2023.

An Experimental Study of Linear Attention in Multi-Agent Transformer

Xuan Duy Ta

VNU University of Engineering and Technology
Hanoi, Vietnam

Vu Minh Le

VNU University of Engineering and Technology
Hanoi, Vietnam

Viet Cuong Ta*

VNU University of Engineering and Technology
Hanoi, Vietnam
cuongtv@vnu.edu.vn

Abstract—Transformers have shown great promise in Multi-Agent Reinforcement Learning through architectures like the Multi-Agent Transformer (MAT), which leverages the self-attention mechanism to model complex inter-agent interactions. However, the quadratic complexity of Softmax attention hinders scalability in environments with large number of agents or long episodes. To reduce the complexity in standard MAT, we propose to integrate the Linear attention into the standard MAT instead of the Softmax attention. The Linear attention is able to reduce the quadratic complexity in large number of agents thanks to having linear complexity via a kernel based operator. Moreover, the proposed linear operator is expected to better adapt to the dynamics in multi-agent interactions than the Softmax implementation. Through extensive experiments on benchmark MARL environments including MPE, MA-MuJoCo, and SMACLite, we show that the Linear attention architecture achieves competitive performance to its Softmax counterpart. Our findings suggest that Linear attention is a viable drop-in alternative for efficient sequence modeling in MARL, laying the groundwork for future progress in lightweight MARL architectures.

I. INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) has long been recognized as a powerful framework for tackling complex decision-making problems involving multiple interacting agents. It underpins a wide range of real-world applications, including autonomous driving in mixed-traffic environments, cooperative robotic control, distributed energy management, and other domains that require awareness and reasoning about the behavior of multiple intelligent actors [1]. In a fully cooperative setting, one of the main challenge is the credit assignment problem, which involves decomposing the global advantage into agent-level components. Several popular algorithms has been proposed to solve the credit assignment, including value-based [2], [3] and policy gradient-based methods [4]. On the other hand, credit assignment can be tackled by viewing the learning agents' action as a sequence modeling problem [5], [6].

Transformers, proposed by Vaswani et al. [7] originally as a sequence-to-sequence model for neural-machine translation

tasks in Natural Language Processing, have since become foundational in a variety of domains due to their flexible attention mechanism and impressive representational capacity. At the core of the Transformer is the self-attention mechanism, which enables the model to dynamically integrate information across all elements of an input, allowing it to capture rich contextual relationships. They are now the go-to architecture in language [8], [9], vision [10], and audio [11] tasks, with growing adoption across additional domains. The increasing success of Transformer models has inspired their adaptation to MARL, notably through the Multi-Agent Transformer (MAT) [6] architecture. MAT casts MARL settings into sequence-modeling problems, in which the objective is to map agents' observation sequences to their optimal action sequences. It then leverages the Transformer's attention module, allowing for fine-grained modeling of observation-agent and agent-agent interactions, resulting in enhanced coordination and superior learning outcomes. However, MAT also inherits the quadratic $\Theta(n^2)$ time and memory complexity of Softmax attention with respect to the number of agents n . In MARL, this quadratic cost becomes a critical bottleneck - especially in practical scenarios requiring long rollouts, deep networks, or large agent populations.

To address this, we can look at prior work in across various tasks exploring lightweight and computationally efficient approximations to attention mechanisms. These efforts aim to alleviate the quadratic time and memory bottlenecks of standard Softmax-based attention, particularly when applied to long sequences. Prominent examples include $\Theta(n\sqrt{n})$ complexity via sparse factorizations of the attention matrix [12] and $\Theta(n \log n)$ by leveraging Real Fast Fourier Transform to accelerate computations [13]. Among these, Linear attention [14] mechanisms have emerged as a particularly attractive alternative, offering $\Theta(n)$ complexity while preserving much of the representational power of standard Softmax attention. Their linear scalability makes them especially promising for large-scale sequence modeling and real-time applications, where computational overhead is a key concern.

This raises a central question:

*Corresponding author.

Can the canonical Softmax attention in MAT be replaced verbatim by a linear-time kernel, without sacrificing policy performance on MARL tasks?

In this work, we investigate this question through direct substitution of the Softmax attention in MAT [6] with a Linear attention kernel and evaluating its impact across a diverse set of MARL benchmarks. Our results demonstrate that this replacement not only enhances scalability but also delivers performance that is comparative to the original formulation across several challenging multi-agent tasks.

II. BACKGROUND AND RELATED WORKS

A. Multi-Agent Reinforcement Learning

Consider the Multi-Agent Reinforcement Learning (MARL) formulation of an n -agent Partially Observable MDP [15] defined by the tuple $\langle \mathcal{S}, \mathcal{O}, \mathcal{A}, P, r, \gamma, \omega \rangle$. Here, $\mathcal{S}$ and $\mathcal{A}$ are the joint state and action space, respectively. The state at timestep 0, $s_0 \in \mathcal{S}$, is drawn from distribution ω , and $\gamma \in [0, 1)$ is the discount factor. The environments transitions from any state $s \in \mathcal{S}$ with joint action $a \in \mathcal{A}$ to another state $s' \in \mathcal{S}$ following distribution P . Associated with each such transition is the scalar reward $r(s, a, s', i)$ for agent i . Across all timestep t 's, any agent i can only observe a partial view of the true joint state $o_t^i = \Omega(s_t, i) \in \mathcal{O}$.

Let $\tau^i = (o_0^i, a_0^i, r_0^i, o_1^i, a_1^i, r_1^i, \dots, o_{H-1}^i, a_{H-1}^i, r_{H-1}^i, o_H^i)$ be the trajectory of time horizon H induced by the i -th agent's policy π^i from any initial state s_0 , where $r_t^i = r(s_t, a_t, s_{t+1}, i)$. The discounted return $R(\tau^i)$ of the i -th agent over trajectory τ^i is given by $R(\tau^i) = \sum_{h=0}^{H-1} \gamma^h r_h^i$.

The objective is to learn the parameterized joint policy $\pi = \langle \pi^1, \pi^2, \dots, \pi^n \rangle \in \Pi$ that maximizes the expected discounted cumulative reward, solving the optimization problem:

$$\max_{\pi \in \Pi} \left\{ V(\pi) := \mathbb{E}_{\langle \tau^1, \dots, \tau^n \rangle \sim \pi} \left[\sum_{i=1}^n R(\tau^i) \right] \right\}.$$

We denote the ordered set $\{1, 2, \dots, n\}$ as $[n]$ and, for any $m \in [n]$, $i_{1:m} = \{i_1, \dots, i_m\}$ be an arbitrary disjoint ordered subset of m distinct agents. In fully cooperative settings, all agents receive a common reward $r_t = r(s_t, a_t, s_{t+1}) = r_t^i \forall i$. As such, to learn its specific policy π^i , each agent has to decide the contribution of its action $\pi^i(\cdot | o_t^i)$ to the global reward, which is known as the credit assignment problem.

The most popular training paradigm of cooperative MARL is Centralize Training and Decentralize Execution (CTDE) [16] with strong baselines such as QMIX [17], VDN [18], MAPPO [4] and HAPPO [19]. In CTDE settings, the main challenge lies in effectively decomposing the global advantage into agents' specific advantages. To address cooperative decision-making, MAT [6] models the cooperative behaviours as a sequence learning problem. Specifically, given the previous actions of an input order, MAT directly learns to predict the next action of the subsequent agents using the encoder-decoder architecture based on the Transformer. Through this design, MAT captures spacial dependencies and inter-agent coordination patterns more effectively. Empirical

results demonstrate that MAT outperforms both MAPPO and HAPPO in multiple complex benchmarks.

B. The Transformer models

General Transformer architecture. Given a sequence of n feature vectors with embedding dimensionality d , $x \in \mathbb{R}^{n \times d}$, a Transformer is a function $T : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times d}$ constructed by stacking L Transformer layers $T_1(\cdot), \dots, T_L(\cdot)$ where each layer T_l is defined as $T_l(x) = f_l(A_l(x) + x)$. Here, $f_l(\cdot)$ applies element-wise transformation, typically implemented as a two-layer feedforward network; while $A_l(\cdot)$ is the self-attention mechanism, operating across the sequence to enable interactions between different elements. Self-attention plays a critical role in the efficacy of Transformers as it is Transformer's only component that facilitates cross-sequence dependencies.

The self attention function $A_l(\cdot)$ computes, for every position, a weighted average of the feature representations of all other positions. Each weight represents a similarity score between the input feature representation x of each element. Any input sequence x is first projected to three matrices Q, K, V - its corresponding representations. Then the output for all positions, $A_l(x) = O$, is computed as follows:

$$Q = xW_Q, K = xW_K, V = xW_V, \\ O_i = \sum_{j=1}^n \frac{\text{Sim}(Q_i, K_j)}{\sum_{j=1}^n \text{Sim}(Q_i, K_j)} V_j, \quad (1)$$

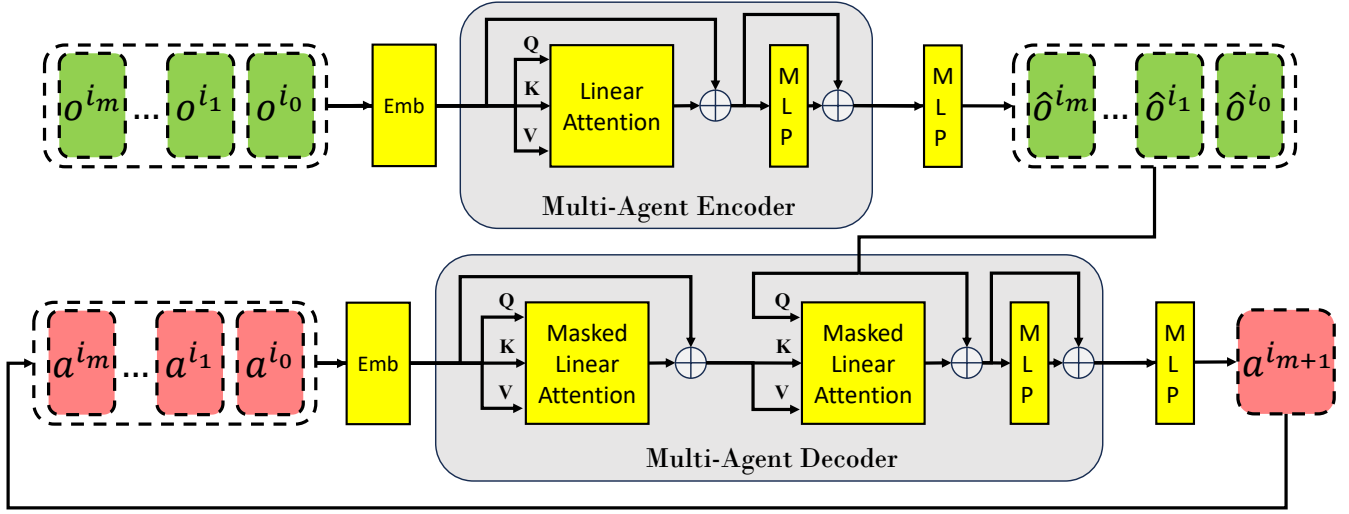
where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are projection matrices and $\text{Sim}(\cdot, \cdot)$ denotes the similarity function. Softmax attention [7] is the default choice, formulated as:

$$\text{Sim}(Q, K) = \exp \left(\frac{QK^T}{\sqrt{d}} \right). \quad (2)$$

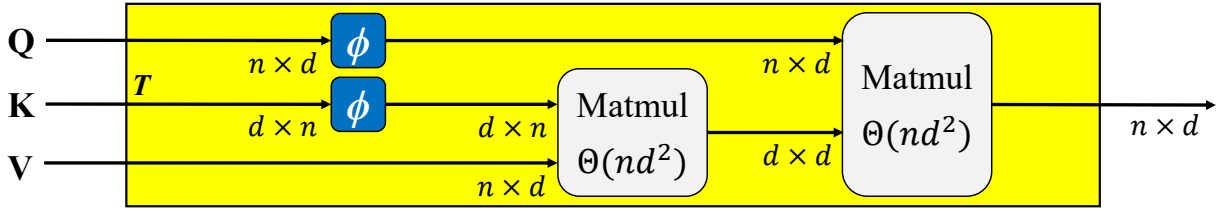
III. MULTI-AGENT TRANSFORMER WITH LINEAR ATTENTION

In this section, we propose our approach to extend the standard Multi-Agent Transformer (MAT) [6] with the Linear attention operator. The overall architecture is described in Fig. 1. Instead of learning to decompose the global reward in cooperative tasks, MAT frames the interactions between multiple agents as a sequence of actions and models it directly, implementing an encoder-decoder module for its value and policy functions. The encoder is designed as a mapping from the input observation space to an interaction-aware observational encoding space, which is later leveraged to train both value and policy components. Our MAT-linear approach replaces the standard self-attention in the encoder and decoder of MAT by the Linear attention (Fig. 1a). Following this design, our Linear attention operator adopt the work of [14] and is illustrated in Fig. 1b, given input Q, K and V matrices.

The encoder module: The encoder of MAT-linear is parameterized by ϕ , processes observation sequences $(o^{i_1}, \dots, o^{i_n})$, where $(i_1, \dots, i_m) \forall m$ denotes an arbitrary ordered subset of agents. Through multiple computational blocks, it produces observational encodings $(\hat{o}^{i_1}, \dots, \hat{o}^{i_n})$ that captures both agent states and their interactions. In the training phase, the encoder



(a) Multi-Agent Encoder-Decoder utilizing Linear attention for agent-wise sequence modeling



(b) Standalone Linear attention mechanism with annotated dimensionalities and computational costs
(n : number of agents, d : embedding size)

Fig. 1: Overall architecture of the Multi-Agent Transformer with Linear attention (MAT-linear).

is made to approximate the value functions by minimizing the empirical Bellman error:

$$L_{\text{Encoder}}^{\text{MAT}}(\phi) = \left[r_t + \gamma V_{\bar{\phi}}(\hat{o}_{t+1}^{i_m}) - V_{\phi}(\hat{o}_t^{i_m}) \right]^2, \quad (3)$$

where $\bar{\phi}$ is the target network's parameter.

The decoder module: The decoder is parameterized by θ and processes the embedded joint action sequence $\mathbf{a}^{i_0:m-1}$ through a series of decoding blocks, in which a^{i_0} representing a start-of-decoding symbol. Vitrally, each decoding block comes with 2 masked self-attention functions that enforce the sequential update scheme by computing attention only between the i_r^{th} and i_j^{th} heads where $r < j$. The first handles agent action interactions while the second computes the action-observation's attention. The decoder outputs a sequence of joint action probability distribution $\pi_{\theta}^{i_m}(a^{i_m} | \hat{o}_{t+1:n}^{i_1:n}, \mathbf{a}^{i_1:m-1})$ and is trained using a clipped PPO [20] objective:

$$L_{\text{Decoder}}^{\text{MAT}}(\theta) = -\min \left(\rho_t^{i_m}(\theta) \hat{A}_t, \text{clip}(\rho_t^{i_m}(\theta), 1 \pm \epsilon) \hat{A}_t \right),$$

$$\text{where } \rho_t^{i_m}(\theta) = \frac{\pi_{\theta}^{i_m}(a_t^{i_m} | \hat{o}_{t+1:n}^{i_1:n}, \hat{\mathbf{a}}_t^{i_1:m-1})}{\pi_{\theta_{\text{old}}}^{i_m}(a_t^{i_m} | \hat{o}_{t+1:n}^{i_1:n}, \hat{\mathbf{a}}_t^{i_1:m-1})}, \quad (4)$$

and $\hat{A}_t$ represents the estimated joint advantage function, applying generalized advantage estimation (GAE) [21] using $\hat{V}_t = \frac{1}{n} \sum_{m=1}^n V_{\phi}(\hat{o}_t^{i_m})$.

A critical implementation distinction exists between inference and training phases. During inference, actions are generated “auto-regressively”, with each action a^{i_m} being fed back into the decoder to produce $a^{i_{m+1}}$, beginning with a^{i_0} and stopping at $a^{i_{n-1}}$. In contrast, training leverages full parallelization by utilizing all stored actions $\mathbf{a}^{i_1:n-1}$ from the replay buffer to compute $\mathbf{a}^{i_1:n}$ simultaneously.

Linear attention: The standard MAT employs Softmax attention's for both the encoder and decoder module. However, Softmax attention requires computing pairwise similarities between all queries and keys, leading to $\Theta(n^2d)$ computational complexity. As a result, applying Softmax attention with a global receptive field incurs high computational costs as the number of agents increases. The Linear attention [14] can be used to address this computation challenge with a linear complexity of $\Theta(nd^2)$. Specifically, the similarity mapping between Q and K , $\text{Sim}(Q, K) = \phi(Q)\phi(K)^T$, allows changing the computation order from $(\phi(Q)\phi(K)^T)V$ to $\phi(Q)(\phi(K)^TV)$, resulting in the self-attention expression:

$$O_i = \frac{\phi(Q_i) \left(\sum_{j=1}^n \phi(K_j)^T V_j \right)}{\phi(Q_i) \left(\sum_{j=1}^n \phi(K_j)^T \right)}. \quad (5)$$

In the proposed MAT-linear, we employ the kernel function $\phi(x) = 1 + \text{elu}(x)$ for the linear operator, which is the popular choice [14]. While the introduction of Linear attention may

Algorithm 1 Single-step MAT-linear Training

Input: observation-action trajectories $\{\mathbf{o}_t, \mathbf{a}_t, r_t\}_{t=0}^T$ from replay buffer, current encoder and decoder parameters ϕ and θ , and old decoder parameters θ_{old}
for each $\mathbf{o}_t, \mathbf{a}_t, r_t$ in sampled trajectory **do**
 Encode $\mathbf{o}_t \rightarrow \hat{\mathbf{o}}_t^{i_{1:n}}$ using encoder ϕ
 Compute target values $r_t + \gamma V_{\phi}(\hat{\mathbf{o}}_{t+1}^{i_m}) - V_{\phi}(\hat{\mathbf{o}}_t^{i_m})$
 Compute $\hat{V}_t = \frac{1}{n} \sum_{m=1}^n V_{\phi}(\hat{\mathbf{o}}_t^{i_m})$
 Estimate advantages $\hat{A}_t$ using GAE [21]
end for
Update encoder:
 $\phi \leftarrow \phi - \eta_{\phi} \nabla_{\phi} L_{\text{Encoder}}^{\text{MAT}}(\phi)$ ▷ See eq. (3)
Update decoder:
 $\theta \leftarrow \theta - \eta_{\theta} \nabla_{\theta} L_{\text{Decoder}}^{\text{MAT}}(\theta)$ ▷ See eq. (4)
Output: Updated encoder, decoder parameters ϕ and θ

reduce the complexity with respect to the agents n , it can introduce unbounded gradients relative to the standard Softmax. This issue can be resolved in practical implementations with a small learning rate and gradient clipping mechanism. There are also other variants of the kernel function [22], [23], which can be considered as potential future works. The overall learning algorithm with MAT-linear is presented in Algorithm 1.

IV. EXPERIMENTS

A. Experiment Setup

The performance of MAT with Linear attention (MAT-linear) is evaluated and contrasted against the Softmax-attention MAT (MAT) and the Multi-Agent GRU (MA-GRU) - a baseline using the Gated Recurrent Unit (GRU) [24] for the sequence-to-sequence modeling task in the MAT framework. Comparison experiments are conducted on 3 canonical MARL benchmarks: Multi-Particle Environment (MPE) [25], Multi-Agent MuJoCo (MA-MuJoCo) [26] and the StarCraft II Multi-Agent Challenge Lite (SMACLite) [27].

In the MPE benchmark, built upon the grounded communication environment [28], where varying agents and landmarks inhabit a two-dimensional world with continuous space and discrete time, we trained teams of 3 and 5 agents on the staple *simple spread* task. The MA-MuJoCo suite features numerous multi-agent continuous control tasks where agents coordinate by controlling joints of biomimetic robots. Among them, the tasks and robot configurations: *Hopper-v2* (3×1), *HalfCheetah-v2* (6×1), *Ant-v2* (8×1) and *Humanoid-v2* (17×1) are utilized for evaluations. For SMACLite, experiments are carried out on three maps: *27m vs 30m*, *3s5z vs 3s6z* and *MMM2* - all categorized as Super Hard difficulty, representing the highest level of challenge in this benchmark.

We report episode rewards of the MPE and MA-MuJoCo environments; for SMACLite, we plot the win rate. Results are averaged over 3 seeds. To account for variability across tasks, we compute the normalized mean performance of each scenario at fixed evaluation intervals.

For all tasks in the MPE, MA-MuJoCo, and SMACLite environments, we use learning rates of 7×10^{-5} , 5×10^{-5} , and 5×10^{-4} , respectively. The PPO clipping ratio is set to $\epsilon = 0.05$ across all environments and tasks, except for the *27m vs 30m* scenario where $\epsilon = 0.2$. All models use the Adam optimizer with a latent embedding dimension of $d = 64$ and a discount factor of $\gamma = 0.99$. Transformer-based models consist of one layer and employ a single attention head.

B. Results

Main results are shown in Fig. 2. In general, the proposed MAT-linear outperform MAT and MA-GRU in the MA-MuJoCo benchmark while having comparable performances in the remaining two benchmarks.

Specifically, in MPE *simple spread* environments, the three models demonstrates M-shaped learning dynamics which can be characterized by repeated phases of progress, overfitting, and recovery. All models have the same learning speed with similar variances in the 3 and 5 agent settings. In the MA-MuJoCo benchmark, MAT-linear achieves the best performance in 3 of 4 testing environments. It has a lower performance than the original MAT in the *Humanoid-v2* (17×1) environment. One possible explanation is the linear operator in Linear attention is being affected by a large number of agents. The softmax operator can be viewed as a smoothing mechanism for extreme gradient conditions. Without the softmax operator, the gradients in MAT-linear is more prone to gradient noise when increasing the number of agents. These noises make the learning of MAT-linear have greater instability [29]. However, this issue only happens when the training continues for a long time, which is near the region of 1M training steps. Meanwhile, the MAT has higher variance than our proposed MAT-linear in *Hopper-v2* (3×1) and *Humanoid-v2* (17×1) environments. In the SMACLite benchmark, the MAT-linear outperforms the standard MAT in 2 of the 3 environments. MAT-linear demonstrates a sharp increase in both the *27m vs 30m* and the *MMM2* environments. On the other hand, MAT has the best performance in *3s5z vs 3s6z* and reaches the winning rate of 1.0 in under 2M training steps.

Based on the previous results, we extend the testing configs of MAT and MAT-linear with two layers and single attention head for *simple spread* (5 agents) of the MPE benchmark and *Hopper-v2* (3×1) from the MA-MuJoCo suite. As illustrated in Fig. 3, behaviour of deeper models appear consistent with the single-layer findings. On *simple spread*, models again display the same characteristic M-shaped learning curves, indicating that this higher depth does not mitigate the environment's innate optimization difficulty. For *Hopper-v2* (3×1), Softmax-attention MAT initially achieves slightly higher returns but simultaneously suffers from extremely high variance. As training progresses and variance subsides, its average performance falls behind that of the Linear attention version. In contrast, MAT-linear demonstrates consistent, steady learning throughout. It has a slower start, but ultimately outperforming MAT, mirroring behaviors observed in single-layer experiments. The MAT-linear also attains a lower variance in these settings.

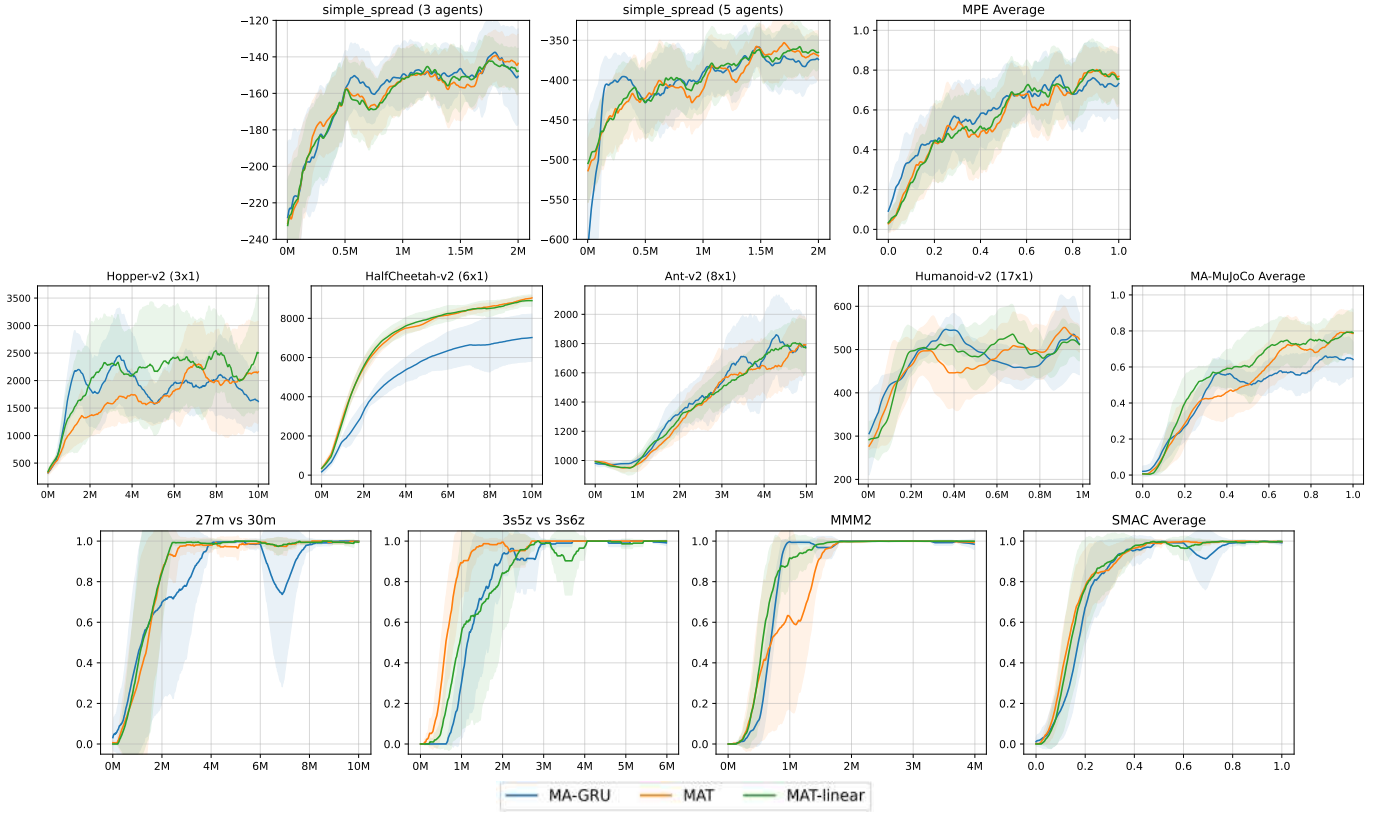


Fig. 2: Results on three benchmarks with different scenarios averaged over 3 seeds, showing the episode rewards for MPE and MA-MuJoCo scenarios and the win rate of SMAC tasks.

We further expand the proposed MAT-linear with different number of layers in *simple spread*. To validate the effects of varying agent numbers to MAT-linear, we iterate the *simple spread* setting with 3, 5 and 10 agents. The results are plotted in Fig. 4. Overall, performances have strong correlation with the number of layers. The best performance of MAT-linear can be achieved with 6 or 12 layers. Both 1 and 2 layer settings have large variance when the number of agents increases to 10. It can be explained that the number of layers plays an important role to effectively encode and decode the interplay between different agents. Therefore, if the interaction is too complex, a low number of MAT-linear layers is not sufficient to capture the cooperative behaviours of the agents.

V. CONCLUSION AND FUTURE WORK

In this work, we have proposed the MAT-linear, which employs Linear attention instead of the Softmax attention in cooperative MARL learning tasks. The Linear attention can reduce the quadratic complexity when the number of agents increases while simultaneously maintains solid approximations of the value and policy functions. We carry out experiments on three popular environments, in which MAT-linear has comparable or better performance than the normal MAT counterpart. Based on the results of this work, we plan to extend these analysis to verify other aspects of Linear attention in MAT, including the stability of gradients without the smoothing

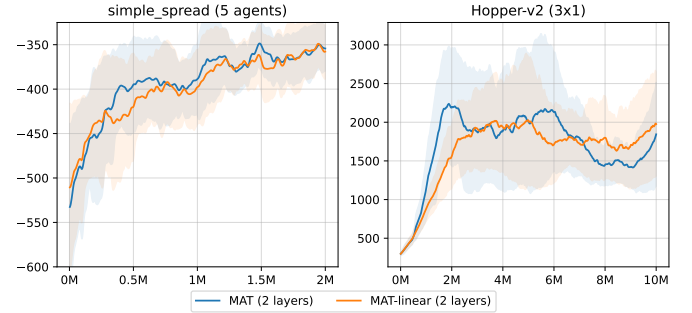


Fig. 3: Average rewards of 2-layer MAT with Softmax and Linear attention.

effects from Softmax and the models' learning capability with increasing number of agents.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-24-1-4012. We thank NT Duyen, LD Trong, TQ Long for helpful discussions.

REFERENCES

- [1] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 895–943, 2022.

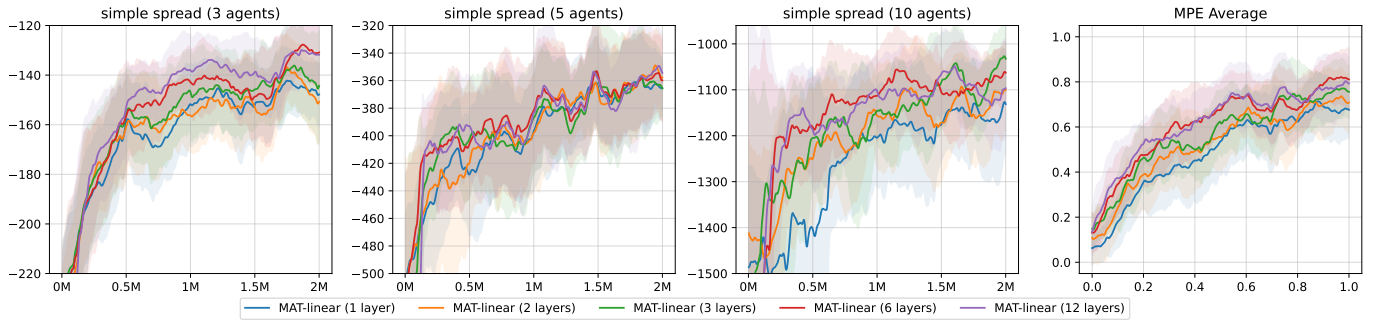


Fig. 4: Results on MPE with different number of layers of MAT-linear.

- [2] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, “Value-decomposition networks for cooperative multi-agent learning,” *arXiv preprint arXiv:1706.05296*, 2017.
- [3] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, “Monotonic value function factorisation for deep multi-agent reinforcement learning,” *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.
- [4] C. Yu, A. Velu, E. Vinitzky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, “The surprising effectiveness of ppo in cooperative multi-agent games,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 24611–24624, 2022.
- [5] M. Janner, Q. Li, and S. Levine, “Offline reinforcement learning as one big sequence modeling problem,” *Advances in neural information processing systems*, vol. 34, pp. 1273–1286, 2021.
- [6] M. Wen, J. Kuba, R. Lin, W. Zhang, Y. Wen, J. Wang, and Y. Yang, “Multi-agent reinforcement learning is a sequence modeling problem,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 509–16 521, 2022.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [8] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, “Mixtral of experts,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.04088>
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.13971>
- [10] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, “Image transformer,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 4055–4064. [Online]. Available: <https://proceedings.mlr.press/v80/parmar18a.html>
- [11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [12] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating long sequences with sparse transformers,” *ArXiv*, vol. abs/1904.10509, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:129945531>
- [13] J. Fein-Ashley, N. Gupta, R. Kannan, and V. Prasanna, “Spectre: An fit-based efficient drop-in replacement to self-attention for long contexts,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.18394>
- [14] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers are rnns: Fast autoregressive transformers with linear attention,” in *International conference on machine learning*. PMLR, 2020, pp. 5156–5165.
- [15] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*, ser. SpringerBriefs in Intelligent Systems. Springer International Publishing, 2016.
- [16] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.
- [17] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, “Monotonic value function factorisation for deep multi-agent reinforcement learning,” *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.
- [18] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, “Value-decomposition networks for cooperative multi-agent learning,” 2017. [Online]. Available: <https://arxiv.org/abs/1706.05296>
- [19] J. G. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, and Y. Yang, “Trust region policy optimisation in multi-agent reinforcement learning,” *arXiv preprint arXiv:2109.11251*, 2021.
- [20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017. [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [21] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” 2018. [Online]. Available: <https://arxiv.org/abs/1506.02438>
- [22] D. Han, Y. Pu, Z. Xia, Y. Han, X. Pan, X. Li, J. Lu, S. Song, and G. Huang, “Bridging the divide: Reconsidering softmax and linear attention,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 79 221–79 245, 2024.
- [23] K. Choromanski, V. Likhoshervstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller, “Rethinking attention with performers,” 2022. [Online]. Available: <https://arxiv.org/abs/2009.14794>
- [24] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” 2014. [Online]. Available: <https://arxiv.org/abs/1409.1259>
- [25] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Neural Information Processing Systems (NIPS)*, 2017.
- [26] C. S. de Witt, B. Peng, P.-A. Kamienny, P. Torr, W. Böhmer, and S. Whiteson, “Deep multi-agent reinforcement learning for decentralized continuous cooperative control,” *arXiv preprint arXiv:2003.06709*, vol. 19, 2020.
- [27] A. Michalski, F. Christianos, and S. V. Albrecht, “Smaclite: A lightweight environment for multi-agent reinforcement learning,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.05566>
- [28] I. Mordatch and P. Abbeel, “Emergence of grounded compositional language in multi-agent populations,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [29] C. Tamayo-Rousseau, Y. Zhao, Y. Zhang, and R. Balestriero, “Your attention matters: to improve model robustness to noise and spurious correlations,” *arXiv preprint arXiv:2507.20453*, 2025.

ABSTRACT-THEN-CONCRETIZE

Automated Design of Multi-Agent System

Trung-Hieu Nguyen, Huy-Thinh Tran, Quang-Hung Trinh, Tung-Thuy Pham,
Thu-Trang Nguyen, Son Nguyen*, and Hieu Dinh Vo

Faculty of Information Technology, VNU University of Engineering and Technology, Hanoi, Vietnam
{21020017, 23020162, 23021587, 25025009, trang.nguyen, sonnguyen, hieuvd}@vnu.edu.vn

Abstract—Recent advances in multi-agent systems (MAS) powered by Large Language Models (LLMs) have shown promising results in solving complex tasks. However, existing approaches either rely on handcrafted designs or generate MAS from scratch for each query, limiting scalability and efficiency. In this paper, we propose MASAC, a novel automated MAS design framework based on an ABSTRACT-THEN-CONCRETIZE paradigm. MASAC first abstracts common design features from existing MAS across a domain, then dynamically concretizes them to construct query-specific MAS for new inputs. Extensive experiments on widely used benchmarks demonstrate that MASAC not only outperforms state-of-the-art baselines in accuracy by up to 200%, but also significantly reduces generation time and cost. Our results highlight the potential of reusing design abstractions to enable scalable, high-quality MAS generation.

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated impressive performance across a wide range of natural language processing tasks [1]–[5]. However, certain complex problem domains, such as mathematical problem solving [6], domain-specific question answering [7], and software engineering tasks [8], remain challenging for standalone LLMs to solve effectively. To address these limitations, recent research has increasingly explored Multi-Agent Systems (MAS), where multiple LLM-based agents collaborate and communicate to jointly complete tasks [9]–[18].

Early MAS designs were predominantly handcrafted, with humans manually configuring components such as prompt templates, agent roles, and inter-agent communication protocols. While effective to a degree, these manual approaches hinder scalability and adaptability to diverse domains or dynamic task requirements. Therefore, there is a growing interest in automating the design and optimization of MAS to reduce reliance on human engineering and enable more generalizable and scalable solutions.

Recent efforts in automated MAS design can be broadly categorized into *task-level* and *query-level* approaches. *Task-level* methods aim to generate a MAS tailored for a specific task (e.g., code generation or math problem solving) and reuse the same MAS for all user queries within that task [9]–[15]. However, such approaches often adopt a one-size-fits-all strategy, lacking the flexibility to adapt to individual query requirements. This limits their effectiveness when handling

diverse or nuanced user inputs. To improve adaptability, *query-level* methods have been proposed to automatically generate a customized MAS for each individual query [16], [17]. While these approaches show promising results in various domains, they often overlook reusable design components that shared across tasks in the same domain and could potentially improve efficiency and performance. Moreover, since query-level methods typically generate MAS configurations from scratch for every query, the associated computational cost can be substantial, which restricts their applicability in real-time or time-sensitive scenarios.

In this paper, we introduce MASAC, a novel automated approach for multi-agent system design based on an ABSTRACT-THEN-CONCRETIZE paradigm, operating in two phases. In *Offline MAS Abstraction* phase, MASAC analyzes existing MAS designs across queries within the same domain to identify and abstract common design components that can potentially benefit the construction of new MAS instances. In *Online MAS Concretization* phase, given a new, unseen query, MASAC selects relevant abstracted components from the offline abstraction phase, adapts them to the specific query context, and synthesizes a complete, query-specific MAS. This process includes evaluating candidate configurations to ensure effectiveness.

We evaluate MASAC on widely adopted MAS benchmarks. The results demonstrate that the proposed ABSTRACT-THEN-CONCRETIZE approach is effective at leveraging reusable design patterns from prior MAS instances to construct new, high-performing systems. Compared to state-of-the-art query-level methods, MASAC achieves significantly higher task accuracy, improving performance by up to 200%, while simultaneously reducing MAS generation time and computational cost by up to 20% and 32%, respectively.

In brief, this paper makes the following contributions¹:

- 1) MASAC, a novel automated MAS design framework based on an ABSTRACT-THEN-CONCRETIZE paradigm, which identifies and reuses common MAS design components to generate high-quality, query-specific systems.
- 2) An extensive empirical evaluation showing that MASAC significantly outperforms existing task-level and query-level MAS approaches in both accuracy and efficiency.

¹MASAC’s implementation and results could be found on our website: github.com/iSE-UET-VNU/MASAC.

*Corresponding author.

II. LLM-BASED MULTI-AGENT SYSTEMS

This section defines key concepts used throughout the paper, including *Task*, *Query*, *LLM-based Agent*, *Operator*, and *LLM-based Multi-Agent System (MAS)*.

A *task* T denotes a domain-level problem category, such as mathematical problem solving, software engineering, or question answering, comprising a set of independent queries $\{q_1, q_2, \dots, q_n\}$ that share a common topical focus. Each query $q_i \in T$ is a standalone instance solvable in isolation, though reusable design strategies may emerge across queries within the same task. A *query* q is an individual user input drawn from a task T , representing a self-contained problem instance requiring reasoning or generation. For example, a query in a software debugging task may ask: “Identify and fix the error in the following Python function.”

An *LLM-based agent* a is a modular reasoning unit powered by a LLM, responsible for executing a specific subtask (e.g., planning, critiquing, or verifying). It is modeled as a function $a : \mathcal{I} \rightarrow \mathcal{O}$, where $\mathcal{I}$ and $\mathcal{O}$ are the spaces of inputs and outputs, respectively. An *operator* o is a composite unit comprising a set of agents $\{a_1, \dots, a_k\}$ organized under a shared reasoning pattern, e.g., *Chain-of-Thought (CoT)*, *Self-Consistency*, *Debate*, or *Reflexion*. An operator can be viewed as a higher-order function $o : \mathcal{I} \rightarrow \mathcal{O}$ that coordinates its agents to perform structured, multi-step inference.

An *LLM-based Multi-Agent System* $\mathcal{M}$ is a structured composition of operators and/or standalone agents designed to collaboratively solve a query q or a set of queries T . Formally, a MAS is defined as a tuple $\mathcal{M} = (O, F)$, where:

- $O = \{o_1, o_2, \dots, o_m\}$ is the set of operators.
- F is a unified flow graph over O encoding both:
 - *Control flow*: execution ordering, branching, and looping semantics among operators.
 - *Data flow*: directed edges indicating output-input dependencies between operators.

In practice, MASs are implemented as programs, with their structure and behavior encoded in source code. This code representation implicitly defines the operators, their configurations, and the flow graph F . In our work, we analyze these MAS programs to extract the underlying components (O, F) , enabling systematic understanding, comparison, and optimization of LLM-based multi-agent designs.

III. ABSTRACT-THEN-CONCRETIZE APPROACH FOR AUTOMATED DESIGN OF MULTI-AGENT SYSTEM

Figure 1 shows the ABSTRACT-THEN-CONCRETIZE framework for automated design of MAS in MASAC. Particularly, MASAC consists of two main phases: (1) *Offline MAS Abstraction* (Section III-A), where common MAS design components are extracted and abstracted from previously constructed MAS in the same domain, and (2) *Online MAS Concretization* (Section III-B), where these relevant abstracted components are selectively adapted and concretized to build an effective, query-specific MAS for a new query input. This two-phase process enables MASAC to reuse shared design knowledge while maintaining adaptability to individual queries.

A. Offline MAS Abstraction

In this phase, MASAC extracts a compact collection of reusable MAS templates, each capturing a common design components, from a corpus of existing MASs. Let $T = \{(M_i, q_i)\}_{i=1}^N$ be the set of N MAS-query pairs within a single task domain. We perform three main steps. First, *Stage Decomposition* decomposes MAS-query pairs into fine-grained reasoning stages. Then, *Stage Clustering* groups these stages based on structural and functional similarity. Finally, *Abstracted MAS Creation* constructs an abstract MAS by identifying and unifying shared reasoning patterns across clusters.

1) *Stage Decomposition*: Each MAS M_i is implemented in code (e.g., Python) which is analyzed together with the corresponding query to recover its structure and decompose it into interpretable reasoning stages.

Formally, each MAS M_i is represented as a directed graph (O_i, F_i) of operators and flows. We decompose M_i into a sequence of reasoning stages $S_i = \{s_{i1}, s_{i2}, \dots, s_{iL_i}\}$, where each stage $s_{ij} = (\varphi_{ij}, \iota_{ij}, O_{ij}, F_{ij})$ is defined by:

- φ_{ij} is a description of the goal for that stage (e.g., “generate candidate solutions,” “verify consistency”);
- $\iota_{ij} = (I_{ij}, O_{ij})$ is the stage interface with input set I_{ij} and output set O_{ij} consumed and produced by the stage;
- $O_{ij} \subseteq O_i$ is the subset of operators engaged in that stage;
- $F_{ij} \subseteq F_i$ is the subgraph of flow edges among O_{ij} .

After decomposing M_i into S_i , we construct an inter-stage flow graph F_i^{stage} over S_i by analyzing F_i to identify cross-stage dependencies. Specifically, if an output from an operator in stage s_{ij} is consumed by an operator in a subsequent stage s_{ik} ($j < k$), then we include a directed edge $(s_{ij} \rightarrow s_{ik})$ in F_i^{stage} , indicating a flow of data or control between stages.

2) *Stage Clustering*: Let $\mathcal{S} = \bigcup_{i=1}^N S_i$ be all stages. We cluster $\mathcal{S}$ into K clusters $\{C_1, \dots, C_K\}$ by maximizing intra-cluster similarity under the composite metric:

$$\text{sim}(s, s') = w_\varphi \text{sim}_\varphi(\varphi, \varphi') + w_\iota \text{sim}_\iota(\iota, \iota') + w_o \text{sim}_o(\mathcal{O}, \mathcal{O}') + w_f \text{sim}_f(\mathcal{F}, \mathcal{F}'), \quad (1)$$

where $w_\varphi + w_\iota + w_o + w_f = 1$. Specifically:

- $\text{sim}_\varphi(\varphi, \varphi') = \cos(\text{emb}(\varphi), \text{emb}(\varphi'))$, where emb returns embedding vector of corresponding string.
- $\text{sim}_\iota(\iota, \iota') = 1$ if $\iota = \iota'$, otherwise $\text{sim}_\iota(\iota, \iota') = 0$
- $\text{sim}_o(\mathcal{O}, \mathcal{O}') = \frac{|\mathcal{O} \cap \mathcal{O}'|}{|\mathcal{O} \cup \mathcal{O}'|}$ and $\text{sim}_f(\mathcal{F}, \mathcal{F}') = \frac{|\mathcal{F} \cap \mathcal{F}'|}{|\mathcal{F} \cup \mathcal{F}'|}$.

In this work, we set $w_\varphi = w_\iota = w_o = w_f$. We apply a standard clustering algorithm (e.g., k-means on the similarity-embedded stage vectors or DBSCAN on pairwise similarities) to partition $\mathcal{S}$. For each cluster C_k , we derive a representative stage $\hat{s}_k = (\hat{\varphi}_k, \hat{\iota}_k, \hat{\mathcal{O}}_k, \hat{\mathcal{F}}_k)$, where $\hat{\varphi}_k$ is the medoid objective, $\hat{\iota}_k$ identical interface, and $\hat{\mathcal{O}}_k, \hat{\mathcal{F}}_k$ contain the most frequent operators and flows.

3) *Abstracted MAS Creation*: Using the representative stages $\{\hat{s}_1, \dots, \hat{s}_K\}$, we transform each original MAS M_i into an abstracted MAS M'_i :

- 1) **Stage Replacement**. For each original stage $s_{ij} \in S_i$, let C_k be its assigned cluster. Replace s_{ij} with its

1. Offline MAS Abstraction

2. Online MAS Concretization

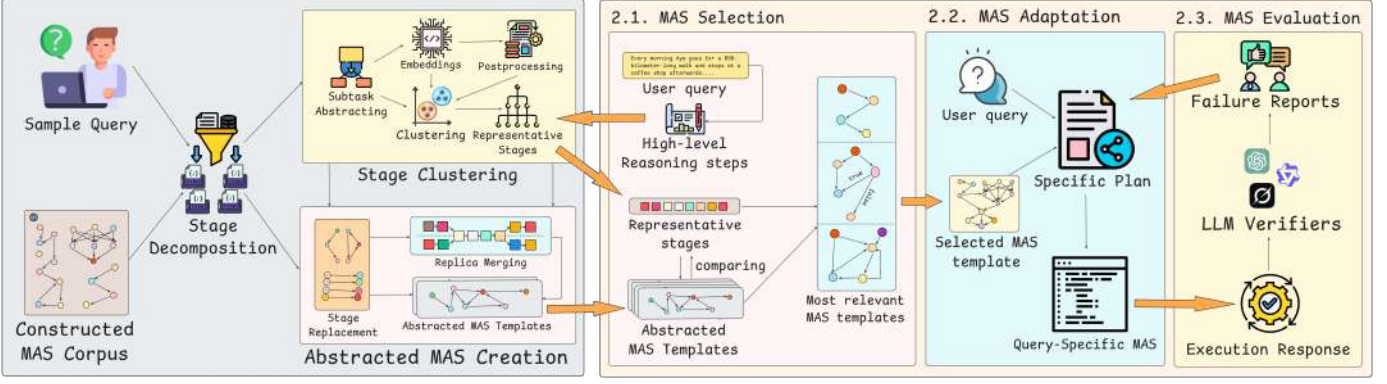


Fig. 1: ABSTRACT-THEN-CONCRETIZE paradigm in MASAC

corresponding cluster representative $\hat{s}_k$ in $\hat{S}_i$, preserving all inter-stage edges from F_i^{stage} between s_{ij} and others.

- 2) **Replica Merging.** Collect $\{M'_i\}_{i=1}^N$ and merge any two M'_i, M'_j whose MAS-level similarity exceeds a predefined threshold. The merged template records:

- *Mandatory components:* operators and flows common to all replicas.
- *Variational components:* operators or flows present only in a subset of replicas.

The final output is a set of $L \leq N$ abstracted MAS templates $\{\mathcal{B}_1, \dots, \mathcal{B}_L\}$, each encoding both control- and data-flow patterns along with their optional variants.

B. Online MAS Concretization

Given the set of abstracted MAS templates $\{\mathcal{B}_1, \dots, \mathcal{B}_L\}$ produced by the Offline MAS Abstraction phase, MASAC enters the *Online MAS Concretization* phase. Its objective is to instantiate a *query-specific MAS* $\mathcal{M}_q$ for a novel query q by leveraging and adapting reusable design components. The concretization proceeds in three stages: *MAS Selection*, *MAS Adaptation*, and *MAS Evaluation*.

1) *MAS Selection:* In the *MAS Selection* stage, the input query q is first analyzed and decomposed into a sequence of high-level reasoning steps $S_q = \{s_1^q, s_2^q, \dots, s_n^q\}$. Each step s_j^q is defined by a natural language description φ_j^q that expresses its reasoning objective. This decomposition is performed using the same methodology employed in the *Offline MAS Abstraction* phase. For each step s_j^q , MASAC measures the similarity between φ_j^q and the objective descriptions $\hat{\varphi}_k$ of the representative stages $\hat{s}_k$ obtained in *Stage Clustering*. Based on this similarity, the top- k most relevant representative stages are retrieved per step. Aggregating the matched stages across all steps, MASAC ranks all abstracted MAS templates $\mathcal{B}_1, \dots, \mathcal{B}_L$ and selects the most structurally and semantically relevant templates as candidates for instantiation.

2) *MAS Adaptation:* In the *MAS Adaptation* stage, MASAC constructs a query-specific MAS $\mathcal{M}_q$ by adapting one of the selected abstracted templates. Each query step s_j^q is mapped to a representative stage $\hat{s}_k$ present in the selected

template, and the corresponding operators and flow structures are instantiated. For each operator $o \in \hat{s}_k$, MASAC adapts its configuration, such as prompt templates, agent roles, and tool usage strategies, according to the content and reasoning needs of the input query q . After all relevant operators are instantiated and configured, they are assembled into a complete MAS $\mathcal{M}_q = (\mathcal{O}_q, \mathcal{F}_q)$, which preserves the control-flow and data-flow structure defined in the abstracted template while being concretely tailored to the query.

3) *MAS Evaluation:* MASAC then executes the generated MAS $\mathcal{M}_q$ on the input query. During execution, MASAC continuously monitors the system for potential failures, such as incorrect outputs, flawed reasoning chains, or execution errors. If any such issues arise, MASAC invokes a set of LLM-based verifiers and analysis agents to identify failure causes and generate a structured *failure report*. This report highlights misconfigured components, inconsistencies in reasoning, or invalid operator behavior. The failure report is fed back into the adaptation stage, enabling MASAC to revise and regenerate the MAS. This *adaptation-evaluation loop* continues iteratively until either (i) $\mathcal{M}_q$ produces valid and acceptable outputs, or (ii) a predefined resource limit (e.g., time, number of iterations, token usage, or monetary cost) is reached.

IV. EVALUATION METHODOLOGY

To assess the effectiveness and efficiency of the proposed approach, we formulate the following research questions:

- **RQ1: Effectiveness of MAS Generation.** How effective is MASAC in automatically generating MAS designs? How does its performance compare to state-of-the-art approaches in terms of task success and output quality?
- **RQ2: Efficiency and Computational Cost.** What are the computational resources, including inference time and monetary cost, required by MASAC to generate and execute a MAS for a given query? How does this cost compare to that of existing baselines?
- **RQ3: Model Configuration Analysis.** How sensitive is MASAC's performance to the choice of underlying large language models? Does it maintain competitive effectiveness when different LLMs are used in its pipeline?

A. Benchmarks and Metrics

Benchmarks. We evaluate MASAC on two benchmarks:

- **AIME24** [19]: A collection of high-difficulty math problems modeled after the American Invitational Mathematics Examination (AIME), designed to assess complex symbolic reasoning and logical deduction.
- **GPQA_diamond** [7]: A benchmark containing university-level, domain-specific science questions across disciplines such as biology, physics, and chemistry.

For *GPQA_diamond*, we split the data using a 3:1 ratio, allocating 75% for validation and 25% for testing, following standard practice for query-specific evaluation. Due to its small size, *AIME24* is used in its entirety for both validation and testing to ensure meaningful comparison. To construct the available MAS sets, we run existing methods such as MAS-Zero on benchmark and collect the success MASs which return the expected outputs (high-quality MASs).

Evaluation Metrics. We assess each approach based on effectiveness and efficiency using the following primary metrics:

- **Accuracy:** Measures the proportion of correctly solved queries. A query is marked correct if the MAS output matches the ground truth.
- **Cost:** Is the average monetary cost incurred during inference per query. This is computed based on token usage across all LLM calls in both generation and execution.
- **Time:** Captures the average wall-clock time required to generate and execute a MAS per query.

B. Experimental Procedure

RQ1: Effectiveness of MAS Generation. We compare MASAC against two main categories of MAS design methods:

(1) *Manual MAS Design.* This includes widely adopted human-engineered interaction patterns [20]: *Chain-of-Thought*, *Self-Consistency Chain-of-Thought*, *Debate*, and *Reflection*. These strategies are commonly used to enhance reasoning quality and incorporated into the operator library in the existing work [9], [10], [17].

(2) *Automated MAS Design.* We evaluate against the state-of-the-art methods in two categories:

- **Task-level methods:** *ADAS* [9] and *MaAS* [14], which generate fixed MAS designs per task domain.
- **Query-level methods:** *MAS-Zero* [17] which dynamically constructs MASs on a per-query basis.

Following the procedure in existing work [9], [14], [17], for MASAC and all baselines, is evaluated on the same set of benchmark queries. We ensure fair comparison by standardizing prompt formatting, using the same backbone LLMs, evaluation metrics, and running all models under identical runtime and hardware constraints. Accuracy is measured as the proportion of correctly solved queries.

RQ2: Efficiency and Computational Cost. We measure the cost and latency of each method in generating and executing MASs, and compare against Query-level methods, which have both generation and execution phases, facilitate comprehensive evaluation. The cost is measured in USD based

TABLE I: Accuracy of generated MASs

	gpt-4.1-mini		gpt-4o		o4-mini	
	GPQA	AIME24	GPQA	AIME24	GPQA	AIME24
CoT	62.50	43.33	47.91	13.33	75.00	83.33
SC-CoT	60.41	36.66	47.91	6.67	79.17	86.66
Debate	64.58	50.00	58.33	10.00	83.33	86.66
Self-Refine	64.58	50.00	47.91	10.00	77.08	83.33
ADAS	68.80	×	45.80	×	76.00	×
MaAS	47.92	56.67	27.08	13.33	37.50	76.67
MAS-ZERO	62.50	33.33	62.50	16.67	81.25	73.33
MASAC	83.33	56.67	70.83	20.00	85.41	83.33

“×” indicates that the baseline is not applied to that benchmark.

on prompt and completion token usage. Time is measured as latency from query input to final output.

RQ3: Model Configuration Analysis. We evaluate MASAC using multiple backend LLMs, including gpt-4.1-mini, gpt-4o, and o4-mini, to analyze performance consistency across model variants. This analysis provides insight into how sensitive MASAC is to changes in LLM capability and pricing, and whether MASAC maintains robust performance under varying LLM constraints.

V. EXPERIMENTAL RESULTS

A. Answer RQ1: Effectiveness of MAS Generation

Table I shows the performance MASAC and other methods using various LLMs on *GPQA_diamond* and *AIME24*. The results show that MASAC *consistently outperforms state-of-the-art methods in both Manual MAS Design and Automated MAS Design, across both benchmarks and various LLMs.* On *GPQA_diamond*, the MASs generated by MASAC are much better than those by the other methods by 21–73%, 13–161%, and up to 127% when using gpt-4.1-mini, gpt-4o, and o4-mini respectively. This trend persisted on *AIME24*.

Compared to SOTA *task-level* methods (*ADAS* and *MaAS*), MASAC achieves superior performance, with improvements ranging from 12% to 55% over *ADAS* and up to 161% over *MaAS*. Meanwhile, compared to SOTA *query-level* method (*MAS-Zero*), MASAC consistently outperforms, demonstrating enhancements ranging from 5% to 70%. These improvements can be attributed to MASAC’s ability to effectively extract useful design components shared across queries within the same domain (*Offline MAS Abstraction*) while still consider query-specific requirements in the MAS generation process (*Online MAS Concretization*).

Especially, for all the studied LLMs, MASAC demonstrates significant improvements over the studied human-engineered interaction patterns (*Manual MAS Design*). For instance, when utilizing gpt-4.1-mini on *GPQA_diamond*, CoT pattern achieves an accuracy of 62.5%, whereas MASAC attains an accuracy of 83.33%, approaching the performance levels of the advanced reasoning model, o4-mini which are inherently designed for complex problem-solving tasks. On *AIME24* which is more challenging, MASAC still consistently outperforms these patterns. When applying reasoning model, o4-mini, MASAC achieves notable improvements on

TABLE II: Efficiency and Computational Cost

LLM	Approach	Acc.	Gen. time	Exe. time	Gen. cost	Exec. cost
gpt-4.1-mini	MAS-Zero	62.50	289.48	99.80	0.05	0.02
	MASAC	83.33	163.16	224.55	0.05	0.04
gpt-4o	MAS-Zero	62.50	293.14	58.39	0.24	0.05
	MASAC	70.83	135.75	124.81	0.11	0.14
o4-mini	MAS-Zero	81.25	452.89	200.65	0.23	0.08
	MASAC	85.41	160.32	279.66	0.10	0.15

the *GPQA_diamond* benchmark and exhibits only a marginal 4% performance gap on *AIME24*.

B. Answer RQ2: Efficiency and Computational Cost

In this experiment, we measured both the latency and cost associated with MASAC during the MAS generation and execution phases on *GPQA_diamond*. As shown in Table II, MASAC consistently achieved approximately **half** the generation time (Gen. time) and cost (Gen.cost) compared to *MAS-Zero* across all evaluated LLMs. This improvement is largely attributed to MASAC’s ability to reuse shared design components across queries within the same domain, rather than generating MAS designs from scratch as *MAS-Zero* does.

However, the computational time (Exe. time) and cost (Exe. cost) required to execute the MAS generated by MASAC are significantly higher than those of *MAS-Zero*. By analyzing the MASs generated by both MASAC and *MAS-Zero*, we found that the MAS generated by MASAC incorporates basic operators similar to those in *MAS-Zero* but includes a greater number of subtasks with more complex operator combinations. MASAC prioritizes task decomposition to clearly and specifically delineate the tasks to be addressed, thereby mitigating model hallucination caused by overly complex tasks. Despite the additional cost during execution, MASAC achieves a much lower overall time-to-completion (generation + execution) while producing higher-quality MASs than *MAS-Zero* [17]. This demonstrates the efficiency of MASAC in balancing generation reuse with execution precision. Such a property is crucial for real-world deployments, where minimizing end-to-end latency while maintaining high-quality MAS generation is often the primary requirement.

C. Answer RQ3: Model Configuration Analysis

To evaluate the impact of different LLM configurations, we varied the generation or execution models of MASAC while keeping the other fixed, using *GPQA_diamond*.

Figure 2a shows that using gpt-4.1-mini or o4-mini for MAS generation yields higher accuracy (up to 8% improvement) and lower cost (up to 35.7% reduction) than gpt-4o, which performs worst (77.08% accuracy, \$0.14/query). Although gpt-4o offers faster generation, it lacks the instruction-following and reasoning abilities necessary for optimal MAS planning. Therefore, we recommend using models with strong instruction-following capabilities, such as gpt-4.1-mini or o4-mini, in the generation phase for better accuracy-cost trade-offs.

As shown in Figure 2b, gpt-4o again performs worst during MAS execution (66.67% accuracy, \$0.258/query). In

contrast, reasoning-capable models like gpt-4.1-mini and o4-mini improve accuracy by 25–34% and reduce cost by 22–65%. Notably, o4-mini achieves the best overall performance (89.58% accuracy, \$0.20/query). While gpt-4o is faster, it lacks the academic and reasoning proficiency required for complex, multi-step tasks. Thus, we recommend using o4-mini or gpt-4.1-mini for execution, especially when precision and reasoning are critical.

Summary. In both phases, LLMs with strong instruction-following and reasoning capabilities (e.g., o4-mini, gpt-4.1-mini) significantly enhance MASAC’s performance. While faster models like gpt-4o may offer time savings, they are less reliable for complex tasks. We recommend prioritizing reasoning-rich models in both generation and execution to maximize accuracy and cost-efficiency.

VI. THREATS TO VALIDITY

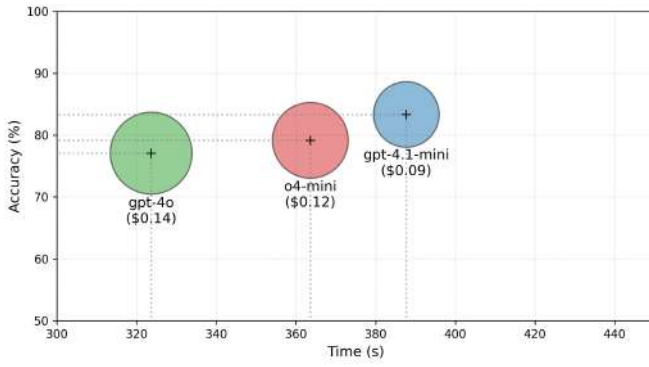
The primary threat to the validity of our study lies in the completeness of the existing MAS corpus used as the foundation for the *Offline MAS Abstraction* (Section III-A) process. To mitigate this threat, during the Stage Decomposition process, the goal of individual stages are deliberately described in a generalized manner to enhance applicability across different domains. In this way, even in the absence of data to extract patterns from new domains, MASAC can still construct a MAS that remains effective in handling queries.

VII. RELATED WORK

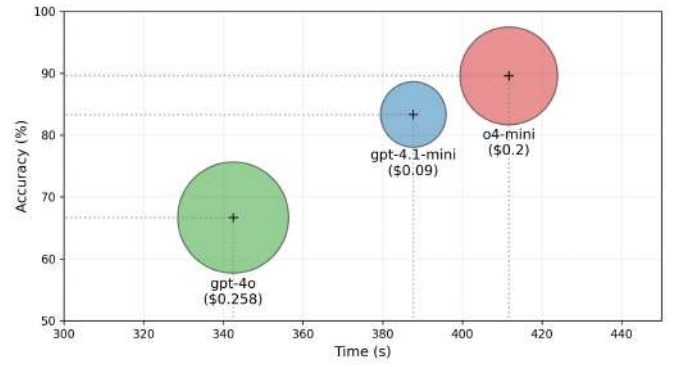
Early approaches to improving LLM reasoning capabilities rely on manually crafted MAS interaction patterns [20]. These include *Chain-of-Thought*, *Self-Consistency Chain-of-Thought*, *Debate*, and *Reflection*, which are designed to enhance reasoning and problem-solving. While effective, these designs lack adaptability and require significant human effort.

To overcome the limitations of manual engineering, recent work has focused on automated MAS design which can be categorized into two types: *task-level* and *query-level* methods.

Task-level approaches aim to generate MAS designs that are fixed for a specific task domain. GPTSwarm [13] formulates the MAS workflow as a graph or network and optimize the inter-agent connections to solve domain-specific problems like code generation or mathematical reasoning. ADAS [9] and AFlow [10] represent both agents and workflows as executable code and treat MAS generation as a code generation problem. ADAS uses a meta-agent search strategy, while AFlow employs Monte Carlo Tree Search to explore the MAS design space. ScoreFlow [15] improves upon this by training a workflow generator through direct preference optimization, while MaAS [14] introduces an agentic supernet followed by multi-agent architecture search. Despite their advances, these methods generate one MAS per task, limiting their adaptability and efficiency when facing diverse query inputs within the same domain. In contrast, *query-level* methods, such as MAS-Zero [16], [17], aim to dynamically construct a customized MAS design for each input query. This enables fine-grained adaptation to query-specific requirements and allows for more



(a) Generation Model Analysis



(b) Execution Model Analysis

Fig. 2: Model Configuration Analysis (X-axis: time per a query (s), Y-axis: accuracy (%), radius of bubble: cost per query (\$))

efficient and context-sensitive reasoning. However, MAS-Zero and other query-level methods generate MASs from scratch with some handcrafted patterns and overlook the useful MAS design components shared between queries in the same domain, which may restrict its generalization and scalability.

MASAC differs from prior work in several key aspects. Compared to task-level methods, MASAC operates at the query level, enabling dynamic adaptation to each individual query. Furthermore, in contrast to MAS-Zero, MASAC performs explicit planning over a rich operator library and introduces a novel generation-execution decoupling mechanism, allowing it to synthesize high-quality, query-specific MAS designs that balance reasoning effectiveness and execution cost.

VIII. CONCLUSION

This paper introduces MASAC, a novel framework for automated multi-agent system (MAS) design generation. By leveraging a modular operator library and a recursive search strategy guided by LLMs, MASAC constructs effective and adaptable MASs tailored to individual input queries. Extensive experiments across multiple reasoning tasks demonstrate that MASAC outperforms both manual and existing automated MAS design approaches in accuracy, efficiency, and robustness across different LLM backends. These results highlight the promise of data-driven, compositional MAS generation for advancing multi-agent reasoning capabilities.

ACKNOWLEDGMENTS

This research has been done under the research project QG.24.37 of Vietnam National University, Hanoi

REFERENCES

- [1] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv:2501.12948*, 2025.
- [2] T. Vu, K. Krishna, S. Alzubi, C. Tar, M. Faruqui, and Y.-H. Sung, “Foundational autoraters: Taming large language models for better automatic evaluation,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 086–17 105.
- [3] T.-D. Bui, D.-T. Luu-Van, T.-P. Nguyen, T.-T. Nguyen, S. Nguyen, and H. D. Vo, “Rambo: Enhancing rag-based repository-level method body completion,” *arXiv preprint arXiv:2409.15204*, 2024.
- [4] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, “Llm-planner: Few-shot grounded planning for embodied agents with large language models,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2998–3009.
- [5] L. Zhong, Z. Wang, and J. Shang, “Debug like a human: A large language model debugger via verifying runtime execution step by step,” in *Findings of ACL 2024*, 2024, pp. 851–870.
- [6] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, “Pal: Program-aided language models,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 764–10 799.
- [7] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, “Gpqa: A graduate-level google-proof q&a benchmark,” in *First Conference on Language Modeling*, 2024.
- [8] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. R. Narasimhan, “Swe-bench: Can language models resolve real-world github issues?” in *ICLR 2024*.
- [9] S. Hu, C. Lu, and J. Clune, “Automated design of agentic systems,” in *The First Workshop on System-2 Reasoning at Scale, NeurIPS’24*.
- [10] J. Zhang, J. Xiang, Z. Yu, F. Teng, X. Chen, J. Chen, M. Zhuge, X. Cheng, S. Hong, J. Wang *et al.*, “Aflow: Automating agentic workflow generation,” *arXiv preprint arXiv:2410.10762*, 2024.
- [11] Z. Liu, Y. Zhang, P. Li, Y. Liu, and D. Yang, “Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization,” *arXiv preprint arXiv:2310.02170*, 2023.
- [12] G. Zhang, K. Chen, G. Wan, H. Chang, H. Cheng, K. Wang, S. Hu, and L. Bai, “EvoFlow: Evolving diverse agentic workflows on the fly,” *arXiv preprint arXiv:2502.07373*, 2025.
- [13] M. Zhuge, W. Wang, L. Kirsch, F. Faccio, D. Khizbullin, and J. Schmidhuber, “Gptswarm: Language agents as optimizable graphs,” in *Forty-first International Conference on Machine Learning*, 2024.
- [14] G. Zhang, L. Niu, J. Fang, K. Wang, L. Bai, and X. Wang, “Multi-agent architecture search via agentic supernet,” *arXiv preprint arXiv:2502.04180*, 2025.
- [15] Y. Wang, L. Yang, G. Li, M. Wang, and B. Aragam, “Scoreflow: Mastering llm agent workflows via score-based preference optimization,” *arXiv preprint arXiv:2502.04306*, 2025.
- [16] H. Gao, Y. Liu, Y. He, L. Dou, C. Du, Z. Deng, B. Hooi, M. Lin, and T. Pang, “Flowreasoner: Reinforcing query-level meta-agents,” *arXiv preprint arXiv:2504.15257*, 2025.
- [17] Z. Ke, A. Xu, Y. Ming, X.-P. Nguyen, C. Xiong, and S. Joty, “Meta-design matters: A self-design multi-agent system,” *arXiv preprint arXiv:2505.14996*, 2025.
- [18] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, “Large language model based multi-agents: a survey of progress and challenges,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 8048–8057.
- [19] “American invitational mathematics examination,” Mathematical Association of America, 2025, <https://maa.org/maa-invitational-competitions/>.
- [20] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen *et al.*, “Siren’s song in the ai ocean: A survey on hallucination in large language models,” *Computational Linguistics*, pp. 1–46, 2025.

Enhance Hyper-Actor Critic Framework in Session-based Recommendation System with Nash Social Welfare

Quoc Huy Nguyen

VNU University of Engineering and Technology
Hanoi, Vietnam

Bang Giang Le

VNU University of Engineering and Technology
Hanoi, Vietnam

Viet Cuong Ta*

VNU University of Engineering and Technology
Hanoi, Vietnam
cuongtv@vnu.edu.vn

Abstract—Recent advances in Reinforcement Learning (RL) have significantly improved session-based recommendation systems by modeling user-item interactions as Markov Decision Processes. The Hyper-Actor Critic (HAC) framework addresses scalability and long-term optimization by decomposing item selection into a latent hyper-action space. However, the original HAC often neglects fairness and welfare across diverse user preferences, often amplifying utility disparities. In this paper, we propose Fair-HAC, an enhanced variant of HAC that incorporates social welfare optimization to ensure equitable recommendation outcomes. Our approach introduces a multi-objective policy which is trained to maximize the Nash Social Welfare over session trajectories. By using an aggregated Q-value sampling over a session, our value function is trained to capture both effectiveness and fairness across user segments. The experiments on three datasets demonstrate that Fair-HAC improves both aggregate performance and the interaction depth metrics, which achieving more stable recommendations in comparison to other baseline RL models.

Index Terms—Reinforcement Learning, Session-based Recommendation, Nash Social Welfare, Hyper-Actor Critic, Multi-objective Optimization

I. INTRODUCTION

Reinforcement learning (RL) has become increasingly effective in modeling dynamic user interactions for session-based recommendation [1]. The Hyper-Actor Critic (HAC) [2] framework, in particular, represents a significant advancement by decomposing decision-making into a latent hyper-action space which is designed to efficiently explore in large item domains. By separating the actor into a latent policy sampler and a decoder, HAC can model complex user interactions while maintaining computational efficiency. However, maximizing expected cumulative rewards in standard HAC can result in systematically unequal treatment of different user groups or session types [3], [4].

To enhance fairness in recommendation systems, researchers have developed a variety of mitigation strategies that can

be broadly categorized by their point of intervention. Data-oriented methods attempt to address bias at its source by adding antidote data to rebalance skewed distributions [5]. Ranking methods integrate fairness directly into the model’s objective by adding regularization terms [6]. In [7], the authors propose a re-ranking method to adjust initial candidate lists to satisfy fairness constraints. Within the specific domain of RL, one prominent approach is to frame the problem as a multi-objective learning, which agents find a Pareto-optimal balance between utility and fairness metrics [8], [9]. Other methods include using adversarial learning to remove biases from representations [10] or allowing for personalized definitions of user-oriented fairness [11].

In this paper, we propose Fair-HAC, an extension of the HAC framework that incorporates fairness via the Nash Social Welfare (NSW) objective [8]. The Fair-HAC is designed to target the group fairness in session-based recommendation by using an modified objective function of the learning actor. Given an input user session, we can derive the geometric mean of utilities as the NSW and integrate into to learning of the actor in HAC. In our formulation, we treat cumulative rewards estimated by the critic network over a batch as the NSW utility signals. Aggregation functions are then applied to highlight reward disparities across users. During actor optimization, we adopt an NSW inspired loss that penalizes policies yielding low minimum rewards. The modified actor encourages the agent to improve not only average performance but also ensure equitable outcomes for underperforming users, aligning with the core fairness objectives of NSW. We evaluate Fair-HAC using the ML1M, KuaiRand1K and RL4RS dataset under a simulated online environment. Experimental results demonstrate that our method achieves a more balanced reward distribution across sessions while maintaining high recommendation performance.

In the remaining section of the paper, the setting and background are presented in Section II. Our proposed method,

*Corresponding author.

Fair-HAC, is presented in Section III. The experiments are discussed in Section VI and the conclusion is presented in Section V.

II. BACKGROUND

A. HAC in Session-Based Recommendation

In session-based recommendation settings, the objective is to recommend the next item given an user interaction trajectory $x_{1:t} = \{x_1, x_2, \dots, x_t\}$ with x_i is the i^{th} item of the session. We can assume that the total available item is the set $\mathcal{I}$ and at each time step t , the system chooses a specific item or a slate of items from $\mathcal{I}$ as the next action to interact with a user. In the HAC framework [2], this task is formulated as a Markov Decision Process (MDP), defined by the tuple $\langle \mathcal{S}, \mathcal{Z}, \mathcal{A}, \mathcal{R}, \gamma \rangle$, where:

- $\mathcal{S}$: the user state space,
- $\mathcal{Z}$: the latent action space (i.e., hyper-action space),
- $\mathcal{A}$: the actual action space (e.g., slate of recommended items),
- $\mathcal{R}$: the reward function,
- γ : the discount factor.

Given a policy π that maps the current user state s_t to an action a_t , the goal of HAC is to learn the optimal policy π^* that maximizes the expected cumulative reward:

$$\pi^* = \arg \max_{\pi} \mathbb{E} \left[\sum_{t=1}^T \gamma^t \cdot r_t \right] \quad (1)$$

To handle the large and discrete action space in recommendation, HAC introduces a low-dimensional latent action space $\mathcal{Z}$. A scoring function maps latent actions $Z_t \in \mathcal{Z}$ to actual item slates $a_t \in \mathcal{A}$:

$$\text{score}(i|Z_t) = \Phi(i)^T Z_t \quad (2)$$

where $\Phi(i)$ is the kernel function. In general, the action a_t contains a list of items i , which are sampled through a softmax function:

$$P(i|Z_t) = \text{softmax}(\text{score}(i|Z_t)) \quad (3)$$

To encourage exploration in the latent space, HAC applies Gaussian noise to the sampled hyper-actions. The shared critic network, inspired by the Deep Q-Network (DQN) architecture [12], is trained using temporal-difference (TD) learning. Additionally, HAC applies a regularization loss to align the decoded action a_t with the corresponding latent representation Z_t .

B. Actor Loss in HAC

The actor in the standard HAC framework is trained to generate latent hyper-actions Z_t from the current user state s_t , which are then decoded into item recommendations. To guide the learning of the actor, the critic estimates the Q-value $Q(s_t, Z_t)$, representing the expected return of executing a particular latent action in the current state.

The actor is updated by minimizing the negative Q-value output from the critic, following a standard value-based optimization approach. The actor loss in HAC is formulated as:

$$\mathcal{L}_{\text{actor}} = -Q(s_t, Z_t) \quad (4)$$

This objective encourages the actor to generate actions that maximize cumulative reward. However, if the actor is over-exploitation, the output hyper action would focus on only a subset of user with high rewards. In this work, we aim to resolve the issue of over-exploitation of HAC by proposing a modification to the actor loss function.

III. METHOD

To reduce the above over-exploitation effects, we propose Fair-HAC, a fairness-aware extension of the HAC framework. While the original HAC optimizes expected cumulative return across sessions, Fair-HAC incorporates a fairness signal that penalizes reward disparity among user sessions. This is achieved by modifying the actor's loss to include a batch-level minimum Q-value, promoting equitable learning across underperforming trajectories. The overall architecture of Fair-HAC is illustrated in Fig. 1. Our NSW module is designed to update the policy with a more balancing target.

A. Q-value Aggregates in Actor Optimization

In HAC, the actor generates latent hyper-actions Z_t , conditioned on the current state s_t , which are then decoded into item recommendations. The actor is optimized via value-based learning, guided by the critic's estimate of $Q(s_t, Z_t)$. In the original formulation, this is done by maximizing the average Q-value over a mini-batch of training samples:

$$\bar{Q} = \frac{1}{n} \sum_{i=1}^n Q(s_i, Z_i), \quad (5)$$

$$\mathcal{L}_{\text{actor}}^{\text{HAC}} = -\bar{Q}$$

In the above equations, the index $i \in [1, n]$ use to denote users in a mini-batch of size n , with each sample corresponding to a user session represented by state s_i and latent action Z_i .

While optimizing $\bar{Q}$ typically leads to policies with high average return, it may overlook sessions that consistently yield poor outcomes. To mitigate this, we introduce a complementary optimization signal: an *aggregate Q-value* that emphasizes specific statistical properties across the batch. It is defined as:

$$Q_{\text{agg}} = \text{agg}_{i \in [1, n]} Q(s_i, Z_i) \quad (6)$$

where agg denotes an aggregation function applied across the Q-values of all users in the mini-batch. This term captures worst-case or low-performing returns and surfaces disparities otherwise masked by the average.

B. Fairness-Aware Actor Loss Based on NSW

To balance performance and fairness, we formulate a new actor loss inspired by *Nash Social Welfare (NSW)* objectives. The modified actor loss is:

$$\mathcal{L}_{\text{actor}}^{\text{Fair}} = -\log(\bar{Q} + \alpha \cdot Q_{\text{agg}}) \quad (7)$$

where $\alpha > 0$ is a hyper-parameter to balance the effect of the aggregated value to the standard average Q value of HAC

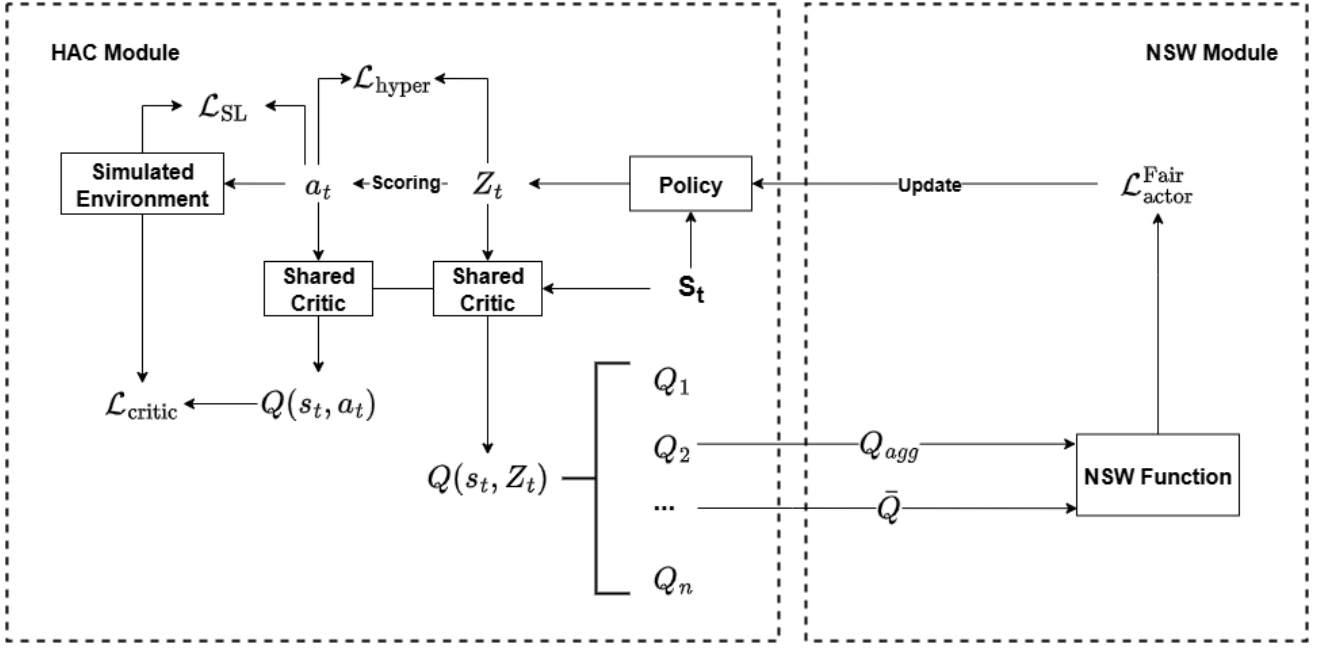


Fig. 1. Pipeline of the proposed Fair-HAC framework. Our NSW module updates the policy in HAC using a fairness-aware loss.

$\bar{Q}$. This log-form encourages both a high average return and improvement in low-reward sessions, reducing the disparity across session-level outcomes. Both $\bar{Q}$ and $Q_{\min}$ are computed from critic estimates in each batch, ensuring sensitivity to training-time reward distribution.

Algorithm 1 Fair-HAC Training with NSW Module

- 1: Initialize actor θ , critic ϕ , item kernel Φ
- 2: Initialize target networks $\theta_{\text{target}} \leftarrow \theta$, $\phi_{\text{target}} \leftarrow \phi$
- 3: Initialize replay buffer $\mathcal{B}$
- 4: **while** not converged **do**
- 5: Collect transitions $(s_t, Z_t, a_t, r_t, s_{t+1}, d)$ using π_θ
- 6: Store transitions in $\mathcal{B}$
- 7: Sample mini-batch $\mathcal{M} \sim \mathcal{B}$
- 8: Compute Q_{target} with Eq.(8)
- 9: Update critic ϕ to minimize Eq.(9)
- 10: Compute batch mean Q-value: $\bar{Q}$ with Eq.(5)
- 11: Compute batch aggregate Q-value (min or max): Q_{agg} with Eq.(6)
- 12: Update actor θ to minimize Eq.(12)
- 13: Update Φ with Eq.(13)
- 14: **end while**

C. Implementation Details

The critic in Fair-HAC is updated using a standard temporal-difference (TD) learning rule to estimate the action-value function. The TD error is computed as:

$$Q_{\text{target}} = r_t + \gamma \cdot (1 - d) \cdot Q(s_{t+1}, Z_{t+1}) \quad (8)$$

$$\mathcal{L}_{\text{critic}} = \mathbb{E} \left[(Q(s_t, Z_t) - Q_{\text{target}})^2 \right] \quad (9)$$

where d is the termination signal. This approach follows deep Q-learning paradigms such as DQN [12], ensuring stability and reliable convergence in value estimation.

To complement this, we apply a hyper-actor regularization term to enforce consistency between the sampled latent action Z and its reconstruction $\hat{Z}$ through the decoder. The loss is defined as:

$$\mathcal{L}_{\text{hyper}} = \lambda_{\text{hyper}} \cdot \|\hat{Z} - Z\|_2^2 \quad (10)$$

where $\lambda_{\text{hyper}} > 0$ is a hyperparameter that controls the strength of the regularization on the latent space. This regularization encourages alignment between the latent policy sampler and the decoded representation, preserving the structural integrity of HAC's two-stage latent-to-actual action pipeline. The overall training algorithm is described in Algorithm 1.

To train in a simulated setting, a session-based simulation environment is built following the user response model (URM) paradigm using the input dataset. A pre-trained URM Ψ is used to generate stochastic binary feedback $y_t \in \{0, 1\}$ for each action slate a_t given user state s_t . The probability is given by:

$$P(y_t | s_t, a_t) = \Psi(s_t, a_t)$$

The reward signal $r(s_t, a_t)$ is computed as the average score over items: +1 for positive response and -0.1 for negative feedback. Session termination is governed by a temper mechanism. Each session starts with a temper budget $B_0 = 20$, and is decremented based on reward satisfaction:

$$B_{t+1} = B_t - ((r(s_t, a_t) - T) \cdot k + 1)$$

where k is the slate size and T is the satisfaction threshold. The session ends if $B_t < 1$.

TABLE I
ONLINE PERFORMANCE COMPARISON ACROSS DATASETS. BEST IN **BOLD**, SECOND BEST IN UNDERLINE.

Model	RL4RS		ML1M		KuaiRand1K	
	Total Reward	Depth	Total Reward	Depth	Total Reward	Depth
SAC	4.71 $\pm$ 1.92	6.51 $\pm$ 1.75	1.96 $\pm$ 0.79	3.79 $\pm$ 0.75	1.40 $\pm$ 2.05	3.33 $\pm$ 1.81
A2C	8.24 $\pm$ 6.74	9.67 $\pm$ 5.93	3.12 $\pm$ 1.13	4.77 $\pm$ 1.03	9.87 $\pm$ 7.66	10.94 $\pm$ 6.85
DDPG	5.17 $\pm$ 1.98	6.74 $\pm$ 1.81	3.05 $\pm$ 1.17	4.72 $\pm$ 1.06	4.91 $\pm$ 5.59	6.45 $\pm$ 4.96
HAC	13.11 $\pm$ 3.86	13.74 $\pm$ 3.47	7.50 $\pm$ 1.82	8.68 $\pm$ 1.69	12.27 $\pm$ 8.67	13.05 $\pm$ 7.76
Fair-HAC	16.35 $\pm$ 3.15	16.66 $\pm$ 2.86	7.56 $\pm$ 1.85	8.74 $\pm$ 1.72	12.34 $\pm$ 8.43	13.12 $\pm$ 7.55

We use a supervised learning loss based on the simulated environment:

$$\mathcal{L}_{SL} = \mathbb{E} \sum_{i \in a_t} [y_{t,i} \log P(i | \mathbf{Z}_t) + (1 - y_{t,i}) \log(1 - P(i | \mathbf{Z}_t))] \quad (11)$$

where $y_{t,i}$ is the ground truth user response for the exposed item i .

Combining Eq. (7), Eq. (10), and Eq. (11), the loss for the actor θ is defined as:

$$\mathcal{L}_{\theta} = \mathcal{L}_{actor}^{Fair} + \mathcal{L}_{hyper} + \mathcal{L}_{SL} \quad (12)$$

The item kernel is updated by minimizing the combination of Eq. (10) and Eq. (11):

$$\mathcal{L}_{\Phi} = \mathcal{L}_{hyper} + \mathcal{L}_{SL} \quad (13)$$

Besides that, to stabilize the training process, we maintain target networks for both the actor and the critic. These target networks are updated using a soft update rule, which slowly blends the parameters of the main network into the target network. This technique helps reduce training variance and prevents divergence in temporal-difference updates [13].

IV. EXPERIMENTS

A. Datasets

In this section, we evaluate the performance of our proposed Fair-HAC model and compare it against other baselines, including original HAC framework. We conduct the experiments on three widely used recommendation datasets:

- **RL4RS**: A synthetic benchmark simulating user interactions with 360-dimensional item features and 10-dimensional user embeddings. It captures the sequential nature of decision-making and supports interaction-based evaluation [14].
- **KuaiRand1K**: A real-world short-video interaction dataset exhibiting sparse and short-term behaviors across 265,000 users [15].
- **MovieLens 1M (ML1M)**: A dense dataset containing explicit feedback from over 6,000 users on approximately 4,000 movies. Each user has a long history of rated items [16].

B. Baseline Models

We compare our method, Fair-HAC, against four actor-critic-based RL baselines:

- **HAC**: Hyper-Actor Critic model which introduces latent action vectors and a shared critic [2].
- **SAC**: Soft Actor-Critic with entropy maximization for efficient exploration [17].
- **A2C**: A synchronous actor-critic model trained with n-step returns and entropy regularization [18].
- **DDPG**: A deterministic policy gradient algorithm using continuous latent actions and experience replay [19].

For our Fair-HAC and all baseline methods, we adopt SASRec [?] as the backbone actor, consisting of an encoder and a scoring module. The encoder is a Transformer with 2 layers, 4 attention heads, and a dropout rate of 0.1. A causal attention mask is applied to prevent access to future positions, and the final state embedding is taken from the output at the last position. Item features are embedded via a shared kernel Φ , implemented as a single linear layer projecting into a 32-dimensional latent space. The scoring module is a dot-product scorer that maps the latent user-action embedding to item probabilities. The critic is a two-layer multilayer perceptron with hidden dimensions of 256 and 64, and a dropout rate of 0.2. We aggregate using the minimum operator in Eq. (6) and set $\alpha = 3$ in Eq. (7).

We use a fixed URM simulation with offline rollouts and temper-based termination to ensure consistency across baselines. Interaction depth is capped at 20, with starting temper $B_0 = 20$ and discount factor $\gamma = 0.9$. The dataset is split into 80% training and 20% testing. Evaluation metrics are Total Reward and Depth: Total Reward is the average cumulative reward per session, and Depth is the average number of interactions before session termination.

C. Results

TABLE II
COMPARISON OF MINIMUM REWARD AND MINIMUM STEP ON RL4RS.

Model	Min Reward	Min Step
A2C	2.27	4.26
DDPG	2.12	4.05
HAC	4.29	<u>5.91</u>
Fair-HAC	10.48	11.39

The overall performance comparison is presented in Table I. Fair-HAC consistently outperforms all baselines on RL4RS, KuaiRand1K and ML1M. Specifically, it achieves the highest total reward and session depth across two of the three datasets. On RL4RS, Fair-HAC improves total reward by over +3.24

TABLE III
PERFORMANCE COMPARISON OF FAIR-HAC VARIANTS ACROSS DATASETS.

Model	RL4RS		ML1M		KuaiRand1K	
	Total Reward	Depth	Total Reward	Depth	Total Reward	Depth
Fair-HAC-Min	16.35 $\pm$ 3.15	16.66 $\pm$ 2.86	7.56 $\pm$ 1.85	8.74 $\pm$ 1.72	12.34 $\pm$ 8.43	13.12 $\pm$ 7.55
Fair-HAC-Max	16.13 $\pm$ 2.58	16.45 $\pm$ 2.36	7.51 $\pm$ 1.84	8.69 $\pm$ 1.71	11.68 $\pm$ 8.48	12.52 $\pm$ 7.58

compared to HAC, and achieves a session depth of 16.66 steps on average, suggesting both higher engagement and effectiveness.

Table II further reports the minimum reward and the corresponding steps required to reach it on RL4RS. Fair-HAC attains the highest minimum reward 10.48 and requires the longest step trajectory 11.39, which indicates that its policy not only delivers more consistent recommendation quality but also retains users longer. This supports our motivation to incorporate fairness via the NSW objective.

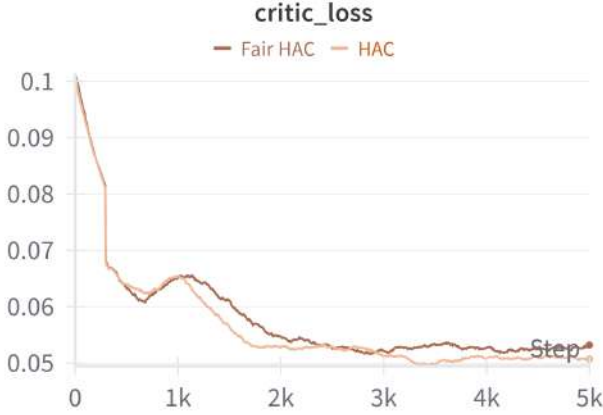


Fig. 2. Critic loss during training. Both HAC and Fair-HAC maintain stable Q-value convergence despite Fair-HAC using a fairness-aware actor loss.

For the choice of aggregation function in Eq. (6), we compares the performance of two fairness-aware variants, which are Fair-HAC-Min and Fair-HAC-Max in Table III. On the RL4RS dataset, Fair-HAC-Min outperforms Fair-HAC-Max in both reward and depth, achieving a total reward of 16.35 ± 3.15 and an average depth of 16.66 ± 2.86 , compared to Fair-HAC-Max’s 16.13 ± 2.58 and 16.45 ± 2.36 . This indicates that penalizing underperforming sessions (via the minimum Q-signal) leads to better overall performance and longer user engagement in a sequential recommendation setting. For ML1M, the two variants perform nearly identically, with Fair-HAC slightly leading in both metrics. On KuaiRand1K, which features shorter videos and noisier user signals, Fair-HAC-Min again demonstrates superior performance with a total reward of 12.34 ± 8.43 and depth of 13.12 ± 7.55 , outperforming Fair-HAC-Max’s respective values of 11.68 ± 8.48 and 12.52 ± 7.58 . In general, the choice of minimum aggregate function provides the most stable performance.

To assess whether the proposed actor loss interferes with

critic stability, we compare the training loss curves of the Q-value network. Figure 2 shows that both HAC and Fair-HAC achieve similar convergence rates. Despite the change in the actor’s optimization objective, the Q-value estimates from the critic remain smooth and stable.

TABLE IV
EFFECT OF FAIRNESS REGULARIZATION COEFFICIENT α ON RL4RS.

α	Total Reward	Depth
1	16.63 $\pm$ 3.22	16.92 $\pm$ 2.93
3	16.35 $\pm$ 3.15	16.66 $\pm$ 2.86
5	16.31 $\pm$ 3.03	16.63 $\pm$ 2.76

To assess the effect of fairness regularization hyper-parameter α , we conduct an ablation study varying the coefficient α in the Fair-HAC framework on the RL4RS dataset. As shown in Table IV, total reward and session depth remain competitive across all values, with the best performance at $\alpha = 1$, where the model achieves a total reward of 16.63 and an average session depth of 16.92. These results suggest that a moderate fairness weight effectively balances reward maximization and user engagement. However, performance gradually declines as α increases, indicating a trade-off: stronger fairness regularization slightly penalizes overall utility. At $\alpha = 5$, both reward and depth drop to their lowest values in the comparison, though still comparable to other configurations.



Fig. 3. Reward variance across training steps for different α values. Higher α reduces variance, indicating fairer outcomes across sessions.

We further plot the reward variance in Figure 3 during training. We observe that Fair-HAC with $\alpha = 5$ maintains consistently lower reward variance compared to Fair-HAC with $\alpha = 1$ and $\alpha = 3$, which confirms that higher α

effectively reduces reward disparity. Fair-HAC with $\alpha = 1$, while yielding the best overall reward, has the highest variance, suggesting that certain user sessions may still receive suboptimal outcomes. Therefore, the hyper-parameter α can be used as a tunable parameter to control the fairness-performance trade-off.

V. CONCLUSION

In this work, we proposed Fair-HAC, a fairness-aware extension of the HAC framework for session-based recommendation. Unlike prior approaches that optimize only for expected reward, our method explicitly incorporates the Nash Social Welfare objective into the actor loss. This design encourages policies that improve the equitable distribution of rewards while maintaining competitive average performance. Experimental results across three recommendation datasets demonstrate that Fair-HAC achieves a significant reduction in reward disparity and superior session retention, while retaining strong total reward. Moreover, our convergence analysis shows that the critic's Q-value learning remains stable, despite changes to the actor loss.

REFERENCES

- [1] M. M. Afsar, T. Crump, and B. Far, "Reinforcement learning based recommender systems: A survey," 2022.
- [2] S. Liu, Q. Cai, B. Sun, Y. Wang, J. Jiang, D. Zheng, P. Jiang, K. Gai, X. Zhao, and Y. Zhang, "Exploration and regularization of the latent action space in recommendation," in *Proceedings of the ACM Web Conference 2023*, WWW '23, p. 833–844, ACM, Apr. 2023.
- [3] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in reinforcement learning," in *International conference on machine learning*, pp. 1617–1626, PMLR, 2017.
- [4] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, "A survey on the fairness of recommender systems," *ACM Transactions on Information Systems*, vol. 41, p. 1–43, Feb. 2023.
- [5] B. Rastegarpanah, K. P. Gummadi, and M. Crovella, "Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, p. 231–239, ACM, Jan. 2019.
- [6] Z. Zhu, X. Hu, and J. Caverlee, "Fairness-aware tensor-based recommendation," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, (New York, NY, USA), pp. 1153–1162, Association for Computing Machinery, 2018.
- [7] M. Zehlke, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa*ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, p. 1569–1578, ACM, Nov. 2017.
- [8] Z. Fan, N. Peng, M. Tian, and B. Fain, "Welfare and fairness in multi-objective reinforcement learning," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1991–1999, International Foundation for Autonomous Agents and Multiagent Systems, 2023.
- [9] Y. Ge, X. Zhao, L. Yu, S. Paul, D. Hu, C.-C. Hsieh, and Y. Zhang, "Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, p. 316–324, ACM, Feb. 2022.
- [10] C. Wu, F. Wu, X. Wang, Y. Huang, and X. Xie, "Fairrec: Fairness-aware news recommendation with decomposed adversarial learning," 2021.
- [11] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang, "User-oriented fairness in recommendation," in *Proceedings of the Web Conference 2021*, WWW '21, p. 624–632, ACM, Apr. 2021.
- [12] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
- [13] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, pp. 1861–1870, Pmlr, 2018.
- [14] K. Wang, Z. Zou, M. Zhao, Q. Deng, Y. Shang, Y. Liang, R. Wu, X. Shen, T. Lyu, and C. Fan, "RI4rs: A real-world dataset for reinforcement learning based recommender system," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2935–2944, 2023.
- [15] C. Gao, S. Li, Y. Zhang, J. Chen, B. Li, W. Lei, P. Jiang, and X. He, "Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos," in *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, p. 3953–3957, ACM, Oct. 2022.
- [16] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, pp. 19:1–19:19, Dec. 2015.
- [17] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," 2018.
- [18] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, pp. 1928–1937, Pmlr, 2016.
- [19] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

Explainable Decomposed Reward Q-learning with Roulette Wheel Selection in Maze Navigation

1st Dac Hoang Nguyen
dept. Information Technology
Tay Do College
Can Tho, Vietnam
dacnguyen@caodangtaydo.edu.vn

2nd Nhu-Tai Do
dept. Information Technology
Saigon University
Ho Chi Minh, Vietnam
dntai@sgu.edu.vn

3rd Huy Quoc Nguyen
dept. Information Technology
Saigon University
Ho Chi Minh, Vietnam
nqhuy@sgu.edu.vn

Abstract—Enhancing interpretability in reinforcement learning remains a significant challenge, particularly in environments requiring efficient navigation. This paper introduces a method designed to increase the interpretability of Decomposed Reward Q-learning (DrQ) in maze navigation tasks. Unlike traditional approaches employing argmax or ϵ -greedy strategies, our method utilizes Roulette Wheel Selection (RWS) to probabilistically select actions based on decomposed Q-values, promoting balanced exploration and exploitation. To enhance explanations, our approach employs two distinct methods: Reward Difference Explanation (RDX), which clarifies how individual reward components influence decisions, and Minimal Sufficient Explanation (MSX), which identifies the minimal set of reward components necessary to justify an action. Experimental validation was performed in Grid Maze Navigation and CliffWorld environments, evaluating both interpretability and navigation efficiency. Results indicate that the DrQ-RWS method achieves performance comparable to or better than traditional Q-learning and Hybrid Reward Architecture (HRA), while significantly enhancing interpretability. Additionally, this paper explores the possibility of using RDX and MSX outputs to generate human-agent dialogues through the current ChatGPT platform.

Index Terms—Explainable Reinforcement Learning, Reward Decomposition, Roulette Wheel Selection policy

I. INTRODUCTION

Reinforcement Learning (RL) has become an influential approach for addressing sequential decision-making in complex environments, allowing agents to acquire optimal behaviors through interactions that aim to maximize cumulative rewards. Among recent advancements, Decomposed Reward Q-learning (DrQ) has been proposed as an effective approach that not only enhances learning in multi-objective environments but also inherently supports interpretability. While the widely adopted ϵ -greedy policy ensures action exploration, alternative methods such as Roulette Wheel Selection (RWS) can provide more balanced exploration and exploitation dynamics.

In this study, we extend the DrQ framework by incorporating an RWS policy and integrating two distinct explanation techniques:

- **Reward Difference Explanation (RDX):** Clarifies the influence of individual reward components on the agent's decisions.
- **Minimal Sufficient Explanation (MSX):** Identifies the smallest set of reward components necessary to justify a given action.

Our proposed approach is evaluated in both CliffWorld and maze navigation environments. Through empirical analysis using Q-value heatmaps and detailed explanation logs, we demonstrate that our methodology achieves comparable or superior performance compared to standard Q-learning and Hybrid Reward Architecture (HRA) [1], while significantly enhancing the interpretability of component-specific rationales behind agent behaviors. Additionally, we simulate human-agent interactions via ChatGPT-style dialogues to assess the clarity and relevance of the explanations generated.

II. LITERATURE REVIEW

A. Reward Decomposition

Reinforcement Learning (RL) is a field of Artificial Intelligence that focuses on training agents through interaction with the environment to maximize cumulative rewards over time [2]. It involves an agent learning to navigate an environment through a cycle of observing states, selecting actions, and receiving rewards.

Q-learning is a model-free RL algorithm that learns the optimal action-value function $Q^*(s, a)$, reflecting the anticipated cumulative reward for taking action a in state s and subsequently adhering to the optimal policy. The Q-value is updated iteratively using the Bellman equation:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[R(s, a, s') + \gamma \max_{a'} Q(s', a') - Q(s, a) \right], \quad (1)$$

where $\alpha \in (0, 1]$ is the learning rate, and s' is the next state. This update rule enables the agent to converge toward the optimal policy without requiring an explicit environment model.

In navigation tasks such as maze solving, Q-learning enables agents to identify optimal action strategies guided by reward signals. To improve the interpretability and transparency of agent behaviors, reward decomposition methods have been proposed. By partitioning the scalar reward into semantically meaningful components (e.g., goal-reaching, safety, efficiency), methods like Decomposed Q-learning (DrQ) [3] maintain separate Q-functions per component, ensuring convergence and enabling Minimal Sufficient Explanations (MSX). This facilitates clearer insight into agent behavior compared

to earlier architectures such as HRA [1]. Empirical studies have shown that RD enhances user understanding of agent decisions [4] and improves alignment in RL from Human Feedback (RLHF) scenarios by disentangling prompt-related rewards [5].

The selection probability $P(i)$ for individual i in a population of N individuals, each with fitness score f_i , is calculated as:

$$P(i) = \frac{f_i}{\sum_{j=1}^N f_j} \quad (2)$$

In the reinforcement learning (RL) setting, this concept can be extended to action selection, where each action's *fitness* is derived from its Q-value. Traditional greedy policies select the action with the maximum Q-value. In contrast, Roulette Wheel Selection (RWS) introduces stochasticity into the decision-making process, enabling agents to explore non-greedy actions based on their relative values.

To address this, Lipowski and Lipowska (2012) proposed a stochastic acceptance variant of RWS [6]. Instead of computing cumulative fitness scores, this approach randomly selects individuals and accepts them based on a scaled probability:

$$P_{\text{accept}}(i) = \frac{f_i}{f_{\text{max}}} \quad (3)$$

This results in an average selection time complexity of $O(1)$, significantly improving runtime efficiency in large-scale applications.

B. Interpretable Reinforcement Learning

Recent studies [7], [8] have aimed at improving the interpretability of RL policies, typically employing post-hoc explanations or inherently interpretable models. Although these methods significantly enhance policy transparency, they often rely on heuristic techniques and lack rigorous guarantees on explanation fidelity.

To enhance interpretability, we integrate two explanation methods that leverage decomposed Q-values: Reward Difference Explanations (RDX) and Minimal Sufficient Explanations (MSX).

RDX explains why the agent demonstrates a preference toward action a_1 over action a_2 at state s by computing the difference for each reward component:

$$\Delta_c(s, a_1, a_2) = Q_c(s, a_1) - Q_c(s, a_2) \quad (4)$$

Each RDX component reflects a positive or negative contribution:

- If $\Delta_c > 0$: component c supports action a_1 .
- If $\Delta_c < 0$: component c favors action a_2 .

Thus, RDX provides a detailed view of the influence of individual reward components on the agent's action selection.

MSX aims to simplify explanations by presenting only the most influential reward components. It consists of two parts:

(a) Positive Minimal Sufficient Explanation (MSX⁺):

First, define d as the total disadvantage of a_1 compared to a_2 :

$$d = \sum_{c \in C} \mathbb{I}(\Delta_c(s, a_1, a_2) < 0) \cdot |\Delta_c(s, a_1, a_2)| \quad (5)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

MSX⁺ identifies the smallest subset of components $M \subseteq C$ such that:

$$\text{MSX}^+ = \arg \min_{M \subseteq C} |M| \quad \text{subject to} \quad \sum_{c \in M} \Delta_c(s, a_1, a_2) > d \quad (6)$$

This subset captures the most critical positive contributions that justify favoring a_1 .

(b) Negative Minimal Sufficient Explanation MSX⁻:

Specify the threshold for insufficient advantage v as:

$$v = \sum_{c \in \text{MSX}^+} \Delta_c(s, a_1, a_2) - \min_{c \in \text{MSX}^+} \Delta_c(s, a_1, a_2) \quad (7)$$

MSX⁻ identifies the smallest subset of negative components $M \subseteq C$ such that:

$$\text{MSX}^- = \arg \min_{M \subseteq C} |M| \quad \text{subject to} \quad \sum_{c \in M} -\Delta_c(s, a_1, a_2) > v \quad (8)$$

If all $\Delta_c > 0$, then $\text{MSX}^- = \emptyset$, indicating that action a_1 is superior across all reward dimensions.

III. METHODOLOGY

Maze navigation tasks provide a rich and interpretable domain for RL due to their spatial complexity and multi-objective nature. This study proposes an interpretable reinforcement learning (XRL) approach using Decomposed Reward Q-learning (drQ) and Roulette Wheel Selection (RWS), enhanced with Reward Difference Explanations (RDX) and Minimal Sufficient Explanations (MSX) in Maze Navigation Tasks. The goal is to address the limitations of traditional reinforcement learning algorithms in terms of transparency and decision justification.

A. Decomposed Reward Q-Learning (drQ)

The Decomposed Reward Q-learning (DrQ) addresses these shortcomings by incorporating off-policy Q-learning updates into its framework.

Traditional Q-learning methods learn a scalar Q-function $Q(s, a)$. In contrast, drQ enhances interpretability by learning a vector of separate Q-functions, one per reward component $c \in C$, such as goal, safety, blocked, and turn components. Each of these Q-values is updated independently. The total Q-value used for decision-making is computed as:

$$Q_{\text{total}}(s, a) = \sum_{c \in C} Q_c(s, a) \quad (9)$$

This method transforms the multi-objective formulation into a scalar single objective signal that can be used for action selection. This design is well suited to balancing multiple goals: the default uniform aggregation treats objectives equally, whereas task-specific trade-offs can be encoded

through weights, allowing the policy to emphasize or down-weight particular objectives without discarding component-level attributions that support post hoc explanation.

$$Q_{\text{total}}(s, a) = \sum_{c \in C} w_c Q_c(s, a), \quad w_c \geq 0, \quad (10)$$

In standard DRQ, the agent determines its action through an ε -greedy policy as follows:

$$a^* = \begin{cases} \text{random action,} & \text{with probability } \epsilon \\ \arg \max_a Q_{\text{total}}(s, a), & \text{otherwise} \end{cases} \quad (11)$$

However, this approach may lead to suboptimal convergence due to insufficient exploration. To address this, we incorporate RWS to enable probabilistic action selection based on Q-values, thereby enhancing exploration while still favoring high-reward actions. In the DRQ-RWS framework, at each decision point, the agent computes the selection probability for each action based on their total Q-values:

- 1) For state s , compute total Q-values $Q_{\text{total}}(s, a)$ for all possible actions $a \in A$.
- 2) Normalize Q-values into probabilities:

$$P(a_i) = \frac{Q_{\text{total}}(s, a_i)}{\sum_j Q_{\text{total}}(s, a_j)} \quad (12)$$

- 3) Sample action $a \sim P(a)$ using roulette wheel selection. Higher Q-values have higher probabilities, but all actions are eligible.
- 4) This encourages more robust exploration while remaining reward-aligned.

B. Explanation Mechanisms

To evaluate and compare two actions a_1 and a_2 in a given state s , Reward Difference Explanation (RDX) computes:

$$\Delta_c(s, a_1, a_2) = Q_c(s, a_1) - Q_c(s, a_2) \quad (13)$$

Interpretation:

- $\Delta_c > 0$: Component c favors action a_1 .
- $\Delta_c < 0$: Component c favors action a_2 .

This computation provides a detailed decomposition of the decision rationale across all components.

Although RDX provides detailed insights, interpreting many reward types can be overwhelming. To address this issue, Minimal Sufficient Explanation (MSX) is employed to highlight the most significant factors justifying an action choice. MSX proceeds as follows:

First, define d , the total disadvantage of action a_1 compared to a_2 :

$$d = \sum_{c \in C} I(\Delta_c(s, a_1, a_2) < 0) \cdot |\Delta_c(s, a_1, a_2)| \quad (14)$$

a) (a) *Minimal Positive Sufficient Explanation (MSX⁺)*: MSX⁺ identifies the smallest subset of positive components $M \subseteq C$ satisfying:

$$MSX^+ = \arg \min_M |M| \quad \text{subject to} \quad \sum_{c \in M} \Delta_c(s, a_1, a_2) > d \quad (15)$$

This subset highlights the most critical positive contributions justifying the selection of a_1 .

b) (b) *Negative Minimal Sufficient Explanation (MSX⁻)*: Define the insufficient advantage threshold v as:

$$v = \sum_{c \in MSX^+} \Delta_c(s, a_1, a_2) - \Delta_c(s, a_1, a_2) \quad (16)$$

MSX⁻ then identifies the smallest subset of negative components $M \subseteq C$ such that:

$$MSX^- = \arg \min_M |M| \quad \text{subject to} \quad \sum_{c \in M} -\Delta_c(s, a_1, a_2) > v \quad (17)$$

If $\Delta_c > 0$ for all components, then $MSX^- = \emptyset$, indicating that action a_1 is superior across all reward components.

C. Implementation Procedure

To enhance both exploration and interpretability in reinforcement learning, we extend the Decomposed Reward Learning (DRQ) framework by integrating Roulette Wheel Selection (RWS). This approach, termed DRQ-RWS, replaces the standard greedy or ε -greedy policy with a probabilistic mechanism that selects actions based on their relative desirability derived from decomposed Q-values.

Q-Value Decomposition

At each state s , the agent maintains separate Q-tables $Q_c(s, a)$ corresponding to distinct reward components $c \in \{\text{goal, safe, blocked, turn}\}$. These are updated independently as follows:

$$Q_c(s, a) \leftarrow Q_c(s, a) + \alpha \left[r_c + \gamma \max_{a'} Q_c(s', a') - Q_c(s, a) \right] \quad (18)$$

The total Q-values used for action selection is computed by summing over all components:

$$Q_{\text{total}}(s, a) = \sum_c Q_c(s, a) \quad (19)$$

Roulette Wheel Selection (RWS) for Action Sampling

Rather than selecting the action with the highest $Q_{\text{total}}(s, a)$, we employ RWS to probabilistically sample actions:

$$P(a_i | s) = \frac{Q_{\text{total}}(s, a_i)}{\sum_j Q_{\text{total}}(s, a_j)} \quad (20)$$

To avoid division by zero or negative values, we normalize Q-values by ensuring a minimum threshold $Q_{\text{total}}(s, a) \geq \epsilon$. This method preserves the ranking of actions while promoting exploration proportional to learned utility. The training loop proceeds as follows:

- 1) **Initialize** all Q-tables for each reward component (*goal, safe, blocked, turn*).

2) For each episode:

- Start at the initial state.
- At each time step:
 - Compute the total Q-value $Q_{\text{total}}(s, a)$ for all actions.
 - Normalize these values to a probability distribution using Roulette Wheel Selection (RWS).
 - Sample an action based on the probabilities.
 - Execute the action, receive decomposed rewards, and transition to the next state.
 - Update each Q-table independently using temporal difference learning.
 - Log explanations (RDX and MSX) for future analysis.

3) Repeat until the maximum number of episodes is reached or convergence is observed.

a) *Maze Environment Design:* The maze environment is represented as a grid-based structure (e.g., 10×10, 20×20). Each maze includes:

- **Start point:** dark green circle at coordinates (0,0).
- **Goal point:** blue X located variably (e.g., 10×10, 20×20).
- **Walls:** strategically placed to form safe pathways, traps, and dead-ends.
- Each cell permits 1–4 valid neighboring actions.

Each state is defined as a tuple (x, y) , with rewards calculated based on proximity to the goal and the consequences of movements.

b) *Interpretable Components:* Reward decomposition separates the total reward into clearly interpretable components to facilitate effective learning:

- **Positive Rewards:**
 - +30 for successfully reaching the goal.
 - +2 for every two consecutive safe moves.
- **Negative Penalties:**
 - -1 for encountering a blocked cell.
 - -1 for reversing the direction of movement (turning back).

This clearly defined reward structure enhances the agent’s learning by explicitly highlighting the consequences of each action.

D. Experimental result

To investigate the impact of probabilistic action selection, we run the DrQ+RWS in the 20×20 grid-based maze. Figure 1 illustrates the learning curves over 3000 episodes. DRQ+RWS achieved rapid convergence, maintaining a stable total reward around -11 after approximately 600 episodes. This result shows that the agent selected actions based on a probability distribution proportional to total Q-values at each state, thereby maintaining alignment with high-reward behaviors while promoting exploratory flexibility.

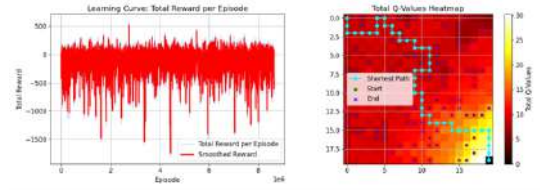


Fig. 1. Heatmap of total Q-values for DRQ+RWS.

IV. EVALUATION

Our evaluation covers three main contributions: (1) integrating Roulette Wheel Selection (RWS) into Decomposed Reward Q-learning (DRQ) for better exploration in maze navigation; (2) utilizing XAI techniques (RDX and MSX) to provide human-interpretable explanations; and (3) enabling human-agent interaction through ChatGPT-generated dialogues.

A. DrQ+RWS Evaluation

We compare the performance of three algorithms: Q-learning, Hybrid Reward Architecture (HRA), and Decomposed Reward Q-learning with Roulette Wheel Selection (DrQ+RWS) in the 4×12 CliffWorld environment over 3000 episodes, averaged across five runs. drQ+RWS consistently outperformed others, achieving near-optimal rewards (-13) within 400 episodes by using reward decomposition and probabilistic action selection. Q-learning improved gradually but had higher variance and lacked interpretability. HRA showed unstable learning due to deterministic action selection, frequently overestimating Q-values. The result highlights the advantages of our approach in convergence speed and interpretability in risky environments. Figure 2 demonstrates the performance of each algorithm. Hyperparameters included:

- Learning rate $\alpha = 0.1$
- Discount factor $\gamma = 0.99$
- ϵ -greedy strategy with ϵ decaying from 1.0 to 0.01
- Total of 3000 training episodes

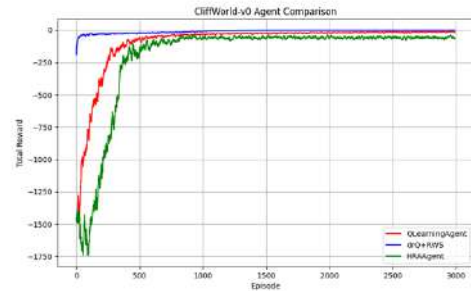


Fig. 2. Timing performance curves among DrQ + RWS, Q-learning, and HRD agents in CliffWorld.

Additionally, in order to evaluate the impact of the RWS mechanism on the agent’s navigation performance, we conducted a comparative analysis between it and ϵ -greedy policy. Both models were trained under identical conditions with

1000, 2000, and 3000 episodes on the same maze configuration. The result in Table I shows that applying (RWS) improves agent performance during early training. With 1000 episodes, DRQ+RWS outperformed ϵ -greedy one in final reward, steps to goal, and turn ratio. However, as training extended to 3000 episodes, the performance gap narrowed. This suggests that RWS is effective for enhancing early exploration but may become less beneficial over time due to its inherent randomness in action selection.

TABLE I
PERFORMANCE COMPARISON BETWEEN DRQ WITH ϵ -GREEDY AND RWS POLICY

Episodes	Model	Final	Avg (All)	Avg (Last 100)	Steps	Turns	Turn Ratio
1000	ϵ -greedy	-107.5	-123.08	-134.77	2984	797	0.267
	RWS	-79.0	-108.94	-118.09	1102	279	0.253
2000	ϵ -greedy	-270.5	-115.78	-128.56	3346	907	0.271
	RWS	-193.0	-116.10	-139.65	3308	886	0.268
3000	ϵ -greedy	-639.0	-118.35	-120.28	7142	1879	0.263
	RWS	-343.0	-121.56	-104.03	4768	1258	0.264

B. Explanation Evaluation

To evaluate the impact of explanation mechanisms on policy learning, we compare two variants of DrQ+RWS with and without explanation (RDX and MSX). As illustrated in Figure 3, the algorithm with no explanation demonstrates rapid convergence, with its smoothed reward stabilizing near 120 after just 3000 episodes. This indicates that even without explicit explanation mechanisms, the combination of decomposed Q-values and stochastic action selection via RWS supports efficient policy learning.

In contrast, Figure 4 shows that the explanation variant learn more slowly. Despite prolonged training over 2.5 million episodes, its smoothed reward remains around 200, accompanied by higher variance. This behavior reflects the influence of MSX and RDX, which introduce additional constraints to promote interpretability and safe action selection potentially at the expense of short-term reward optimization.

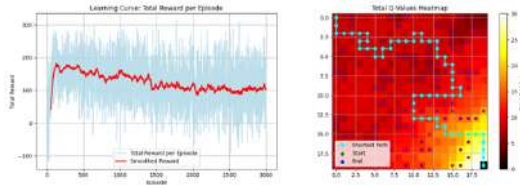


Fig. 3. DrQ+RWS Without Explanation.

C. Human Interpretability Evaluation

To further evaluate the interpretability of our DrQ+RWS framework, we employ ChatGPT as a natural language interface to translate the agent's internal reasoning into human-understandable explanations. Specifically, we provide ChatGPT with both the maze description and the result file,

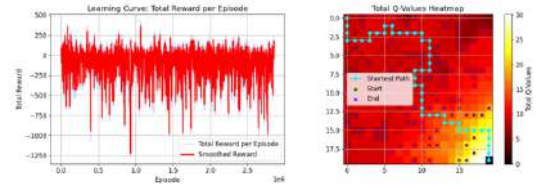


Fig. 4. DrQ+RWS With Explanation.

accompanied by the initial instruction prompt that outlines how to interpret and explain the agent's decisions.

Once initialized, the system allows users to pose queries in natural language, such as:

User Query: "Why did the agent go right at (14,12) instead of down?"

ChatGPT, using the MSX and RDX framework, is then able to generate a response such as:

ChatGPT Answer:

At state (14,12), the agent selected **Right** (action = 3) over **Down** (action = 2). Here's a detailed analysis:

Action Decision:

- **Chosen:** Right (3)
- **Alternative:** Down (2)

RDX Component Differences ($\Delta = Q(\text{Right}) - Q(\text{Down})$):

- $\Delta_{\text{goal}} = +1.79 \rightarrow$ Right advances goal progress
- $\Delta_{\text{turn}} = +0.47 \rightarrow$ Right is more efficient in turning
- $\Delta_{\text{blocked}} = +0.80 \rightarrow$ Right avoids blockages better
- $\Delta_{\text{safe}} = -0.45 \rightarrow$ Right is less safe

Net gain: +2.60 total Q-value (Right is strongly preferred in 3 out of 4 reward components).

MSX Explanation:

- $\text{MSX}^+ = [\text{'goal'}] \rightarrow$ Goal progress is the key component justifying the choice.
- $\text{MSX}^- = [\text{'safe'}] \rightarrow$ Safety was compromised but considered less critical in this decision.

RWS Probabilities (total_P):

```
{
  "Up": 0.2744,
  "Down": 0.0000,
  "Left": 0.1364,
  "Right": 0.5892
}
```

The action **Right** held the highest selection probability ($\sim 58.9\%$), while **Down** was not considered (0%).

Conclusion

The agent chose Right at (14,12) over Down because:

- It offered a major gain in goal reward (+1.79),
- Reduced turn cost and avoided blocks,
- Was most probable under Roulette Wheel Selection ($\approx 59\%$),
- Despite a slight loss in safety.

This interaction demonstrates that, when guided by component-wise explanations, ChatGPT can effectively translate the agent’s internal reasoning into coherent and comprehensible natural language summaries. The explanation reflects the agent’s reward trade-offs, maintains fidelity to the underlying data, and is accessible to non-expert users. This highlights the potential of combining interpretable reinforcement learning with large language models to enhance transparency and user understanding in decision-making systems.

V. CONCLUSION

The integration of Roulette Wheel Selection with Decomposed Reward Q-learning demonstrates a promising policy design for navigation tasks in maze environments. The probabilistic selection driven by decomposed reward structures not only mitigates local optima and promotes more effective exploration, but also facilitates meaningful explanations through RDX and MSX. This combined framework provides both improved navigation performance and enhanced interpretability, indicating its potential applicability in reinforcement learning domains where decision transparency is essential.

REFERENCES

- [1] H. van Seijen, M. Fatemi, J. Romoff, R. Laroché, T. Barnes, and J. Tsang, “Hybrid reward architecture for reinforcement learning,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5398–5408.
- [2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.
- [3] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, “Explainable reinforcement learning via reward decomposition,” 2019.
- [4] Y. Septon, T. Huber, E. André, and O. Amir, “Integrating policy summaries with reward decomposition for explaining reinforcement learning agents,” in *Advances in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection: 21st International Conference, PAAMS 2023, Guimarães, Portugal, July 12–14, 2023, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 320–332. [Online]. Available: https://doi.org/10.1007/978-3-031-37616-0_27
- [5] L. Mao, H. Xu, A. Zhang, W. Zhang, and C. Bai, “Information-theoretic reward decomposition for generalizable rlhf,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.06020>
- [6] A. Lipowski and D. Lipowska, “Roulette-wheel selection via stochastic acceptance,” *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 6, pp. 2193–2196, Jun 2012. [Online]. Available: <https://doi.org/10.1016/j.physa.2011.12.004>
- [7] S. Milani, N. Topin, M. Veloso, and F. Fang, “Explainable reinforcement learning: A survey and comparative review,” *ACM Comput. Surv.*, vol. 56, no. 7, Apr. 2024. [Online]. Available: <https://doi.org/10.1145/3616864>
- [8] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, “Explainability in deep reinforcement learning,” vol. 214, p. 106685. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705120308145>

Reasoning about Coalitional Ability in Network Threshold Models

Thomas Ågotnes^{*†}, Zoé Christoff[‡]

^{*} *Department of Information Science and Media Studies, University of Bergen*

Bergen, Norway

thomas.agotnes@uib.no

[†] *School of Philosophy, Shanxi University*

Taiyuan, China

[‡] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen*

Groningen, The Netherlands

z.l.christoff@rug.nl

Abstract—This paper proposes a formal logic for reasoning about the ability of groups of agents in social networks when it comes to triggering cascades. We focus on threshold models, where agents are influenced into adopting a new product or behaviour as soon as enough of their direct network-neighbours have adopted it already. We introduce the syntax and semantics and illustrate how it captures the ability of (groups of) agents to trigger cascading effects in the network. We study the basic principles of reasoning about this kind of coalitional ability, and the relationship to Coalition Logic.

Index Terms—Logic-based knowledge representation, Reasoning, Threshold models, Coalitional ability

I. INTRODUCTION

In this paper we lay the groundwork for a framework for logic-based knowledge representation and reasoning about the ability of agents and groups of agents to trigger cascading behaviour in social networks. This allows us to study the fundamental logical principles of such ability, and opens the door to the use of formal specification and verification tools for checking, e.g., the level of influence of particular agents in a social network.

When modeling how new technologies, trends, opinions, behaviours, or products spread in social networks, *threshold models* [11] are one of the most ubiquitous types of models. In such models, agents are assigned influenceability thresholds, representing how much conformity pressure is needed to influence them into adopting the new behaviour. According to the corresponding *threshold rule*, an agent will adopt the new behaviour at the next time-step if the fraction of her neighbours who have adopted it already meets her influenceability threshold. For instance, if her threshold is $\frac{1}{2}$, she will be influenced into adopting the new behaviour as soon as at least half of her direct neighbours in the social network have.

Threshold models are synchronous, in the sense that all agents update their behaviour at the same time, and deterministic. Moreover, they are typically restricted to a finite, unchanging non-weighted, undirected, and irreflexive network of influence, coupled with a unique spreading behaviour, that can never be unadopted, and a uniform threshold among all

agents. Nevertheless, there are variants in the literature which relax some of those simplifying assumptions (see, for instance, [16], [14], [12], [4]). In this paper, while we will restrict ourselves to deterministic and synchronous influence on finite, unchanging, non-weighted, undirected and irreflexive networks too, we allow for several behaviours to spread in parallel and for each agent to have a different threshold for each behaviour.

The step-by-step social influence process in threshold models, generated by the iterated application of the above threshold rule, is known to correspond to the best response dynamics of an iterated neighbours-coordination network game (see [9, Chapter 19] or [6, Section 2] for more detail). The resulting snow-balling diffusion dynamics, or “adoption epidemics”, is typically described as a *cascade*, and when an initial group of adopters triggers *all* agents in the network into adopting the new trend, the cascade is said to be *complete*. Otherwise, the cascade is said to be *partial* or *limited*.

The following well-known “cluster-cascade theorem” characterises the conditions for complete cascades: they occur if and only if there is no cluster of non-adopters of density greater than 1 minus the threshold ([14], [9, Chapter 19.3]). For instance, if the (uniform) threshold is $\frac{1}{2}$, any group of agents which have not adopted the behaviour and each have three neighbours in the group but only two outside of the group, can never be triggered into adopting the behaviour.

As illustrated by this theorem, when reasoning about cascades in social networks, the *ability of a group of individuals to trigger a cascade* is a key issue. For example, if we know which agents are able to trigger a complete cascade, or can influence, through a limited cascade, a certain agent whose behaviour we are interested in affecting, then we can focus our entire attention (say, e.g., our marketing budget, or propaganda effort) on getting those very agents to become early adopters.

Consider the network in Figure 1 below as an example, and assume that the threshold rule is applied iteratively (and synchronously) with a uniform threshold of $\frac{1}{2}$. If agents j and k are the first adopters of a new behaviour, they will first trigger group $\{i, l\}$ into adopting the behaviour too, who triggers m , who triggers p , who triggers o , who triggers n , who

triggers s , who triggers r , who triggers q , who triggers h , who finally triggers g . The second part of the figure shows the result of group $\{j, k\}$ initially adopting the new behaviour, after iteratively applying the threshold rule until nothing changes anymore. The resulting cascade is not complete. In light of the cluster-cascade theorem above, this is not surprising a result, since everybody in the the group $\{a, b, c, d, e, f\}$ has more friends inside the group than outside, they can never be influenced by only outsiders into adopting the new behaviour.

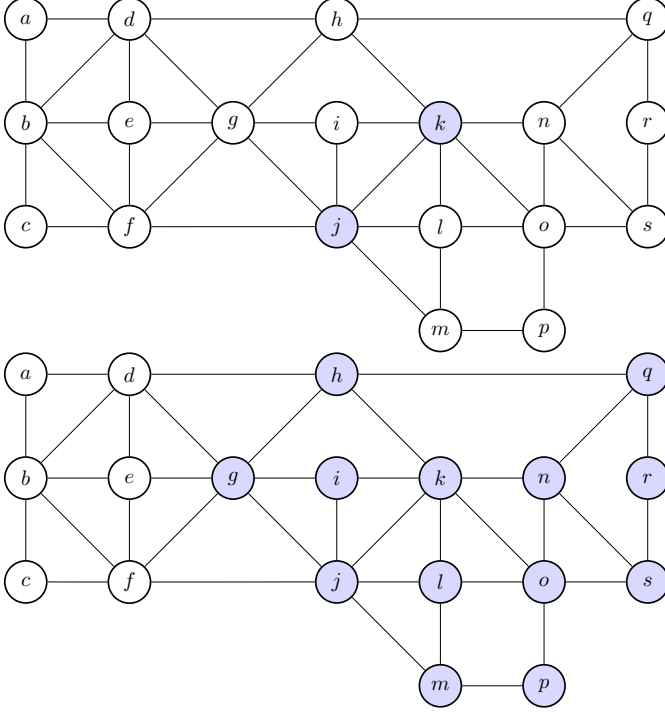


Figure 1. A network with 19 agents. The threshold for adopting a new behaviour p is $\theta(a, p) = \frac{1}{2}$ for all agents a . The top model represents the state where only j and k have adopted the new behaviour, while the bottom model represents the result of applying the threshold rule on the top model, iteratively, until nothing changes anymore.

In terms of the *cascading ability* of groups, we can make the following observations:

- $\{j, k\}$ has the ability to make agent g change her behaviour, i.e., if j and k adopt the new behaviour, it will trigger a (limited) cascade that will make g adopt it too.
- $\{j, k\}$ has the ability to make all direct neighbours of n change their behaviour.
- $\{j, k\}$ does not have the ability to make all direct neighbours of g change their behaviour.
- $\{j, k\}$ has the ability to make all agents in set $\{g, h, i, j, k, l, m, n, o, p, q, r, s\}$ adopt the behaviour but none of the agents in set $\{a, b, c, d, e, f\}$.
- $\{j, k\}$ does not have the ability to trigger a complete cascade.

Note that, while the initial group of adopters $\{j, k\}$ cannot trigger a complete cascade on its own, adding just one initial adopter would be sufficient to trigger one:

- The group $\{d, j, k\}$ has the ability to trigger a complete cascade.

In this paper we introduce a formal logical language for expressing properties such as those above. To formally capture the notion of ability described above, we will use *coalition ability operators* of the form $\langle C \rangle \varphi$, known from, e.g., Coalition Logic [15] and Alternating-time Temporal Logic [3], where φ is a *cascading goal*, e.g., a complete cascade, with the intuitive meaning that group C has the ability to make φ come true, by being initial adopters of certain behaviours. This will allow us to reason about the ability (or *power*) of (groups of) agents in social networks when it comes to triggering cascades by adopting certain behaviours. If we know that $\langle C \rangle \varphi$ holds, then we know that we can achieve the goal φ by controlling initial adoption by agents in C , for instance by focussing our advertisement budget on them.

The rest of the paper is structured as follows: Section II introduces the syntax of our logic, while Section III gives its semantics. Section IV then illustrates the meaning of our formulae with some examples. Section V introduces some validities and validity preserving principles of the newly defined system. The semantics we have given to $\langle C \rangle \varphi$ is that C can adopt behaviours in such a way that φ will be true. In Section VI we define and study a derived operator $\llbracket C \rrbracket \varphi$ with the meaning that C can adopt behaviours in such a way that φ will be true *no matter how the other agents adopt behaviours* – similarly to in Coalition Logic. Finally, Section VII discusses related work and directions for further research.

II. LANGUAGE

The language is parametrised by a finite set N of *agents*, and a set Π of *behaviours* and is defined as follows.

$$\varphi ::= p \mid \neg \varphi \mid \varphi \wedge \varphi \mid \Box \varphi \mid \Box^* \varphi \mid \langle C \rangle \varphi$$

where $p \in \Pi$ and $C \subseteq N$. We use the usual derived propositional connectives, in addition to $\llbracket C \rrbracket \varphi$ for $\neg \langle C \rangle \neg \varphi$ and $\Diamond \varphi$ for $\neg \Box \neg \varphi$. When $C = \{a\}$ is a singleton we sometimes abuse notation and write $\langle a \rangle$ instead of $\langle \{a\} \rangle$.

These formulas will be interpreted as propositions about a given (“current”) agent in a model. Intuitively, p means that the current agent has adopted behaviour p ; $\Box \varphi$ that all of the current agent’s neighbours satisfy φ ; $\Box^* \varphi$ that all agents reachable through a chain (of zero or more links) of neighbours from the current agent satisfy φ ; $\langle C \rangle \varphi$ that C has the ability to make φ true in the following sense: C can adopt behaviours in such a way that φ is guaranteed to be true after the potentially resulting cascade.

III. SEMANTICS

The class of models is also parametrised by N and Π . A model over N and Π is a tuple $M = (R, \theta, V)$ where $R \subseteq N \times N$ is a binary relation on the set of agents representing social connections, such that R is symmetric, irreflexive and connected¹, θ is a threshold function mapping each agent $a \in$

¹There is a path between any pair of nodes. Since we are primarily interested in cascades we assume connectedness.

N and behaviour $p \in \Pi$ to a rational number such that $0 \leq \theta(a, p) \leq 1$ representing agent a 's threshold with respect to adopting behavior p , and $V : \Pi \rightarrow 2^N$ is a valuation function saying which agents currently have adopted which behaviours.

A main concern in defining the semantics of the language is *model updates* caused by, first, a group of agents changing one or more of their behaviours, and, second, the possible cascade that this triggers, under iterative applications of the threshold rule.

We write $V \sim_C V'$ to say that valuation functions V and V' agree for all agents *except* $C \subseteq N$:

$$V \sim_C V' \Leftrightarrow \forall p \in \Pi \forall a \in N \setminus C \\ (a \in V(p) \Leftrightarrow a \in V'(p))$$

Intuitively, in a model $M = (R, \theta, V)$, C can unilaterally change the valuation function to a V' such that $V' \sim_C V$.

Given a model $M = (R, \theta, V)$, a behaviour p and a set of agents C , we are interested in finding the set of agents adopting p after a possible cascade caused by C adopting p . Let $\text{cascade}(R, \theta, B, p)$ be the set of agents who will *eventually* adopt behaviour p if exactly the agents in B initially adopt p , in a network with relation R and threshold assignment θ , through iterated applications of the threshold rule until nothing changes anymore. For example, if N and R are the sets of agents and the relation in Figure 1 and $\theta(a, p) = \frac{1}{2}$ for all $a \in N$, then $\text{cascade}(R, \theta, \{j, k\}, p) = \{g, h, i, j, k, l, m, n, o, p, q, r, s\}$.

More precisely, the function $\text{cascade}(R, \theta, B, p)$ returns the fixed point of the diffusion sequence obtained by iteratively applying the threshold rule wrt. θ in a network with relation R , where the set of agents who have adopted p is B . Let us write $R(a)$ to denote the set of direct neighbours of a , i.e., $R(a) = \{b \in N \mid (a, b) \in R\}$. We define $\text{cascade}(R, \theta, B, p)$ in terms of $\text{step}^n(R, \theta, B, p)$, the set of agents who will have adopted p after exactly n applications of the threshold rule:

$$\begin{aligned} \text{step}^0(R, \theta, B, p) &= B \\ \text{step}^{n+1}(R, \theta, B, p) &= \text{step}^n(R, \theta, B, p) \cup \\ &\quad \{a \in N : \frac{|R(a) \cap \text{step}^n(R, \theta, B, p)|}{|R(a)|} \geq \theta(a, p)\} \\ \text{cascade}(R, \theta, B, p) &= \bigcup_{i \geq 0} \text{step}^i(R, \theta, B, p) \end{aligned}$$

We say that a model $M = (R, \theta, V)$ is *stable* if it does not change under application of the threshold rule, i.e., if for all $p \in \Pi$, $V(p) = \text{cascade}(R, \theta, V(p), p)$.

When $M = (R, \theta, V)$, we write $M^\dagger$ to denote the updated model where all cascades (for all atomic behaviours) have played out. Formally,

$$M^\dagger = (R, \theta, V') \quad V'(p) = \text{cascade}(R, \theta, V(p), p)$$

for any $p \in \Pi$. Note that $M^\dagger$ is a stable model.

Satisfaction of a formula φ at an agent $a \in N$ in a model $M = (R, \theta, V)$ is then defined as follows.

$$\begin{aligned} M, a \models p &\quad \text{iff} \quad a \in V(p) \\ M, a \models \neg \varphi &\quad \text{iff} \quad M, a \not\models \varphi \\ M, a \models \varphi \wedge \psi &\quad \text{iff} \quad M, a \models \varphi \text{ and } M, a \models \psi \\ M, a \models \Box \varphi &\quad \text{iff} \quad \forall b \in N : aRb \Rightarrow M, b \models \varphi \\ M, a \models \Box^* \varphi &\quad \text{iff} \quad \forall b \in N : aR^*b \Rightarrow M, b \models \varphi \\ M, a \models \langle C \rangle \varphi &\quad \text{iff} \quad \exists V' \sim_C V, (R, \theta, V')^\dagger, a \models \varphi \end{aligned}$$

where R^* is the reflexive, transitive closure of R . It is easy to see that $M, a \models \langle C \rangle \varphi$ iff $\forall V' \sim_C V, (R, \theta, V')^\dagger, a \models \varphi$.

As usual, we write $M \models \varphi$ when $M, a \models \varphi$ for all $a \in N$. It is the class of *stable* models we are interested in. We thus say that φ is *valid*, and write $\models \varphi$, when $M \models \varphi$ for all stable models M .

$\langle C \rangle \varphi$ means that C can change their valuation (of zero or more behaviours) such that φ is guaranteed to be true; $\langle C \rangle$ that no matter which behaviours C adopt, φ will be true.

The meaning of $\langle C \rangle \varphi$ plays a natural role when reasoning about the ability of agents in a social network to trigger cascades: if we can control adoption by agents in C , the goal φ is guaranteed to be true. Note that there are potentially many possible cascades that a coalition C can make come about, by making different choices. The quantification over choices here range over the combinations of agents from C and behaviours $p, q, \dots$ mentioned in the formula φ . This semantics can alternatively and equivalently be stated as follows. When $D_1, \dots, D_m \subseteq N$, let $M^{D_1 \hookrightarrow p_1, \dots, D_m \hookrightarrow p_m}$ be the result of all agents in D_1 adopting p_1 , D_2 adopting p_2 , and so on, in M , and the cascade that potentially follows. Formally, when $M = (R, \theta, V)$, $M^{D_1 \hookrightarrow p_1, \dots, D_m \hookrightarrow p_m} = (R, \theta, V')$ where for any $1 \leq i \leq m$, $V'(p_i) = \text{cascade}(R, \theta, D_i \cup V(p), p)$ and $V'(q) = V(q)$ for $q \notin \{p_1, \dots, p_m\}$. We have the following. For any formula φ , let Π^φ be the (finite) set of behaviours occurring in φ .

Lemma 1 For any M, a, C and φ with $\Pi^\varphi = \{p_1, \dots, p_m\}$,

$$M, a \models \langle C \rangle \varphi \text{ iff } \exists D_1, \dots, D_m \subseteq C \\ \text{such that } M^{D_1 \hookrightarrow p_1, \dots, D_m \hookrightarrow p_m}, a \models \varphi.$$

Note that, since we have assumed symmetry and connectedness, R^* in fact plays the role as the universal modality: $M, a \models \Box^* \varphi$ iff $M, b \models \varphi$ for all $b \in N$ iff $M \models \varphi$. In particular, $\Box^* p$ means that everyone has adopted the behaviour p , e.g., the result of a complete cascade.

IV. EXAMPLES

Let N be the set of agents in Figure 1, $p \in \Pi$, and $M = (R, \theta, V)$ be such that R is the relation in the figure, $\theta(x, p) = \frac{1}{2}$ for all $x \in N$ and $V(p) = \emptyset$ (i.e., where no agents have adopted behaviour p yet). We have that:

- $M, i \models \langle \{j, k\} \rangle p$: j and k have the joint ability to make i adopt p .
- $M, i \models \langle \{j, k\} \rangle \Box p$: j and k have the joint ability to make all of i 's neighbours adopt p .
- $M, i \models \langle j \rangle (\langle k \rangle \Box p \wedge \langle k \rangle \neg \Box p)$: j is able to give k the ability to choose whether all of i 's friends adopt p .
- $M, i \models \neg \langle \{j, k\} \rangle \Box \Box p$: j and k do not have the joint ability to make all neighbours of all of i 's neighbours adopt p .
- $M, i \models \neg \langle \{j, k\} \rangle \Box^* p$: j and k do not have the joint ability to trigger a complete cascade.
- $M, i \models \langle \{d, j, k\} \rangle \Box^* p$: d, j, k have the joint ability to trigger a complete cascade.

V. LOGICAL PROPERTIES

We now move on to investigate universally valid properties expressible in the language, in terms of valid formulae (due to lack of space we have to leave out proofs).

Theorem 1 *The following are valid/validity preserving.*

- 1) $[C](\varphi \rightarrow \psi) \rightarrow ([C]\varphi \rightarrow [C]\psi)$ (*K*)
- 2) $\langle C \rangle \top$ (*D*)
- 3) $\langle C \rangle \varphi \rightarrow \langle D \rangle \varphi$, when $C \subseteq D$ (*Coalition Monotonicity*)
- 4) $\langle C \rangle \langle D \rangle \varphi \leftrightarrow \langle C \cup D \rangle \varphi$
- 5) $\langle \emptyset \rangle \varphi \leftrightarrow \varphi$
- 6) $\langle C \rangle [H] \varphi \rightarrow [H] \langle C \rangle \varphi$ (*Church-Rosser generalised*)
- 7) *If φ is valid, then $[C]\varphi$ is valid (Nec)*
- 8) $\langle C \rangle (\varphi \vee \psi) \leftrightarrow (\langle C \rangle \varphi \vee \langle C \rangle \psi)$
- 9) $\neg \langle C \rangle \perp$
- 10) $\langle C \rangle (\varphi \wedge \psi) \rightarrow \langle C \rangle \varphi$
- 11) $\langle C \rangle [C] \varphi \rightarrow [C] \langle C \rangle \varphi$ (*Church-Rosser*)

K says that $[C]$ distributes over implication (and it follows that it distributes over conjunction as well). K and Nec are standard in normal modal logics. D is the well-known seriality axiom; here it means that any coalition can always make some choice (including the choice of doing nothing). Coalition Monotonicity says that anything a coalition can do, any superset can do as well (the additional members just need to do nothing), it is a well-known axiom of, e.g., Coalition Logic [15]. Validity (4) says that two sequential choices by two coalitions are equivalent to a single joint choice by the combined coalition (even if they are overlapping). Validity (5) says that the empty coalition can do nothing. (6) is a generalised variant of the Church-Rosser property (which follows as an instance). The four last validities are derivable from the others.

One commonly considered property of general coalitional ability operators, like the ones in Coalition Logic, is so-called superadditivity: $(\langle C_1 \rangle \varphi_1 \wedge \langle C_2 \rangle \varphi_2) \rightarrow \langle C_1 \cup C_2 \rangle (\varphi_1 \wedge \varphi_2)$ when $C_1 \cap C_2 = \emptyset$ (when the two coalitions overlap it is easy to see that the formula doesn't hold, because in that case achieving φ_1 and φ_2 might require different actions for the same agent). In our case, this doesn't hold even when $C_1 \cap C_2 = \emptyset$: let M be the model in Figure 2 (left) where no-one has adopted yet; we have that $M, a \models \langle c \rangle p \wedge \langle i \rangle \neg p$ but $M, a \not\models \langle \{c, i\} \rangle (p \wedge \neg p)$. This is not a surprise: our modalities are modal diamonds ($\exists$) while Coalition Logic modalities are combined diamonds and boxes ($\exists \forall$). In the next section we consider a variant with the latter type of semantics, for which superadditivity does hold.

There are also other validities, for *atomic* behaviours.

Theorem 2 *The following are valid (for any $p, q \in \Pi$).*

- 1) $\langle N \rangle p$
- 2) $\langle C \rangle \neg p \leftrightarrow \neg p$
- 3) $\langle C \rangle (p \wedge q) \leftrightarrow (\langle C \rangle p \wedge \langle C \rangle q)$

The first validity says that the grand coalition can always make the current agent adopt p (if everyone adopts p then the current agent's threshold will necessarily be met). The second says that first, if the current agent has not adopted p then C can

make sure that doesn't change (by doing nothing), and, second, that if C can make sure that the current agent does not adopt p then they can do it by doing nothing (in which case p must be false initially). While the conjunctive closure axiom (often called C in the literature) does not hold in general, it does hold for atomic behaviours as the third validity (direction towards the left) shows. This is because p and q are independent: C can combine the choices for p and q to a choice for $p \wedge q$. Neither of these hold if we replace the atomic behaviours p and q with general formulae: the first is obvious, for the second consider, e.g., $\neg p$ instead of p , for the third consider the case that $q = \neg p$ and the direction towards the left and a situation where the current agent adopts p if and only if C adopts p .

We now move on to properties of $\Box$ and $\Box^*$ and their interaction with the $\langle C \rangle$ modalities. We have the following.

Theorem 3 *The following are valid ($p \in \Pi$).*

- 1) $\Box p \rightarrow p$ (*upper threshold*)
- 2) $\neg \varphi \rightarrow \Box \neg \Box \varphi$ (*symmetry*)
- 3) $\langle C \rangle \Diamond \varphi \leftrightarrow \Diamond \langle C \rangle \varphi$
- 4) $\langle C \rangle \Box \varphi \rightarrow \Box \langle C \rangle \varphi$
- 5) $\langle C \rangle \Box^* \varphi \rightarrow \Box^* \langle C \rangle \varphi$

The first of these properties says that the upper limit of the threshold (fraction) is 1; in a stable model, if all your neighbours have adopted then so have you. The second is the standard axiom for symmetry of the social relation. For the third property, the left hand side holds iff the right hand side holds – for the same friend of the current agent and the same choice for C . Also for the fourth property, the same choice for C that ensures that each of the current agent's friends has property φ can be used to show that for each of her friends, C can ensure property φ . This, however does not hold in the other direction – because $\Box \langle C \rangle \varphi$ might hold when a *different* choice is necessary for two different friends. To see this, consider the model M' which is like M in figure 2 (left) except that $\theta(b, p) = \frac{1}{2}$, and let $\varphi = p \wedge \Diamond \neg p$, and $C = \{c, d\}$. We have that $M, i \models \Box \langle C \rangle \varphi$: for a, c but not d can choose to adopt p , in which case a (but not b) will adopt p and φ will hold in a , and for b, d but not c can choose to adopt p , in which case b (but not a) will adopt p and φ will hold in b . But $M, i \not\models \neg \langle C \rangle \Box \varphi$: none of the four choices available to $\{c, d\}$ will ensure that φ holds in *both* a and b . This distinction is similar to the de re/de dicto distinction that comes up in reasoning about knowledge and coalitional ability [2]. The case of the fifth validity is similar.

The reader might have noticed that several of the properties above have the form of *reduction axioms*, known from, e.g., dynamic epistemic logic [18]. In particular, we see that the $\langle C \rangle$ modalities have several *distributive* properties. For example, they distribute (in both directions) over disjunction in general, and over conjunction of atomic behaviours, and there is a reduction axiom for $\langle C \rangle$ and $\Diamond$. However, as we have seen, $\langle C \rangle$ does not distribute in both directions over conjunction in general, and there is no reduction axiom for negation (except in the special case of atomic behaviours) – and no reduction axiom for $\langle C \rangle$ and $\Box$.

VI. NO MATTER WHAT OTHERS DO: A COALITION LOGIC

$\langle C \rangle \varphi$ means that C can choose to initially adopt certain combinations of behaviours in such a way that φ is guaranteed to be true. It is implicitly assumed that agents outside of C are not initial adopters. This makes sense (1) in situations where we can control the initial adopters and the only way for others to adopt is through influence, or (2) as a “worst-case” guarantee that the goal will hold even if no other agents are initial adopters. In this section we look at a derived operator $\llbracket C \rrbracket \varphi$, with the intuitive meaning that C can choose to initially adopt behaviours in such a way that *no matter whether or not agents outside of C initially adopts*, φ will be true after the potentially resulting cascade. This is of interest in situations where (1) and/or (2) do not hold.

Define:

$$\llbracket C \rrbracket \varphi \equiv \langle C \rangle [\bar{C}] \varphi$$

where $\bar{C} = N \setminus C$. This “no matter what the other agents do” notion of ability is well known from game theory, and is the defining notion of coalitional ability in Coalition Logic [15]. As mentioned, it also plays a natural role when reasoning about the ability of agents in a social network to trigger cascades in settings where initial adoption can occur autonomously and/or randomly.

With the above definition, the semantics of the $\llbracket C \rrbracket$ operator becomes:

$$M, a \models \llbracket C \rrbracket \varphi \text{ iff } \exists V' \sim_C V, \forall V'' \sim_{\bar{C}} V', (R, \theta, V'')^\dagger, a \models \varphi$$

This is in fact equivalent to

$$M, a \models \llbracket C \rrbracket \varphi \text{ iff } \exists V' \forall V'' \sim_{\bar{C}} V', (R, \theta, V'')^\dagger, a \models \varphi$$

– also showing that the meaning of $\llbracket C \rrbracket \varphi$ is in fact independent from the current valuation.

There are now even more possible cascades resulting from a choice of initial adoption of behaviours by C , depending on what the other agents do, and $\llbracket C \rrbracket \varphi$ means that φ is guaranteed to be true after *any* of them, for some choice by C . If φ is an atom p , then the “no matter what the other agents do” part does not matter – then adoption through cascades is *monotonic*, the other agents can never do anything to spoil it.

Lemma 2 For any M, a, C , and p ,

$$M, a \models \llbracket C \rrbracket p \text{ iff } M^{C \hookrightarrow p}, a \models p.$$

In other words, $\llbracket C \rrbracket p$ means that if everyone in C adopts p then the current agent will adopt p . Thus, for single behaviours $\varphi = p$, the exists-forall semantics of the $\llbracket C \rrbracket \varphi$ collapses – not to just a single quantifier, but to no quantifiers at all. This is not the case, however, for more complex formulas φ , as we will see. In particular, agents outside of C can spoil the achievement of *lack* of adoption, $\varphi = \neg p$. Furthermore, “no matters what the other agents do” generally can lead to several possible outcomes, as in Figure 2 (right) and discussed below.

Consider the model M based on the network in Figure 2 (left), where no one has adopted any behaviour. Let $\varphi = \Diamond p \wedge \Diamond \neg p$. Observe that $M, i \models \llbracket \{i, b, c\} \rrbracket \varphi$: the only strategy that works here is for (only) c to adopt p . It will not work for

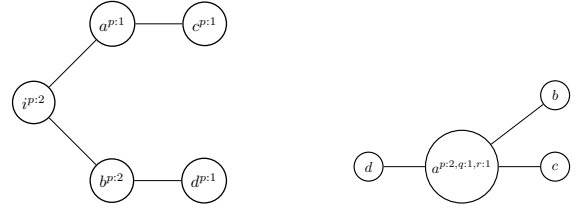


Figure 2. Left: M with $\theta(i, p) = \theta(b, p) = \theta(c, p) = \theta(d, p) = 1$, $\theta(a, p) = \frac{1}{2}$. Right: M with $\theta(a, p) = \frac{2}{3}$, $\theta(a, q) = \theta(a, r) = \frac{1}{3}$ (thresholds for the other agents do not matter). The nodes in the figure are marked with the number (rather than the fraction) of friends needed.

i to adopt p – even though we have both $M, i \models \langle c \rangle \varphi$ and $M, i \models \langle i \rangle \varphi$. Note also that $M, i \not\models \llbracket c \rrbracket \varphi$, illustrating the difference between the two types of ability.

Moving on to properties of the $\llbracket C \rrbracket$ modalities, we first observe that they satisfy all axioms/rules of Coalition Logic [15].

Theorem 4 The following are valid/validity preserving.

- 1) $\llbracket C \rrbracket \top$ (Liveness)
- 2) $\neg \llbracket C \rrbracket \perp$ (Safety)
- 3) $\neg \llbracket \emptyset \rrbracket \neg \varphi \rightarrow \llbracket N \rrbracket \varphi$ (N -maximality)
- 4) $\llbracket C \rrbracket (\varphi \wedge \psi) \rightarrow \llbracket C \rrbracket \varphi$ (Outcome monotonicity)
- 5) $(\llbracket C_1 \rrbracket \varphi_1 \wedge \llbracket C_2 \rrbracket \varphi_2) \rightarrow \llbracket C_1 \cup C_2 \rrbracket (\varphi_1 \wedge \varphi_2)$ when $C_1 \cap C_2 = \emptyset$ (Superadditivity)
- 6) If $\varphi \leftrightarrow \psi$ is valid, then $\llbracket C \rrbracket \varphi \leftrightarrow \llbracket C \rrbracket \psi$ is valid (RG)

The $\llbracket C \rrbracket$ modalities have additional properties that the standard coalition modalities do not have.

Theorem 5 The following are valid (for any $p, q \in \Pi$).

- 1) $\llbracket N \rrbracket p$
- 2) $\neg \llbracket \emptyset \rrbracket p$
- 3) $\llbracket C \rrbracket \neg p \leftrightarrow \neg \llbracket \bar{C} \rrbracket p$
- 4) $\llbracket C \rrbracket (p \wedge q) \leftrightarrow (\llbracket C \rrbracket p \wedge \llbracket C \rrbracket q)$
- 5) $\llbracket C \rrbracket (p \vee q) \leftrightarrow (\llbracket C \rrbracket p \vee \llbracket C \rrbracket q)$
- 6) $\llbracket C \rrbracket \llbracket D \rrbracket \varphi \leftrightarrow \llbracket D \rrbracket \varphi$

Note that 1-5 do not hold if the atomic behaviours p and q are replaced by general formulae. In particular, while $\llbracket C \rrbracket$ distributes over disjunction in both directions for atomic behaviours, it does not distribute over disjunction in general, in contrast with $\langle C \rangle$. To see this, let M be as in Figure 2 (right). $M, a \models \llbracket \{b\} \rrbracket ((p \wedge q) \vee (\neg p \wedge r))$: b can adopt both q and r , which will make a adopt both q and r as well. a might adopt p as well, depending on what $\{c, d\}$ do (and on whether a chooses to be an initial adopter herself), but in either case $p \wedge q$ or $\neg p \wedge r$ will be true for a . However, neither $\llbracket \{b\} \rrbracket (p \wedge q)$ nor $\llbracket \{b\} \rrbracket (\neg p \wedge r)$ holds in a : b cannot control whether or not a adopts p . Thus, $\llbracket C \rrbracket$ does not distribute over disjunction.

Consider now the interaction between $\llbracket C \rrbracket$ and $\Diamond/\Box$.

Theorem 6 The following are valid.

- 1) $\llbracket C \rrbracket \Box \varphi \rightarrow \Box \llbracket C \rrbracket \varphi$
- 2) $\llbracket C \rrbracket \Box^* \varphi \rightarrow \Box^* \llbracket C \rrbracket \varphi$
- 3) $\Diamond \llbracket C \rrbracket \varphi \rightarrow \llbracket C \rrbracket \Diamond \varphi$

The relationship to Coalition Logic also sheds light on the computational complexity of the satisfiability problem for our logic. NP-hardness follows trivially from the fact that it contains propositional logic, but from Theorem 4 (and the fact that there is a constant time encoding of $\langle C \rangle \varphi \equiv \langle C \rangle [\bar{C}] \varphi$) we immediately get the following.

Lemma 3 *The satisfiability problem for the logic is PSPACE-hard.*

VII. DISCUSSION

In [1] the power of individuals and coalitions in terms of being sufficient and/or necessary for triggering cascades, and the computational complexity of related problems, is studied in detail – without any formal logic for reasoning about coalitional ability. When it comes to related work on logics for social networks, this paper relates most closely to [6] on the one hand and to [8], [5] on the other hand. The former, [6], offers the first logical model of threshold diffusion in social networks. It differs from our approach by the fact that it uses atomic propositions to describe who is a neighbour of whom, and uses a unique dynamic modality $[adopt]\varphi$ to model explicitly the *step-by-step* diffusion process. In contrast, our language can only talk explicitly about the *fixed point* of such diffusion processes. This is also the case of [8] and [5], both of which model the ability of a set of agents, in case they all adopt a new behaviour, to eventually influence all agents in another set into adopting it too. There is a trade-off between expressivity and simplicity of the corresponding logics: the succinct languages of “ability systems”, such as ours, [8], and [5], cannot capture what happens “after two time-steps” for instance, contrary to the more fine-grained model of [6]. However, since the ability systems abstract further away from the details of thresholds, they avoid having a different axiom for each assignment of thresholds, which is needed in [6].

Note that, contrary to the systems discussed so far, ours also includes the social network modality quantifying over network-neighbours, to be able to talk about (the behaviour of) one’s neighbour’s neighbours, for instance, as first introduced by [17] and further used in, among others, [7], [13], [10].

Most importantly, to the best of our knowledge, this is the first use of coalitional ability logic for social networks. Given both the commonplace use of coalition logic in the context of games, and the typical game-theoretic interpretation of threshold models as best response dynamics of iterated coordination games on networks, it is natural to apply such operators to threshold models. The resulting logic is simple, yet expressive enough to describe, e.g., complete cascades.

By studying valid properties we have shed some light on the principles of reasoning about coalitional ability in threshold models. The most obvious open problems for future work are completeness results. It would be natural to start with a fragment of the language. The fragment without $\square$ or $\square^*$ is of particular interest: while it cannot express ability to trigger a complete cascade, it can still express interesting properties. In a sense, a complete axiomatisation of the validities of this

language would give us the most basic principles of reasoning about ability in networks. The language with $\langle C \rangle$ as a primary operator, and without $\langle C \rangle$, is also of interest – the resulting logic properly containing Coalition Logic.

Going in the other direction, increasing the expressive power, is also of interest. One idea is to generalise $\langle C \rangle$ to $\langle C : \alpha \rangle$, where α is a propositional formula over Π . For example, $\langle C : p \vee q \rangle \varphi$ is true if φ is guaranteed to be true if each agent in C adopts either p or q . This could be useful for reasoning about situations where, e.g., I know that I can make everyone in a certain group adopt either product p or product q (but not necessarily who would adopt which), and want to reason about the possible cascades that this might trigger.

Another open problem is to obtain a complete characterisation of the computational complexity of the satisfiability problem. We conjecture that (1) the language without the $\diamond/\square$ modalities is PSPACE-complete, and (2) the full language is EXPTIME-complete. EXPTIME hardness is likely since several modal logics with a transitive closure operator on non-transitive frames are EXPTIME-hard, although there does not seem to be a result for symmetric and irreflexive frames reported in the literature.

REFERENCES

- [1] Thomas Ágotnes and Zoé Christoff. Cascading power. In Dong Hao, Bin Li, Swaprava Nath, Taiki Todo, and Dengji Zhao, editors, *Mechanism Design in Social Networks*, pages 36–51. Springer Nature, 2025.
- [2] Thomas Ágotnes, Valentin Goranko, Wojciech Jamroga, and Michael Wooldridge. Knowledge and ability. In *Handbook of Epistemic Logic*, pages 543–589. College Publications, 2015.
- [3] Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49(5):672–713, 2002.
- [4] K. R. Apt and E. Markakis. Diffusion in Social Networks with Competing Products. In G. Persiano, editor, *SAGT 2011*, LNCS 6982, pages 212–223. Springer, 2011.
- [5] S. Azimipour and P. Naumov. Lighthouse principle for diffusion in social networks. *FLAP*, 5(1):97–120, 2018.
- [6] A. Baltag, Z. Christoff, R. K. Rendsvig, and S. Smets. Dynamic epistemic logics of diffusion and prediction in social networks. *Studia Logica*, 107(3):489–531, 2019.
- [7] Z. Christoff and J. U. Hansen. A logic for diffusion in social networks. *Journal of Applied Logic*, 13(1):48–77, 2015.
- [8] Z. Christoff and P. Naumov. Diffusion in social networks with recalcitrant agents. *Journal of Logic and Computation*, 29(1):53–70, 12 2018.
- [9] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
- [10] P. Girard, J. Seligman, and F. Liu. General dynamic logic. In *Advances in Modal Logic*, Vol. 9, pages 239–260. College Publ., 2012.
- [11] M. Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.
- [12] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD 03: Proceedings of the 9th ACM SIGKDD int. conference*, pages 137–146. ACM, 2003.
- [13] F. Liu, J. Seligman, and P. Girard. Logical dynamics of belief change in the community. *Synthese*, 191(11):2403–2431, 2014.
- [14] S. Morris. Contagion. *Review of Economic Studies*, 67:57–78, 2000.
- [15] M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.
- [16] T. C. Schelling. *Micromotives and Macrobehavior*. Norton, N.Y., 1978.
- [17] J. Seligman, F. Liu, and P. Girard. Logic in the community. In M. Banerjee and A. Seth, editors, *Logic and Its Applications*, volume 6521 of *Lecture Notes in Computer Sci.*, pages 178–188. Springer, 2011.
- [18] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic epistemic logic*, volume 337 of *Synthese Library*. Springer, 2008.

Manifold-Preserving Social Recommendation in Hyperbolic Space

Tra My Nguyen Huynh[†]

*Department of Computational Science and Engineering
VNU University of Engineering and Technology
Hanoi, Vietnam
mynht@vnu.edu.vn*

Viet Quan Le[†]

*Department of Computer Science
VNU University of Engineering and Technology
Hanoi, Vietnam
lvquan@vnu.edu.vn*

Abstract—Hyperbolic space has demonstrated strong potential in representing hierarchical networks for social recommendation. However, many state-of-the-art hyperbolic recommenders depend on the tangent space for graph learning, which distorts the embedding quality, resulting in poor performance. To overcome the weakness of tangent space, we propose a novel recommendation framework called Hyperbolic Manifold-Preserving Graph-based Social Recommendation (HMPGSR). Specifically, we combine three different geometric models to ensure that the framework operates directly in hyperbolic space. HMPGSR contains two main stages: social pretraining and preference learning. The social pretraining stage extracts users’ social features, which serve as informative inputs for the subsequent phase. The preference learning stage integrates both user-item interactions and users’ social networks to create a heterogeneous graph. By combining heterogeneous graph structure with pre-trained social embeddings, HMPGSR extracts more complex relationships between users and items, thereby improving social recommendation quality. The experimental results on real-world datasets demonstrate that HMPGSR consistently outperforms existing hyperbolic recommendation methods.

Index Terms—Social recommendation, hyperbolic geometry, graph neural networks

I. INTRODUCTION

Recommender systems aim to deliver personalized item suggestions by analyzing user preferences [1]. As a foundational approach, Collaborative Filtering (CF) has been extensively applied in recommender systems to predict users’ preferences based on user-item historical interaction [2], [3]. However, the effectiveness of CF remains limited due to the insufficient user interaction data. To address this limitation, social networks have been incorporated into recommendation frameworks to enhance personalization [4], [5]. Social recommendation leverages user-user relationships from social networks to alleviate data sparsity and improve recommendation accuracy.

Researchers have utilized the expressiveness of graph convolutional networks (GCNs) [6] for social recommendation and achieved robust performances. Graph-based recommendation methods propagate user and item representations with graph convolutions, effectively capturing high-order dependencies between users and items [7], [8]. Most of the graph-based approaches are formulated in Euclidean space, which supports

the definitions of linear algebra operations. Alternatively, social graphs and user-item graphs have been shown to follow a power-law distribution and exhibit hierarchical topologies, which are not well-represented in Euclidean space [5], [9]. In contrast to Euclidean geometry, hyperbolic geometries such as the Poincaré ball and the Lorentz model are more suitable for embedding hierarchical graphs because of their exponential expansion property [10].

In recent years, there has been an increasing interest in graph-based recommendation with hyperbolic geometry [5], [11]. Despite achieving state-of-the-art performance, hyperbolic methods are often inefficient because of their dependence on tangent space. The induced distance among two neighbor nodes’ embedding on the hyperbolic manifold, which reflects the graph’s hierarchical structure, is not preserved after the mapping to the tangent space [12]. The distortion errors caused by the tangent mappings tend to increase with graph size, which reduces the learning capability of hyperbolic models in general. To avoid distortion issue, the authors in [12] propose the manifold-preserving graph architecture which works directly on hyperbolic manifold by using the Poincaré ball model. This approach demonstrates significant improvements in dynamic graph learning tasks.

Motivated by the advantage of manifold preserving [12], we introduce a novel model called Hyperbolic Manifold-Preserving Graph-based Social Recommendation (HMPGSR) to address distortion issues arising from tangent space. HMPGSR combines three geometric models: the Poincaré ball, the Lorentz model, and the Klein model to operate fully within hyperbolic space. There are two main stages in HMPGSR: social pretraining and preference learning. In the social pretraining stage, we capture users’ social features by optimizing a social network reconstruction task, producing embeddings that serve as informative inputs for the subsequent phase. During preference learning, we construct a heterogeneous graph that integrates both user-item interactions and users’ social networks. By combining heterogeneous graph structure with pre-trained social embeddings, HMPGSR extracts more complex relationships between users and items, thereby improving social recommendation quality. Experimental results on two public benchmarks demonstrate that HMPGSR consistently outperforms existing hyperbolic recommendation methods.

[†]Equal Contribution

II. RELATED WORKS

Recently, researchers have embraced the fundamental concept of graph convolution networks (GCNs) [6] in graph-based recommendations, achieving state-of-the-art performance. For instance, LightGCN [7] employs solely neighborhood message-passing in graph networks for embedding learning, ensuring the effectiveness and efficiency. Furthermore, to learn about user representation, DiffNet++ [8] simulates the diffusion of connections between users while taking into account the diffusion of high-order social influence. Although these graph-based social recommendation techniques are able to learn user and object embeddings in Euclidean space, they encounter representation distortion in tree-like and scale-free graphs.

Hyperbolic geometries have recently been used for recommendation tasks since recommendation graphs inherit hierarchical topologies. HGCF [11] suggests using hyperbolic skip-connections in graph convolution architectures for collaborative filtering. HGSR [5] proposes to learn the heterogeneous relationships between users and items in hyperbolic space, which enhances recommendation performance. However, HGSR still relies on tangent space for graph operations, which negatively affects the performance of the hyperbolic model.

III. PRELIMINARIES

A. Social Recommendation

A social recommendation system consists of a list U of M users and a collection V of N items, where connections between them are classified as: *user-item interactions* and *user-user social connections*. Following the work [5], we represent user-item interactions with a binary matrix $\mathbf{R} \in \mathbb{R}^{M \times N}$ where $r_{ai} = 1$ if user a has interacted with item i , and $r_{ai} = 0$ otherwise. In addition, user-user social connections are modeled by a social adjacency matrix $\mathbf{S} \in \mathbb{R}^{M \times M}$, where $s_{ba} = 1$ indicates that user a follows user b , and $s_{ba} = 0$ otherwise. Given both the interaction matrix $\mathbf{R}$ and the social network matrix $\mathbf{S}$, the objective of graph-based social recommendation is to learn a predictive function f that estimates unknown user preferences, denoted as $\hat{\mathbf{R}} = f(\mathbf{R}, \mathbf{S})$.

B. Hyperbolic Geometry in Social Recommendation

A Riemannian manifold $(\mathcal{M}, g)$ is a differentiable manifold $\mathcal{M}$ equipped with a Riemannian metric tensor g . A hyperbolic space is a specific type of Riemannian manifold characterized by a constant negative curvature. The tangent space $\mathcal{T}_x \mathcal{M}$ at a point $x \in \mathcal{M}$ serves as a first-order Euclidean approximation of the manifold around that point. Due to the compatibility with linear algebra operations, most hyperbolic methods perform graph computations in the tangent space rather than directly on the curved manifold. As mentioned in [12], the reliance on tangent space for graph operations is considered suboptimal. Hence, to build our manifold-preserving recommendation method in hyperbolic space, we integrate three complementary geometric models: the Poincaré ball, the Lorentz model, and the Klein model [13].

Let $\|\cdot\|$ be the L2-norm. The Poincaré ball of an d -dimensional hyperbolic space with constant negative curvature $-c$ ($c > 0$) is defined by the Riemannian manifold $\mathcal{B}_c^d = (\mathcal{H}^{d,c}, g^{\mathcal{H}^{d,c}})$, where $\mathcal{H}^{d,c} = \{x \in \mathbb{R}^d : c\|x\|^2 < 1\}$ is an open d -dimensional ball and $g^{\mathcal{H}^{d,c}}$ is the hyperbolic metric tensor which is conformal to the Euclidean metric tensor g^E :

$$g_x^{\mathcal{H}^{d,c}} = (\lambda_x^c)^2 g^E, \text{ where } \lambda_x^c = \frac{2}{1 - c\|x\|^2} \text{ and } g^E = \mathbf{I}_d. \quad (1)$$

In our proposed framework, the *Möbius addition* in the Poincaré ball [14] is used to aggregate hidden representations across layers, which is given by:

$$x \oplus_c y = \frac{(1 + 2c\langle x, y \rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y \rangle + c^2\|x\|^2\|y\|^2}, \quad (2)$$

where $x, y \in \mathcal{B}_c^d$. When $c = 0$, one becomes the Euclidean addition of two vectors in $\mathbb{R}^d$.

The Lorentz model $\mathcal{L}_c^d$ and the Klein model $\mathcal{K}_c^d$ are two other models of the hyperbolic space that are isometrically equivalent to the Poincaré ball $\mathcal{B}_c^d$. The isometric bijections between a point $x \in \mathcal{B}_c^d$ and its corresponding point $k \in \mathcal{K}_c^d$ are defined as:

$$p_{\mathcal{B} \rightarrow \mathcal{K}}(x) = \frac{2x}{1 + c\|x\|^2}, p_{\mathcal{K} \rightarrow \mathcal{B}}(k) = \frac{k}{1 + \sqrt{1 - c\|k\|^2}}. \quad (3)$$

And the isometric bijections between the Poincaré ball $\mathcal{B}_c^d$ and the Lorentz model $\mathcal{L}_c^d$ are formulated as:

$$p_{\mathcal{B} \rightarrow \mathcal{L}}(x) = \frac{[c^{-1} + \|x\|^2, 2c^{\frac{-1}{2}}x]}{c^{-1} - \|x\|^2}, p_{\mathcal{L} \rightarrow \mathcal{B}}(l) = \frac{c^{\frac{-1}{2}}[l_1, \dots, l_d]}{l_0 - c^{-1}}, \quad (4)$$

where $l = [l_0, l_1, \dots, l_d] \in \mathcal{L}_c^d$ is the corresponding point of x .

The Klein model facilitates signal propagation between users and items within hyperbolic manifolds by leveraging the Einstein midpoint [12]. Given a set of points in Poincaré ball $X^{\mathcal{B}_c^d} = \{x_1^{\mathcal{B}_c^d}, \dots, x_M^{\mathcal{B}_c^d}\}$ and a positive weight set $\mathcal{W} = \{w_1, \dots, w_M\}$, the message passing function built on Einstein midpoint is expressed as:

$$\Phi(X^{\mathcal{B}_c^d}, \mathcal{W}) = \begin{cases} x_i^{\mathcal{K}_c^d} = p_{\mathcal{B} \rightarrow \mathcal{K}}(x_i^{\mathcal{B}_c^d}), \\ \tilde{x} = (\sum_{i=1}^M w_i \gamma_i x_i^{\mathcal{K}_c^d}) / (\sum_{i=1}^M w_i \gamma_i), \\ \tilde{x}^{\mathcal{B}_c^d} = p_{\mathcal{K} \rightarrow \mathcal{B}}(\tilde{x}^{\mathcal{K}_c^d}), \end{cases} \quad (5)$$

where $\gamma_i = (1 - c\|x_i^{\mathcal{K}_c^d}\|^2)^{-1/2}$ is the Lorentz factor.

The Lorentz model produces distance-based predictions for efficient recommendation and robust optimization by taking advantage of its numerical stability and advanced social regularization [5], [11]. Given a pair of points $x, y \in \mathcal{L}_c^d$, the induced distance between them is computed as follows:

$$d_{\mathcal{L}}(x, y) = \frac{1}{\sqrt{c}} \operatorname{arcosh}(-c\langle x, y \rangle_{\mathcal{L}}), \quad (6)$$

where $\langle x, y \rangle_{\mathcal{L}} = -x_0 y_0 + x_1 y_1 + \dots + x_d y_d$ is the Lorentz inner product.

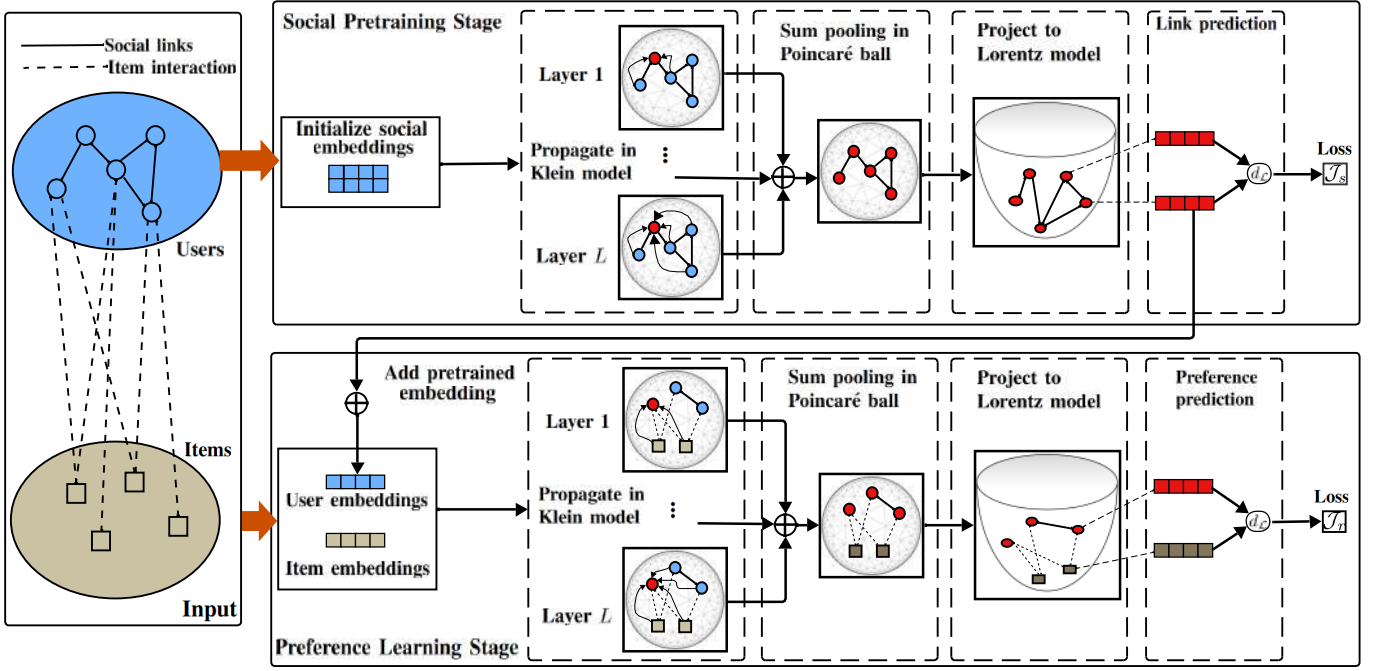


Fig. 1. Illustration of HMPGSR architecture which consists of two stages: **social pretraining** (upper part) and **preference learning** (bottom part)

IV. METHODOLOGY

To overcome the tangent space issues, we propose the Hyperbolic Manifold-Preserving Graph-based Social Recommendation (HMPGSR) which is a recommender system operating fully in hyperbolic space. Figure 1 illustrates the architecture of HMPGSR, consisting of two phases: social pretraining and preference learning. The social pretraining phase is designed to capture general social features while preserving the underlying hierarchical structure. Subsequently, the preference learning stage integrates the pretrained social embeddings with the heterogeneous user-item networks to extract high-quality representations in the recommender system.

A. Social Pretraining Stage

As shown in Figure 1, the social pretraining phase contains four key components: embedding initialization, social propagation, sum pooling, and social link optimization. User hyperbolic embeddings are initialized from a hyperbolic Gaussian distribution [11], resulting in vectors in the Lorentz model $\mathcal{L}_c^d$. Let $\mathbf{Z}^{\mathcal{L}_c^d} = \{z_i^{\mathcal{L}_c^d}\}_{i=1,\dots,M}$ denote the user initial features in the Lorentz model, we project the features to the Poincaré ball with the isometric relations in (4):

$$\mathbf{Z}^{0,\mathcal{B}_c^d} = p_{\mathcal{L} \rightarrow \mathcal{B}}(\mathbf{Z}^{\mathcal{L}_c^d}). \quad (7)$$

The user embeddings $\mathbf{Z}^{0,\mathcal{B}_c^d}$ are then feed into a diffusion process to capture the connections between users in higher orders. This approach is well-known for social recommendation such as HGSR [5]. However, our proposed diffusion is designed to work directly on the hyperbolic manifold instead of the tangent space, by using the Einstein midpoint. Specifically, given the

user a embedding $z_a^{l,\mathcal{B}_c^d}$ in l^{th} diffusion layer in Poincaré ball, the propagated features $z_a^{l+1,\mathcal{B}_c^d}$ is computed as:

$$z_a^{l+1,\mathcal{B}_c^d} = z_a^{l,\mathcal{B}_c^d} \oplus_c \Phi(\{z_b^{l,\mathcal{B}_c^d}\}_{b \in S_a}, \{w_b\}_{b \in S_a}), \quad (8)$$

where S_a indicates a set of users that social connected to the user a , and $\{w_b\}_{b \in S_a}$ are the diffusion weights. In general, $\{w_b\}_{b \in S_a}$ can be treated as trainable parameters; however, for computational efficiency, we simplify by setting $w_b = 1$ for all $b \in S_a$. The Einstein midpoint Φ in (5) is employed to cumulate the neighbor representations $\{z_b^{l,\mathcal{B}_c^d}\}_{b \in S_a}$. The neighbor feature is then combined with user a feature using Möbius addition $\oplus_c$ in (2). User embedding $\mathbf{Z}^{l,\mathcal{B}_c^d}$ in each l^{th} layer is saved, and the Möbius addition sum-pooling is applied to a sequence of matrices $[\mathbf{Z}^{1,\mathcal{B}_c^d}, \dots, \mathbf{Z}^{L,\mathcal{B}_c^d}]$, to achieve multi-hop information for social recommendation:

$$\bar{\mathbf{Z}}^{\mathcal{B}_c^d} = \sum_{l=1}^L \oplus_c \mathbf{Z}^{l,\mathcal{B}_c^d}. \quad (9)$$

Finally, the output $\bar{\mathbf{Z}}^{\mathcal{B}_c^d}$ is transformed to the Lorentz model by isometric mappings in (4) for social prediction and optimization

$$\bar{\mathbf{Z}}^{\mathcal{L}_c^d} = p_{\mathcal{B} \rightarrow \mathcal{L}}(\bar{\mathbf{Z}}^{\mathcal{B}_c^d}), \quad (10)$$

$$\hat{s}_{ab} = (d_{\mathcal{L}}^2(\bar{z}_a^{\mathcal{L}_c^d}, \bar{z}_b^{\mathcal{L}_c^d}))^{-1}. \quad (11)$$

The propensity score $\hat{s}_{ab}$, representing the likelihood of a link between user a and user b , is computed as the inverse of the squared Lorentz distance between their corresponding

nodes. We leverage the social objective function in [5] in social pretraining stage, which is described as:

$$m_{ua} = \delta \left(d_{\mathcal{L}}^2(\bar{z}_u^{\mathcal{L}^d}, \mathbf{o}) + d_{\mathcal{L}}^2(\bar{z}_a^{\mathcal{L}^d}, \mathbf{o}) - d_{\mathcal{L}}^2(\bar{z}_u^{\mathcal{L}^d}, \bar{z}_a^{\mathcal{L}^d}) \right), \quad (12)$$

$$\mathcal{J}_s = \sum_{u=0}^{M-1} \sum_{D_u^S} \Phi \left(d_{\mathcal{L}}^2(\bar{z}_u^{\mathcal{L}^d}, \bar{z}_a^{\mathcal{L}^d}) - d_{\mathcal{L}}^2(\bar{z}_u^{\mathcal{L}^d}, \bar{z}_b^{\mathcal{L}^d}) + m_{ua} \right), \quad (13)$$

where $D_u^S = \{(a, b) \mid a \in \mathbf{S}_u \wedge b \notin \mathbf{S}_u\}$ indicates the collection of positive sample (u, a) and negative data (u, b) , $\delta(\cdot)$ denotes the sigmoid function, and $\Phi(x) = \max(x, 0)$. The adaptive margin m_{ua} is defined by the distance of positive pair (u, a) to the hyperbolic origin $\mathbf{o}$, encouraging recommender systems to focus on less-attended items.

B. Preference Learning Stage

Inspired by the approach in [5], we construct a heterogeneous network consisting of user-user social connections and user-item preferences, as illustrated in the lower part of Figure 1. The structure of the preference learning phase includes four components, analogous to those in the preceding stage. We first initialize user and items embeddings $\mathbf{U}^{\mathcal{L}^d} = \{u_i^{\mathcal{L}^d}\}_{i=1, \dots, M}$, $\mathbf{V}^{\mathcal{L}^d} = \{v_j^{\mathcal{L}^d}\}_{j=1, \dots, N}$ with a hyperbolic Gaussian sampling, then project them to Poincaré ball:

$$\mathbf{U}^{-1, \mathcal{B}^d} = p_{\mathcal{L} \rightarrow \mathcal{B}}(\mathbf{U}^{\mathcal{L}^d}), \quad \mathbf{V}^{0, \mathcal{B}^d} = p_{\mathcal{L} \rightarrow \mathcal{B}}(\mathbf{V}^{\mathcal{L}^d}). \quad (14)$$

The pretrained social embeddings in the first stage $\bar{\mathbf{Z}}^{\mathcal{L}^d}$ are incorporated as an auxiliary input to enhance recommendation performance:

$$\mathbf{U}^{0, \mathcal{B}^d} = \mathbf{U}^{-1, \mathcal{B}^d} \oplus_c p_{\mathcal{L} \rightarrow \mathcal{B}}(\bar{\mathbf{Z}}^{\mathcal{L}^d}). \quad (15)$$

In social recommendation, it is generally assumed that a user's preferences are influenced by socially connected friends and past interactions with items. Therefore, a diffusion process is employed on our heterogeneous network to extract user multi-hop embeddings. Given the user embedding a and the item i embedding $u_a^{l, \mathcal{B}^d}, v_i^{l, \mathcal{B}^d}$ in l^{th} layer in the Poincaré ball, we update their embeddings as follows:

$$v_i^{l+1, \mathcal{B}^d} = v_i^{l, \mathcal{B}^d} \oplus_c \Phi(\{u_a^{l, \mathcal{B}^d}\}_{a \in \mathbf{R}_i^T}, \{1\}), \quad (16)$$

$$u_a^{l+1, \mathcal{B}^d} = u_a^{l, \mathcal{B}^d} \oplus_c \alpha \Phi(\{v_b^{l, \mathcal{B}^d}\}_{b \in \mathbf{S}_a}, \{1\}) \oplus_c (1 - \alpha) \Phi(\{v_j^{l, \mathcal{B}^d}\}_{j \in \mathbf{R}_a}, \{1\}), \quad (17)$$

where $\mathbf{S}_a$ is a social neighborhood of user a , $\mathbf{R}_a$ is a list of interacted items, and $\Phi(\cdot, \{1\})$ is the Einstein midpoint in (5) with weight parameters equal to 1. In (17), we integrate a hyper-parameter α to prevent bias toward either social information or item preferences. After L diffusion layers, we receive a set of user embeddings $[\mathbf{U}^{0, \mathcal{B}^d}, \dots, \mathbf{U}^{L, \mathcal{B}^d}]$ and a set of item embeddings $[\mathbf{V}^{0, \mathcal{B}^d}, \dots, \mathbf{V}^{L, \mathcal{B}^d}]$, which are combined using a sum pooling in the social pretraining stage:

$$\bar{\mathbf{U}}^{\mathcal{B}^d} = \sum_{l=1}^L \oplus_c \mathbf{U}^{l, \mathcal{B}^d}, \quad \bar{\mathbf{V}}^{\mathcal{B}^d} = \sum_{l=1}^L \oplus_c \mathbf{V}^{l, \mathcal{B}^d}. \quad (18)$$

Finally, the learned embeddings are projected to the Lorentz model via the isometric bijection in (4), and the Lorentz distance $d_{\mathcal{L}}(u_a, v_i)$ from an user a to an item i is computed for the preference prediction:

$$\bar{\mathbf{U}}^{\mathcal{L}^d} = p_{\mathcal{B} \rightarrow \mathcal{L}}(\bar{\mathbf{U}}^{\mathcal{B}^d}), \quad \bar{\mathbf{V}}^{\mathcal{L}^d} = p_{\mathcal{B} \rightarrow \mathcal{L}}(\bar{\mathbf{V}}^{\mathcal{B}^d}), \quad (19)$$

$$\hat{r}_{ai} = (d_{\mathcal{L}}^2(\bar{u}_a^{\mathcal{L}^d}, \bar{v}_i^{\mathcal{L}^d}))^{-1}. \quad (20)$$

Similar to social pretraining stage, we employ the objective function in [15] with the adaptive margins in this stage, which is computed as:

$$m_{ai} = \delta \left(d_{\mathcal{L}}^2(\bar{u}_a^{\mathcal{L}^d}, \mathbf{o}) + d_{\mathcal{L}}^2(\bar{v}_i^{\mathcal{L}^d}, \mathbf{o}) - d_{\mathcal{L}}^2(\bar{u}_a^{\mathcal{L}^d}, \bar{v}_i^{\mathcal{L}^d}) \right), \quad (21)$$

$$\mathcal{J}_r = \sum_{a=0}^{M-1} \sum_{D_a^R} \Phi \left(d_{\mathcal{L}}^2(\bar{u}_a^{\mathcal{L}^d}, \bar{v}_i^{\mathcal{L}^d}) - d_{\mathcal{L}}^2(\bar{u}_a^{\mathcal{L}^d}, \bar{v}_j^{\mathcal{L}^d}) + m_{ai} \right), \quad (22)$$

where $D_a^R = \{(i, j) \mid i \in \mathbf{R}_a \wedge j \notin \mathbf{R}_a\}$ includes positive and negative item pairs (i, j) associated with user a , where i denotes an interacted item and j a non-interacted one. The adaptive margin m_{ai} is formulated by a positive sample (a, i) .

V. EXPERIMENTS AND RESULTS

TABLE I
STATISTICS OF TWO BENCHMARKS

Dataset	Epinions	Ciao
Users	18,202	7,375
Items	47,449	91,091
Ratings	298,173	226,307
Links	381,559	111,781
Rating Density	0.035%	0.034%
Link Density	0.115%	0.206%

A. Datasets

We evaluate our proposed model and other baselines on two popular product review-based social platforms [5]: Epinions and Ciao. To know which items should be recommended, we convert the ratings to binary form, if the rating value less than 3 we will take it as negative feedback and the remaining is positive. For training phase, we randomly separate the dataset into 80 : 20 with 80% interactions are training set and 20% interactions are test data. For more detail, we present a summary of two benchmarks in Table I.

B. Baselines and Evaluation Metrics

We compare the performance of HMPGSR with recent recommendation methods in both Euclidean and hyperbolic space. Two Euclidean approaches are LightGCN [7] and DiffNet++ [8], while two hyperbolic models consist of HGCF [11] and HGSR [5].

We concentrate on the item ranking task, and use Recall@N and NDCG@N to assess the recommendation performances of different approaches. In particular, NDCG@N assigns a higher score to the N top-ranked items, while Recall@N

TABLE II
EXPERIMENTAL RESULTS WITH DIFFERENT TOP-N VALUES ON EPINIONS DATASET

Models	N=10		N=20		N=30		N=40		N=50	
	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
LightGCN	0.0428	0.0310	0.0665	0.0380	0.0844	0.0428	0.0994	0.0467	0.1123	0.0499
DiffNet++	0.0427	0.0306	<u>0.0667</u>	0.0377	0.0838	0.0423	0.0982	0.0461	0.1113	0.0493
HGCF	0.0392	0.0288	0.0621	0.0356	0.0797	0.0404	0.0946	0.0444	0.1071	0.0474
HGSR	<u>0.0491</u>	<u>0.0351</u>	<u>0.0667</u>	<u>0.0428</u>	<u>0.0953</u>	<u>0.0483</u>	<u>0.1117</u>	<u>0.0525</u>	<u>0.1258</u>	<u>0.0559</u>
HMPGSR	0.0511	0.0365	0.0795	0.0448	0.1011	0.0506	0.1187	0.0551	0.1336	0.0587
Gain(%)	+4.07	+3.99	+19.19	+4.67	+6.09	+4.76	+6.27	+4.95	+6.20	+5.01

TABLE III
EXPERIMENTAL RESULTS WITH DIFFERENT TOP-N VALUES ON CIAO DATASET

Models	N=10		N=20		N=30		N=40		N=50	
	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
LightGCN	0.0348	0.0295	0.0542	0.0349	0.0676	0.0388	0.0784	0.0418	0.0866	0.0439
DiffNet++	0.0353	0.0302	0.0545	0.0354	0.0674	0.0392	0.0790	0.0423	0.0871	0.0444
HGCF	0.0305	0.0266	0.0497	0.0322	0.0644	0.0366	0.0764	0.0398	0.0854	0.0422
HGSR	<u>0.0394</u>	<u>0.0336</u>	<u>0.0612</u>	<u>0.0399</u>	<u>0.0750</u>	<u>0.0439</u>	<u>0.0873</u>	<u>0.0472</u>	<u>0.0971</u>	<u>0.0497</u>
HMPGSR	0.0395	0.0341	0.0614	0.0401	0.0755	0.0443	0.0876	0.0474	0.0974	0.0497
Gain(%)	+0.25	+1.49	+0.33	+0.50	+0.67	+0.91	+0.34	+0.42	+0.31	+0.00

calculates the proportion of correctly identified items in the top N recommendations. For this setting, we will choose all users who do not have an interactive history to compute all evaluation metrics. After three iterations of the experiment, the average values for each metric are provided.

C. Parameter Settings

We set the embedding dimension to 64 and all model embeddings are initialized with a Gaussian distribution with a mean of 0 and a standard deviation of 0.01. For the baselines, we use the configurations stated in the original studies for fair comparisons. In our proposed HMPGSR, curvature parameter of Poincaré ball c is set to -1 and the number of propagation layers is 3. Riemannian SGD [16] is utilized for training HMPGSR with weight decay $1e^{-5}$, and learning rate $1e^{-3}$.

D. Results

Overall results. The results on two benchmarks are reported in Tables II and III. The best results are in **bold** and the second-best results are in underline for each dataset. On both datasets, Euclidean models are outperformed by all hyperbolic recommendation models. On Epinions, our proposed method consistently achieves the highest performance across all baselines. For instance, HMPGSR achieves gains of 19% in the Recall@20 metric compared to the strongest Euclidean-space baseline, DiffNet++. Furthermore, our model makes notable improvements compared to the best hyperbolic social recommendation model, HGSR. For example, it improves Recall@20 and NDCG@20 by 19.19% and 4.67%, respectively. On the Ciao dataset, our model maintains robust performance against all baselines. These findings demonstrate that HMPGSR effectively leverages both user-item interactions and social network information within hyperbolic space,

highlighting the benefits of manifold-preserving properties for social recommendation tasks.

TABLE IV
THE IMPACT OF MANIFOLD-PRESERVING PROPERTY IN EACH HYPERBOLIC MODEL OF HMPGSR

Models	Epinions		Ciao	
	NDCG@20	Recall@20	NDCG@20	Recall@20
HMPGSR	0.0795	0.0448	0.0614	0.0401
w/o $\mathcal{K}$	0.0651	0.0367	0.0463	0.0286
Gain(%)	-18.11	-18.08	-24.59	-28.68
w/o $\mathcal{B}$	0.0730	0.0432	0.0602	0.0387
Gain(%)	-8.18	-3.57	-1.95	-3.49
w/o $\mathcal{L}$	0.0623	0.0370	0.0464	0.0298
Gain(%)	-21.64	-17.41	-24.43	-25.69

Effectiveness of manifold-preserving property. We assess the effectiveness of manifold-preserving property by performing an ablation study on three variants derived from HMPGSR:

- **w/o $\mathcal{K}$:** Removing the Klein model in HMPGSR. This variant employs diffusion process in the tangent space to propagate neighborhood features.
- **w/o $\mathcal{B}$:** Removing the Poincaré ball in HMPGSR. The hidden embeddings are fed into sum pooling layer in tangent space.
- **w/o $\mathcal{L}$:** Removing the Lorentz model in HMPGSR. The final graph representation is projected to tangent space for preference prediction and optimization.

We compare three variants in Top-20 ranking recommendation task with the same configuration as the original HMPGSR. The ablation results are reported in Table IV. In general, discarding any hyperbolic manifold causes a significant decrease in performance. The Klein and Lorentz models are especially important to the learning capacity of HMPGSR. **w/o $\mathcal{K}$** leads to an average drop of 22.37% in performance because

the hierarchical information of heterogeneous graph is not preserved in the tangent space. **w/o** $\mathcal{L}$ makes the performance decline 22.23% in average, which demonstrates the negative effect of tangent distortion in optimizing hyperbolic models. Similarly, in the case of **w/o** $\mathcal{B}$, the model's performance also drops on both datasets. These observations highlight the cruciality of manifold-preserving in hyperbolic graph learning.

TABLE V
COMPARISON OF THE AVERAGED TRAINING TIME (SECOND/EPOCH)

Models	Epinions	Ciao
LightGCN	4.4(s)	2.6(s)
DiffNet++	5.2(s)	3.3(s)
HGCF	2.2(s)	1.7(s)
HGSR	2.9(s)	2.0(s)
HMPGSR	2.8(s)	1.9(s)

Runtime comparison. For a better complexity comparison, we provide the running time evaluation between our HMPGSR and the baselines in Table V. The results indicate that, despite adopting a similar two-stage design, HGSR exhibits slower performance than our proposed model. This highlights that manifold preserving framework does not introduce computational overhead compared to tangent approaches, while still delivering superior performance.

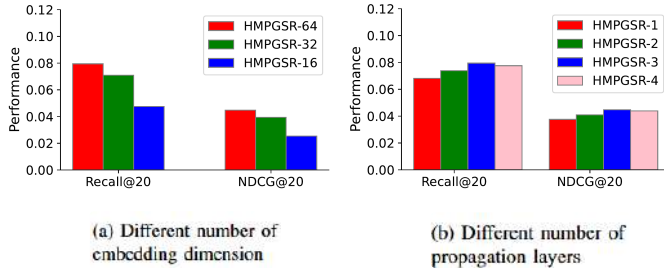


Fig. 2. Performance comparisons under different embedding sizes and number of propagating layers.

Hyperparameters sensitivity. To explore the influence of different embedding sizes and different number of propagation layers of the graph encoder, we conduct additional experiments on the Epinions dataset, with the results shown in Figure 2. The embedding dimensions are selected from $\{16, 32, 64\}$, and the number of propagation layers changes from 1 to 4. As shown in Figure 2a, HMPGSR's performance improves significantly with larger embedding dimensions. In Figure 2b, when the number of propagation layers increases, HMPGSR's performance initially increases but declines beyond a certain depth, suggesting that deeper propagation layers introduce over-smoothing.

VI. CONCLUSION

In this work, we propose a novel manifold-preserving method HMPGSR for social recommendation. By integrating three geometric models, HMPGSR operates entirely within

hyperbolic space without relying on tangent space approximations. HMPGSR consists of two stages, enabling it to effectively capture high-order and complex relationships between users and items. Extensive experiments on real-world datasets show that HMPGSR consistently surpasses the existing hyperbolic recommendation methods.

REFERENCES

- [1] Y. Li, K. Liu, R. Satapathy, S. Wang, and E. Cambria, "Recent developments in recommender systems: A survey [review article]," *IEEE Computational Intelligence Magazine*, vol. 19, pp. 78–95, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259224532>
- [2] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," *2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272, 2008. [Online]. Available: <https://api.semanticscholar.org/CorpusID:10537313>
- [3] R. Zhang, Q. dong Liu, C. Gui, J. Wei, and H. Ma, "Collaborative filtering for recommender systems," *2014 Second International Conference on Advanced Cloud and Big Data*, pp. 301–308, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14151737>
- [4] W. Fan, Y. Ma, Q. Li, Y. He, Y. E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," *The World Wide Web Conference*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:67769538>
- [5] Y. Yang, L. Wu, K. Zhang, R. Hong, H. Zhou, Z. Zhang, J. Zhou, and M. Wang, "Hyperbolic graph learning for social recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, pp. 8488–8501, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:266325212>
- [6] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3144218>
- [7] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:211043589>
- [8] L. Wu, J. Li, P. Sun, R. Hong, Y. Ge, and M. Wang, "Diffnet++: A neural influence and interest diffusion network for social recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, pp. 4753–4766, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:211010740>
- [9] G. Pickering, J. M. Bull, and D. J. Sanderson, "Sampling power-law distributions," *Tectonophysics*, vol. 248, pp. 1–20, 1995. [Online]. Available: <https://api.semanticscholar.org/CorpusID:8795640>
- [10] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá, "Hyperbolic geometry of complex networks," *Phys. Rev. E*, vol. 82, p. 036106, Sep 2010. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevE.82.036106>
- [11] J. Sun, Z. Cheng, S. Zuberi, F. Pérez, and M. Volkovs, "Hgc: Hyperbolic graph convolution networks for collaborative filtering," *Proceedings of the Web Conference 2021*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:235324826>
- [12] V. Q. Le and V.-C. Ta, "Toward a manifold-preserving temporal graph network in hyperbolic space," in *International Joint Conference on Artificial Intelligence*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:271497627>
- [13] J. C. William, W. J. Floyd, R. Kenyon, Walter, and R. Parry, "Hyperbolic geometry," *Mathematical Sciences Research Institute (MSRI) Publications*, 1997.
- [14] A. Ungar, "Hyperbolic trigonometry and its application in the poincaré ball model of hyperbolic geometry," *Computers & Mathematics with Applications*, vol. 41, pp. 135–147, 04 2000.
- [15] M. Yang, Z. Li, M. Zhou, J. Liu, and I. King, "Hicf: Hyperbolic informative collaborative filtering," *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:250643957>
- [16] S. Bonnabel, "Stochastic gradient descent on riemannian manifolds," *IEEE Transactions on Automatic Control*, vol. 58, pp. 2217–2229, 2011. [Online]. Available: <https://api.semanticscholar.org/CorpusID:429631>

SCIGENT: Multi-agent for End-to-end Question Answering on Scientific Knowledge Graph

Dieu-Hien Nguyen
JAIST

Ishikawa, Japan
ndhien@jaist.ac.jp

Nguyen-Khang Le
JAIST

Ishikawa, Japan
lnkhang@jaist.ac.jp

Shinobu Hasegawa
JAIST

Ishikawa, Japan
hasegawa@jaist.ac.jp

Minh Le Nguyen
JAIST

Ishikawa, Japan
nguyenlm@jaist.ac.jp

Abstract—Effective question answering over scientific literature demands up-to-date knowledge, a challenge exacerbated by the rapid growth of publications. Traditional Retrieval-Augmented Generation (RAG) methods, reliant on text-based retrieval, struggle with structured knowledge inherent in knowledge graphs. We introduce SCIGENT, a multi-agent platform for end-to-end question answering over scientific knowledge graphs, addressing the challenge of dynamic entity resolution and SPARQL query translation from raw user queries. Unlike traditional methods, SCIGENT operates in open-ended scenarios, resolving entities and relationships in real-time. Alongside this, we present SCIGENT-QA, a benchmark built on DBLP and ORKG knowledge graphs, to evaluate end-to-end question-answering systems. This benchmark provides a comprehensive testbed for assessing real-world applicability and robustness. Additionally, we release a full pipeline for constructing and continuously updating scientific knowledge graphs, ensuring the system remains current with emerging publications. This work advances knowledge-driven AI, enabling robust and scalable solutions for complex scientific queries. Empirical results demonstrate that SCIGENT, when used with popular large language models, achieves up to 65% accuracy on the end-to-end evaluation of SCIGENT-QA, proving its practicality for real-world applications.

Index Terms—Knowledge Graph, Chatbot Application, Question-Answering.

I. INTRODUCTION

In the rapidly evolving field of artificial intelligence, chatbot agents have become an essential tool for retrieving information and assisting users in diverse domains. Consider a user querying a chatbot with: “Find the papers written by Yann LeCun that were published in ICLR.” To fulfill this request, the chatbot must possess the necessary knowledge to answer the query accurately. However, given the continuous expansion of scientific literature, maintaining an up-to-date knowledge base presents a significant challenge. One common approach to addressing this issue is Retrieval-Augmented Generation (RAG), which augments the model’s input with relevant information retrieved from a textual knowledge source. While effective for text-based knowledge, this approach falls short when the required information is structured as entities and their relationships, as seen in knowledge graphs.

Knowledge graphs provide a structured representation of entities and their relationships, making them a powerful tool for answering complex queries. To accurately retrieve information from a knowledge graph, a query must be translated into a specialized query language such as SPARQL. In the given

example, the chatbot needs to resolve entity representations for “Yann LeCun” and “ICLR” before constructing a valid SPARQL query. However, traditional benchmarks and prior works predominantly focus on translating structured queries into SPARQL, assuming that the entity representations are readily available. In real-world applications, chatbots operate in an open-ended context where user queries lack explicit entity annotations, making end-to-end question answering over knowledge graphs a more complex challenge.

To bridge this gap, we introduce SCIGENT, a multi-agent platform designed for end-to-end question answering over scientific knowledge graphs. Unlike existing approaches, SCIGENT addresses the challenge of entity resolution in an open-ended setting, ensuring that entities and relationships are inferred dynamically before query translation. Alongside SCIGENT, we propose SCIGENT-QA, a new benchmark designed to evaluate end-to-end question answering on scientific knowledge graphs. Our benchmark is built upon well-established knowledge bases such as DBLP and ORKG, providing a comprehensive testbed for evaluating real-world applicability. Furthermore, we release a complete pipeline for constructing and maintaining scientific knowledge graphs, allowing for continuous updates as new publications emerge. By providing this platform and benchmark, we aim to advance the capabilities of knowledge-driven AI systems and facilitate more effective question-answering solutions in scientific domains.

The contribution of this paper is as follows.

- We introduce SCIGENT-QA, a benchmark and dataset for evaluating end-to-end scientific question answering over knowledge graphs, built on well-established knowledge bases such as DBLP and ORKG.
- We present SCIGENT, a multi-agent chatbot system designed to assist users in querying scientific knowledge, leveraging sources like DBLP and ORKG while supporting continuous updates as new publications emerge.
- Empirical results demonstrate that SCIGENT, when integrated with various large language models (LLMs), achieves promising performance on SCIGENT-QA, highlighting its practical applicability. Additionally, the system shows strong results in entity/relation extraction and SPARQL translation tasks, underscoring its robustness and effectiveness.

TABLE I
COMPARISON OF SCIGENT-QA AND EXISTING KNOWLEDGE GRAPH DATASETS. THESE DATASETS INVOLVE DIFFERENT ASPECTS OF EVALUATION: ENTITY AND RELATION EVALUATION (ENTITY-EVAL), SPARQL EVALUATION (SPARQL-EVAL), AND END-TO-END EVALUATION (E2E-EVAL).

Dataset	Knowledge Source	Manner	Size	Entity-eval	SPARQL-eval	E2E-eval
<i>Wiki Domain</i>						
KQA Pro [7]	Wikidata	Manual+Auto	117,000	✗	✓	✓
LC-QuAD 1.0 [19]	DBpedia	Manual+Auto	5,000	✗	✓	✗
LC-QuAD 2.0 [8]	DBpedia+Wiki	Manual+Auto	30,000	✗	✓	✗
<i>General Domain</i>						
Free917 [6]	Freebase	Manual	917	✗	✗	✗
QALD [20]	Freebase	Manual	806	✗	✓	✓
WebQuestionsSP [22]	Freebase	Manual+Auto	-	✗	✓	✗
ComplexWebQues [18]	Freebase	Manual+Auto	34,689	✗	✓	✗
GraphQuestions [17]	Freebase	Manual+Auto	5,166	✗	✗	✗
GraillQA [9]	Freebase	Manual+Auto	64,331	✗	✓	✗
CFQ [12]	Freebase	Auto	239,357	✗	✓	✗
<i>Scientific Domain</i>						
ORKG-QA [10]	ORKG	Auto	100	✗	✗	✗
SciQA [3]	ORKG	Auto	2,565	✗	✓	✗
DBLP-Quad [5]	DBLP	Auto	10,000	✗	✓	✗
SCIGENT-QA	DBLP+ORKG	Auto	11,500	✓	✓	✓

II. PRELIMINARIES

A. Knowledge Graph Representation

A *Knowledge Graph* (KG) is a directed labeled graph $G = (V, E, R)$, where V is the set of entities (nodes), R the set of relations (predicates), and $E \subseteq V \times R \times V$ the set of edges (triples). Each triple $(h, r, t) \in E$ represents a fact, with head entity h , tail entity t , and relation r .

DBLP¹ and ORKG² are bibliographic databases modeled as knowledge graphs. Nodes represent entities such as *Author*, *Paper*, and *Conference*, while relations include predicates like *authoredBy* and *publishedIn*. For example, the triple ("A Study on Knowledge Graphs", *authoredBy*, John_Doe) indicates that John Doe authored the specified paper. This structure facilitates efficient querying and reasoning over academic data.

B. SPARQL: Query Language for Knowledge Graphs

SPARQL is the standard query language for RDF knowledge graphs, enabling information retrieval via graph pattern matching. A typical query uses **SELECT** to specify variables and **WHERE** to define patterns. For example, to find papers by <author> published in <venue> in DBLP:

```
SELECT DISTINCT ?answer WHERE {
  ?answer
  <https://dblp.org/rdf/schema#authoredBy>
  <author> .
  ?answer
  <https://dblp.org/rdf/schema#publishedIn>
  <venue>
}
```

SPARQL supports various query forms—**SELECT**, **ASK**, **CONSTRUCT**, and **DESCRIBE**—providing flexible means to retrieve, transform, and analyze knowledge graphs, making it well-suited for large-scale datasets like DBLP and ORKG.

¹<https://dblp.org>

²<https://orkg.org>

III. SCIGENT-QA DATA GENERATION

A. Overview

This work presents a pipeline for generating evaluation data for end-to-end QA over scientific knowledge graphs. As shown in Table I, SCIGENT-QA differs from existing datasets (e.g., ORKG-QA, SciQA, DBLP-Quad) by integrating both DBLP and ORKG, offering broader coverage. It also uniquely supports entity-level, SPARQL-level, and end-to-end evaluation, enabling more comprehensive benchmarking.

B. Data Collection Process

Figure 1 shows the pipeline for generating SCIGENT-QA. We start with 98 SPARQL and question templates, adapted from prior work [5].

DBLP and ORKG are normalized into a unified RDF knowledge graph with triplets (h, r, t) , using DBLP's schema. To generate a single QA instance, a template is randomly selected from the store. Each template contains a question and a SPARQL query with placeholders. These placeholders are instantiated using a subgraph sampled around a central publication node, including related authors, venues, and other metadata. Author nodes are further linked to affiliation information and attributes such as primary names. The populated template yields a complete question-answer pair.

To ensure correctness, we validate each SPARQL query by executing it against our SPARQL endpoint. We also derive bidirectional mappings between human-readable names (e.g., paper titles, author names) and entity IDs. Specifically, from the knowledge graph $G = (V, E, R)$ (see Section II-A), we extract mappings for $\#title$ and $\#primaryCreatorName$ relations:

$$\begin{aligned}
 E_{\text{title}} &= \{(h, r, t) \in E \mid r = \#title\} \\
 F_{\text{title}}(h) &= \{t \mid (h, \#title, t) \in E_{\text{title}}\} \\
 F_{\text{title}}^{-1}(t) &= \{h \mid (h, \#title, t) \in E_{\text{title}}\}
 \end{aligned}$$

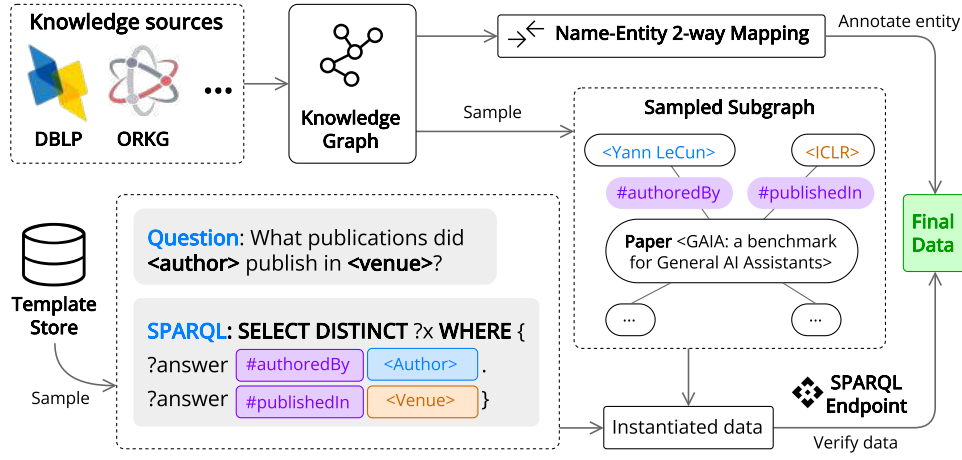


Fig. 1. Data generation process in SCIGENT-QA

$$\begin{aligned}
 E_{\text{name}} &= \{(h, r, t) \in E \mid r = \#primaryCreatorName\} \\
 F_{\text{name}}(h) &= \{t \mid (h, \#primaryCreatorName, t) \in E_{\text{name}}\} \\
 F_{\text{name}}^{-1}(t) &= \{h \mid (h, \#primaryCreatorName, t) \in E_{\text{name}}\}
 \end{aligned}$$

These mappings annotate entities with names, completing the data collection process.

C. Data Splits

The automated pipeline enables large-scale data generation without human annotation. We create a dataset of 11,500 instances, split into 7,000 for training, 1,000 for validation, and 3,500 for testing, supporting zero-shot, few-shot, and fine-tuning evaluations. In SCIGENT, only the training set is used as few-shot exemplars for prompting, without model fine-tuning, ensuring flexibility and fair comparison.

IV. SCIGENT

A. Conversation Handling

User input is first classified by the Conversation Agent as either casual or knowledge-seeking. Casual inputs receive chatbot-style replies, while structured queries are routed to specialized modules. This enables SCIGENT to handle both informal dialogue and complex information retrieval, ensuring a unified conversational experience.

B. Multi-Agent Pipeline

Figure 2 shows the SCIGENT workflow. The Graph Agent first extracts entities (e.g., authors, papers, venues) and their relations from a user query. These are sent to the SPARQL Agent, which constructs and executes a SPARQL query. The Conversation Agent then converts the query and its results into a natural language response, ensuring smooth interaction over the scientific knowledge graph.

1) *Graph Agent*: The Graph Agent uses a LLM with few-shot prompting to extract entities and relations from user queries. Relevant examples are selected via BM25 based on lexical overlap (Table II). Extracted entities are linked to the knowledge graph using mappings F_{name}^{-1} and F_{title}^{-1} . The identified graph entities and relations are then passed to the SPARQL Agent for SPARQL query generation.

TABLE II

PROMPTS USED IN SCIGENT'S AGENTS. PROMPTS ARE FORMATTED USING THE CHAT TEMPLATE SPECIFIC TO EACH LLM.

Prompt for Entity Extraction

System Prompt: You are a helpful AI assistant. Given a sentence, extract named entities of persons, paper titles, or venues, if present. Answer in the format:
persons: ['<person_1>', '<person_2>', ...]
titles: ['<title_1>', '<title_2>', ...]

Few-shot Examples

User: Which papers did John Doe publish in ICLR?

Assistant:

persons: ['John Doe']

titles: ['ICLR']

End few-shot examples

Query

User: What is the ORCID of the person Yann LeCun?

Prompt for Relation Extraction

System Prompt: You are an AI assistant. For each question, identify relevant relations from the following list. If none are applicable, respond with None.
#wikidata, #authoredBy, #publishedIn, #affiliation, ...

Few-shot Examples

User: Which papers did John Doe publish in ICLR?

Assistant: ['#authoredBy', '#publishedIn']

End few-shot examples

Query

User: What is the ORCID of the person Yann LeCun?

Prompt for SPARQL Translation

System Prompt: You are an expert in translating natural language questions into SPARQL queries. Each example includes a question, relevant entities, and relations.

Few-shot Examples

User: What is the primary affiliation of the author ...?

Assistant:

```
SELECT DISTINCT ?ans WHERE { <author_uri>
<#affiliation> ?ans }
```

End few-shot examples

Query

User: What is the ORCID of the person Yann LeCun?

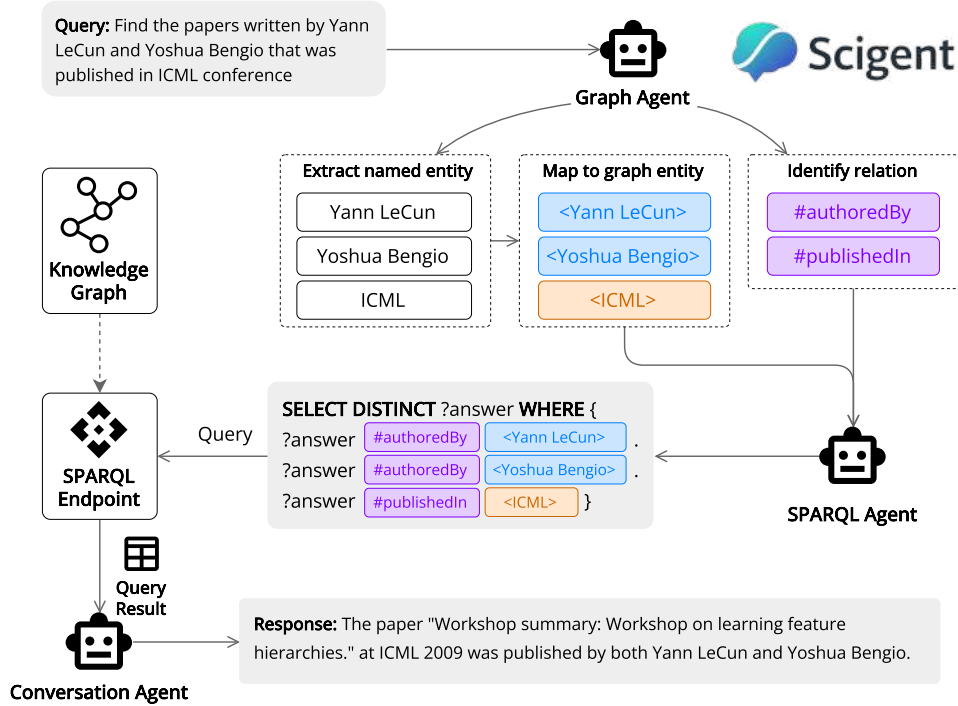


Fig. 2. Overview of SCIGENT’s multi-agent framework

2) *SPARQL Agent*: The SPARQL Agent also relies on an LLM with few-shot prompting, using BM25 to retrieve relevant examples (Table II, bottom). It generates a response containing a SPARQL query, from which the query is post-processed and extracted for execution.

3) *Conversation Agent*: The Conversation Agent processes user inputs, distinguishing between casual dialogue and knowledge graph queries. For queries, it coordinates with other agents to retrieve information from the SPARQL endpoint. Using the mappings F_{name} and F_{title} , it translates graph entities into human-readable names before generating and delivering a natural language response.

4) *Knowledge Graph Update*.: SCIGENT’s knowledge graph is designed for continuous updates, ensuring it remains synchronized with evolving knowledge sources such as DBLP and ORKG. The system can be scheduled to periodically fetch new data dumps from these sources and rebuild the knowledge graph. This update process is flexible and can be extended to incorporate additional data sources beyond DBLP and ORKG, enhancing the system’s coverage and relevance.

C. User Interface

Figure 3 provides a screenshot of SCIGENT’s user interface. The system features an end-to-end chatbot interface designed to enable seamless querying of scientific domain information. When a user submits a query, the system dynamically resolves entities and relationships, generates the corresponding SPARQL query, and retrieves the requested information from the scientific knowledge graph. The results are displayed

in a structured table format, accompanied by the generated SPARQL query for transparency and user understanding.

This end-to-end functionality ensures an intuitive and user-friendly experience, allowing users to interact with complex scientific knowledge without requiring technical expertise. The interface is designed for practicality, supporting real-time updates and contextual refinements, making it a robust tool for researchers and practitioners seeking accurate and up-to-date scientific information.

V. EXPERIMENTS

1) *Models*: The backbone of SCIGENT is an LLM. We evaluate several publicly available, competitive pre-trained LLMs, including Qwen2.5-Instruct (7B), QwQ-Preview (32B) [4], Llama3.1-Instruct (8B) [2], Phi-4 [1], CodeLlama-Instruct (7B) [15], and Mistral-Instruct-v0.3 (7B) [11].

2) *Tasks*: We evaluate SCIGENT on the end-to-end task of SCIGENT-QA, which requires answering scientific questions based on a knowledge graph. We further assess the Graph Agent’s performance on an entity-relation extraction task using the SCIGENT-QA dataset. To evaluate generalization capability, we also test the SPARQL Agent on DBLP-Quad [5], a benchmark designed for SPARQL query translation.

3) *Metrics*: For the end-to-end QA task, we use Exact Match (EM) as the metric, requiring an exact match between the predicted and gold answers. For evaluating entity and relation extraction, we report Precision, Recall, and F1 score. SPARQL translation performance on DBLP-Quad is measured using EM, after normalizing variable names and syntax.

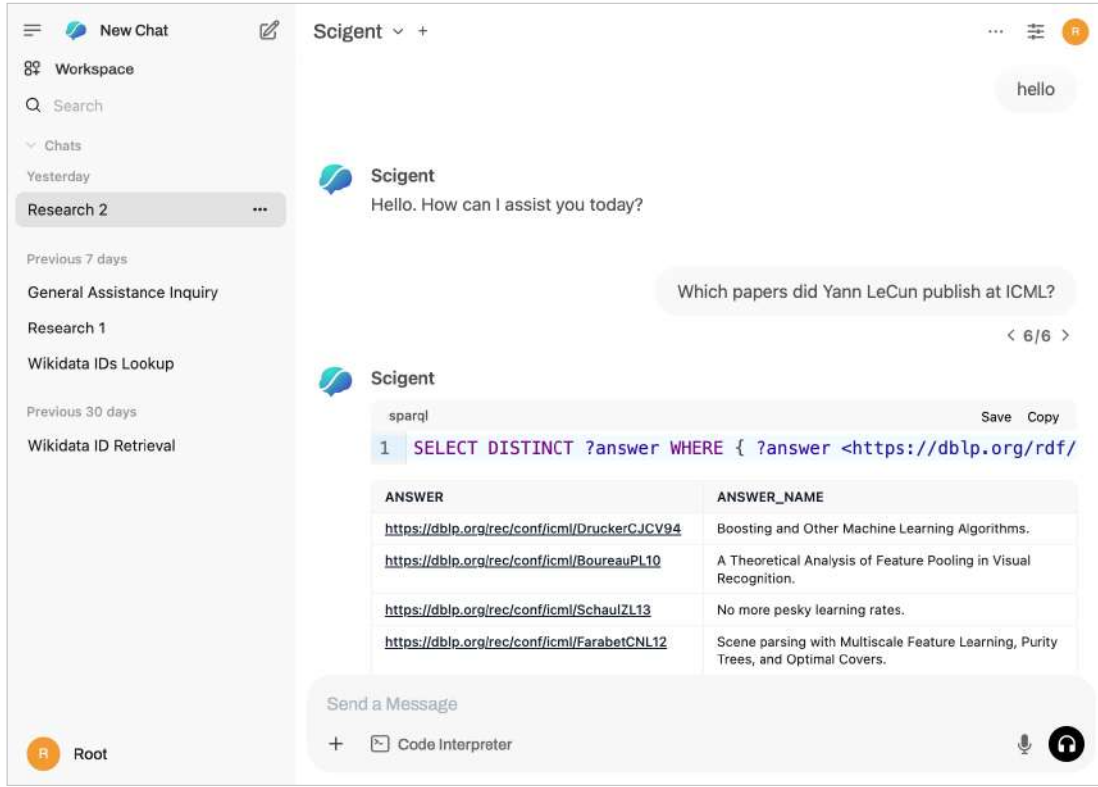


Fig. 3. Screenshot of SCIGENT user interface.

VI. RESULTS

1) *Results on SCIGENT-QA:* Table III presents the end-to-end performance of various models on SCIGENT-QA. QwQ-32B-Preview achieves the highest EM score, highlighting its strong capabilities in querying knowledge graphs. Phi-4 also performs competitively. Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct achieve similar results despite smaller size. In contrast, Mistral-7B-Instruct-v0.3 and CodeLlama-7B-Instruct demonstrate significantly lower performance, suggesting limitations of general-purpose instruction tuning for this task.

TABLE III
END-TO-END PERFORMANCE ON SCIGENT-QA WITH VARIOUS LLM BACKBONES. THE SAME LLM IS USED ACROSS ALL AGENTS.

SCIGENT Backbone	EM
Qwen2.5-7B-Instruct	55.34
Llama-3.1-8B-Instruct	54.89
Llama-3.2-3B-Instruct	33.71
QwQ-32B-Preview	64.14
Phi-4	61.69
CodeLlama-7B-Instruct	2.69
Mistral-7B-Instruct-v0.3	25.06

2) *Entity and Relation Evaluation:* Table IV shows entity and relation extraction results on SCIGENT-QA. QwQ-32B and Phi-4 lead in F1. Qwen2.5-7B performs strongly, especially in relation precision. Llama3.1-8B excels at entity recognition but lags in relations. CodeLlama-7B performs poorly; Mistral-7B yields moderate results.

TABLE IV
ENTITY AND RELATION EVALUATION ON SCIGENT-QA, REPORTED AS PRECISION (PREC), RECALL (REC), AND F1 SCORE.

Graph Agent Backbone	F1	Entity Rec	Prec	F1	Relation Rec	Prec
Qwen2.5-7B-Instruct	80.73	81.42	80.37	96.77	96.87	97.36
Llama-3.1-8B-Instruct	79.37	80.27	78.91	86.98	86.79	87.53
QwQ-32B-Preview	81.17	81.47	81.01	97.29	97.67	97.56
Phi-4	81.65	81.90	81.52	96.80	96.81	97.31
CodeLlama-7B-Instruct	33.52	31.44	38.05	10.47	10.43	10.70
Mistral-7B-Instruct-v0.3	51.67	49.83	55.77	81.75	81.69	82.80

3) *Results on SPARQL Translation:* Table V reports SPARQL translation accuracy on SCIGENT-QA and DBLP-Quad. QwQ-32B-Preview leads, followed by Phi-4 and Llama-3.1-8B. Mistral-7B performs moderately, Qwen2.5-7B is slightly lower, and CodeLlama-7B struggles, highlighting its limitations in structured query generation.

VII. RELATED WORK

ORKG-QA [10] is an early scholarly KGQA dataset based on the ORKG API, organizing publications into comparison tables but offering limited QA coverage. QA datasets exist for both IR-based [13], [14] and KGQA tasks [16], [19], with the latter differing primarily by the inclusion of logical forms, which often require expert annotation.

Manually curated datasets like Free917 [6] and QALD [20] offer small yet high-quality samples. WebQuestionsSP [22] and ComplexWebQuestions [18] extend prior datasets by

TABLE V
SPARQL TRANSLATION ACCURACY (EM) OF THE SPARQL AGENT ON
SCIGENT-QA AND DBLP-QUAD [5].

Backbone	SCIGENT-QA	DBLP-Quad
Qwen2.5-7B-Instruct	71.29	74.65
Llama-3.1-8B-Instruct	76.86	78.35
Llama-3.2-3B-Instruct	66.26	68.05
QwQ-32B-Preview	84.17	87.50
Phi-4	78.57	80.25
CodeLlama-7B-Instruct	48.94	54.40
Mistral-7B-Instruct-v0.3	75.11	77.35

refining questions and generating complex SPARQL queries, respectively, followed by manual verification.

The OVERNIGHT framework [21] creates semantic parsing datasets by extracting logical forms from KGs, converting them into canonical questions, and rewriting via crowdsourcing. Several datasets follow this approach: GraphQuestions [17] (Freebase), LC-QuAD 1.0 [19] and 2.0 [8] (DBpedia/Wikidata), GrailQA [9] (Freebase Commons), KQA Pro [7] (Wikidata), and DBLP-Quad [5] (DBLP). CFQ [12] is a synthetic dataset designed to evaluate compositional generalization using handcrafted grammar and inference rules.

VIII. CONCLUSIONS

This study presents SCIGENT, a novel multi-agent framework for end-to-end question answering over scientific knowledge graphs, tackling key challenges such as dynamic entity resolution and SPARQL generation. Alongside, we introduce SCIGENT-QA, a benchmark dataset built on DBLP and ORKG, supported by a pipeline for continuous knowledge graph construction and updates.

Together, these contributions advance knowledge-driven AI by enabling robust, scalable, and real-time scientific QA. SCIGENT supports open-ended queries and achieves strong performance on SCIGENT-QA when paired with standard LLMs, offering a foundation for improved information retrieval and reasoning in scientific domains.

REFERENCES

- [1] Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R.J., Javaheripi, M., Kauffmann, P., Lee, J.R., Lee, Y.T., Li, Y., Liu, W., Mendes, C.C.T., Nguyen, A., Price, E., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Wang, X., Ward, R., Wu, Y., Yu, D., Zhang, C., Zhang, Y.: Phi-4 technical report (2024), <https://arxiv.org/abs/2412.08905>
- [2] et al., D.: The Llama 3 Herd of Models (2024), <https://arxiv.org/abs/2407.21783>
- [3] Auer, S., Barone, D.A.C., Bartz, C., Cortes, E.G., Jaradeh, M.Y., Karras, O., Koubarakis, M., Mourmoutsev, D., Pliukhin, D., Radyush, D., Shilin, I., Stocker, M., Tsalapati, E.: The scia scientific question answering benchmark for scholarly knowledge. *Scientific Reports* **13**(1), 7240 (May 2023). <https://doi.org/10.1038/s41598-023-33607-z>, <https://doi.org/10.1038/s41598-023-33607-z>
- [4] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen Technical Report (2023)
- [5] Banerjee, D., Awale, S., Usbeck, R., Biemann, C.: Dbp-quad: A question answering dataset over the dblp scholarly knowledge graph. In: BIR@ECIR. pp. 37–51 (2023), <https://ceur-ws.org/Vol-3617/paper-05.pdf>
- [6] Cai, Q., Yates, A.: Large-scale semantic parsing via schema matching and lexicon extension. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 423–433 (2013)
- [7] Cao, S., Shi, J., Pan, L., Nie, L., Xiang, Y., Hou, L., Li, J., He, B., Zhang, H.: Kqa pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. pp. 6101–6119 (2022)
- [8] Dubey, M., Banerjee, D., Abdelkawi, A., Lehmann, J.: Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In: The Semantic Web – ISWC 2019. pp. 69–78. Springer (2019)
- [9] Gu, Y., Kase, S., Vanni, M., Sadler, B., Liang, P., Yan, X., Su, Y.: Beyond i.i.d.: Three levels of generalization for question answering on knowledge bases. In: Proceedings of the Web Conference 2021. pp. 3477–3488 (2021)
- [10] Jaradeh, M.Y., Stocker, M., Auer, S.: Question answering on scholarly knowledge graphs. In: International Conference on Theory and Practice of Digital Libraries. pp. 19–32. Springer (2020)
- [11] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavau, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B (2023)
- [12] Keyser, D., Schärli, N., Scales, H., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., et al.: Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713* (2020)
- [13] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al.: Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7**, 453–466 (2019)
- [14] Rajpurkar, P., Jia, R., Liang, P.: Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822* (2018)
- [15] Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bittou, J., Bhatt, M., Ferrer, C.C., Grattafiori, A., Xiong, W., Défossez, A., Copet, J., Azhar, F., Touvron, H., Martin, L., Usunier, N., Scialom, T., Synnaeve, G.: Code Llama: Open Foundation Models for Code (2024), <https://arxiv.org/abs/2308.12950>, [_eprint: 2308.12950](https://arxiv.org/abs/2308.12950)
- [16] Sen, P., Aji, A.F., Saffari, A.: Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv:2210.01613* (2022)
- [17] Su, Y., Sun, H., Sadler, B., Srivatsa, M., Gur, I., Yan, Z., Yan, X.: On generating characteristic-rich question sets for qa evaluation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 562–572 (2016)
- [18] Talmor, A., Berant, J.: The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643* (2018)
- [19] Trivedi, P., Maheshwari, G., Dubey, M., Lehmann, J.: Lc-quad: A corpus for complex question answering over knowledge graphs. In: The Semantic Web – ISWC 2017. pp. 210–218. Springer (2017)
- [20] Usbeck, R., Ngomo, A.C.N., Haarmann, B., Krithara, A., Röder, M., Napolitano, G.: 7th open challenge on question answering over linked data (qald-7). In: Semantic Web Challenges. pp. 59–69. Springer (2017)
- [21] Wang, Y., Berant, J., Liang, P.: Building a semantic parser overnight. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1332–1342 (2015)
- [22] Yih, W.T., Richardson, M., Meek, C., Chang, M.W., Suh, J.: The value of semantic parse labeling for knowledge base question answering. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 201–206 (2016)

Learning Parameter Efficient Coordination in Autonomous Vehicles with Online Knowledge Distillation

Vu Minh Le

VNU University of Engineering and Technology
Hanoi, Vietnam

Xuan Duy Ta

VNU University of Engineering and Technology
Hanoi, Vietnam

Viet Cuong Ta*

VNU University of Engineering and Technology
Hanoi, Vietnam
cuongtv@vnu.edu.vn

Abstract—Coordinated control of multiple autonomous vehicles (AVs) in complex traffic scenarios can be viewed as a Multi-Agent Reinforcement Learning (MARL) problem with centralized training and execution. While recent advances in MARL and Transformer-based architectures have shown promising results, their high model complexity limits applicability for real-world deployment in AV settings. In this work, we propose a parameter-efficient learning framework based on the Multi-Agent Transformer architecture. The framework employs a knowledge distillation approach with a teacher-student agent pair. The teacher agent is a large Transformer model architecture for learning AV coordination, which is then leveraged to distill knowledge into a smaller student model by a co-learning mechanism. Our proposed framework is able to unite the learning stability of large-scale models with the speed and scalability of compact policies and provide efficient, high-quality multi-agent coordination. Experimental results across six traffic environments with varying traffic densities, distilled student policy achieves performance levels that rival or exceed those of significantly larger models.

I. INTRODUCTION

Autonomous vehicle (AV) technology demands real-time decision-making under strict hardware and energy constraints. These challenges become even more critical in automated driving systems where multiple AVs must coordinate their actions in complex, dynamic environments [1], [2]. Therefore, one of the key challenges in designing multi-AV system is to reduce the complexity overhead with parameter-efficient models.

From the perspective of MARL problems, popular approaches such as QMIX [3] or MAPPO [4] only focus on the training efficiency. To improve the model efficiency, one has to carefully finetune the model size to balance the learning and the inference speed [5]. Transformer-based architectures have recently gained traction in MARL, with Multi-Agent Transformer (MAT) [6] emerging as a popular alternative to

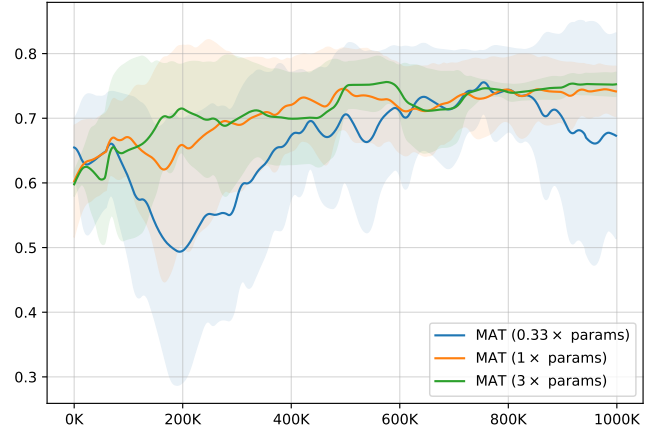


Fig. 1: Trade-off between parameter efficiency and performance in MAT variants. Higher parameter-counts correlate positively with lower variance, yet lower-parameter models ($0.33\times$) can outperform larger ones despite their high volatility.

traditional methods through reformulating the MARL problem as a sequence modeling problem. Despite its strong performance, MAT exhibits a crucial limitation regarding parameter efficiency. Testing AV-control performance using a simulator, shown in Fig. 1, smaller MAT variants (with only 33% of the original parameters) can occasionally outperform their larger counterparts, but their performance is highly volatile. Conversely, scaling up the model improves stability but comes at the cost of significantly increased computational overhead and may result in suboptimal performance. This trade-off between efficiency, stability, and peak performance highlights the need for a more parameter-efficient approach that preserves MAT’s expressiveness without incurring excessive cost.

In this work, we aim to tackle the challenge of effective

*Corresponding Author.

and parameter-efficient traffic control for AVs by proposing a simul-learning distillation paradigm, where the teacher model is not frozen during training. We integrate into MAT and leverage its powerful capabilities. We evaluate our approach across 6 scenarios with varying configurations, difficulty levels, and traffic density. The result demonstrates a stable, high-performing, and parameter-efficient policy that inherits both the robustness of the teacher and the efficiency and high performance potential of a lightweight model.

II. RELATED WORK AND BACKGROUNDS

MARL framework. We formulate the mixed traffic multi-autonomous vehicle control as an n -agent Partially Observable Markov Decision Process [7] defined by the tuple $\langle \mathcal{S}, \mathcal{O}, \mathcal{A}, P, r, \gamma, \omega \rangle$. Where, $\mathcal{S}$ is the global state space shared by all agents, while $\mathcal{O}$ and $\mathcal{A}$ is the observation and action space of any singular agent. The initial state at timestep 0, s_0 , is drawn from distribution ω . The discount factor $\gamma \in [0; 1]$ is used to control the future rewards.

At any state $s_t \in \mathcal{S}$ of timestep t , each agent i can only observe its local observation $o_t^i = O(s_t, i)$. The joint observation of all agents at timestep t is $\mathbf{o}_t = \langle o_t^1, o_t^2, \dots, o_t^n \rangle$. Each agent, using its own policy π^i to choose an action $a_t^i \in \mathcal{A}$. After executing the joint action $\mathbf{a}_t = \langle a_t^1, a_t^2, \dots, a_t^n \rangle$, the next state s^{t+1} is selected following the distribution $P(s_t, \mathbf{a})$ and agents all receive the common scalar reward $r_t = r(s_t, \mathbf{a}, s_{t+1})$.

Let $\tau^i = \langle o_0^i, a_0^i, r_0, o_1^i, a_1^i, r_1, \dots, o_{H-1}^i, a_{H-1}^i, r_{H-1}, o_H^i \rangle$ be the trajectory of time horizon H induced by the i -th agent's policy π^i . The discounted return $R(\tau^i)$ of the i -th agent over trajectory τ^i is given by $R(\tau^i) = \sum_{h=0}^{H-1} \gamma^h r_h$. The objective is to learn the parameterized joint policy $\pi = \langle \pi^1, \pi^2, \dots, \pi^n \rangle \in \Pi$ that maximizes the expected discounted cumulative reward $\sum_{i=1}^n \mathbb{E}_{\tau^i \sim \pi^i} R(\tau^i)$.

MARL methods. Popular approaches are value-based and policy gradient-based methods. In value-based methods, QMIX [3] generalizes the additive mixing of VDN [8] using a monotonic value function decomposition. The authors of [9] extended the DDPG [10] method to the multi-agent setting via a centralized critic, introducing the Centralized Training with Decentralized Execution (CTDE) paradigm. MAPPO [4], extended from PPO [11] following the CTDE framework, is currently the most popular on-policy MARL method. MAPPO utilizes trust region optimization to stabilize training while minimizing the computational cost associated with Hessian calculation through a clipping mechanism.

Multi-Agent Transformer (MAT) [6] is a successful implementation of the encoder-decoder structure of the Transformer [12] into MARL. Let $[n]$ be the ordered set $\{1, 2, \dots, n\}$ and arbitrary disjoint, ordered subsets of agents be denoted as, for any $m \in [n]$, $i_{1:m} = \{i_1, \dots, i_m\}$.

The encoder, parameterized by ϕ , processes observation sequences $(o^{i_1}, \dots, o^{i_n})$ in arbitrary order to produce observational encodings $(\hat{o}^{i_1}, \dots, \hat{o}^{i_n})$ that captures both agent states and their interactions. In the training phase, the encoder is

made to approximate the value functions by minimizing the empirical Bellman error:

$$\mathcal{L}_{\text{Enc}}^{\text{MAT}}(\phi) = \mathbb{E}_{\mathbf{o}_t, \mathbf{a}_t, m \in [n]} [R(\mathbf{o}_t, \mathbf{a}_t) + \gamma V_{\bar{\phi}}(\hat{\mathbf{o}}_{t+1}^{i_m}) - V_{\phi}(\hat{\mathbf{o}}_t^{i_m})]^2, \quad (1)$$

where $\bar{\phi}$ is the target network's parameter.

The decoder, parameterized by θ , processes the embedded joint action sequence $\mathbf{a}^{i_{0:m-1}}$ (where $m \in [n]$ and a^{i_0} is a start-of-decoding symbol) through a series of decoding blocks. Vitally, each decoding block comes with 2 *masked* self-attention functions that enforce the sequential update scheme by computing attention only between the i_j^{th} and i_k^{th} heads where $j < k$. The first handles action-action interactions while the second computes the attention values between action-interactions and encoded observations. The decoder outputs a sequence of joint action probability distribution $\pi_{\theta}^{i_m}(a^{i_m} | \hat{\mathbf{o}}^{i_{1:n}}, \mathbf{a}^{i_{1:m-1}}) \forall m \in [n]$. Overloading the notation $\text{clip}(x, \epsilon) = \text{clip}(x, 1 \pm \epsilon)$, the decoder loss is:

$$\mathcal{L}_{\text{Dec}}^{\text{MAT}}(\theta) = -\mathbb{E}_{\mathbf{o}_t, \mathbf{a}_t, m \in [n]} \left[\min \left(\rho_t^{i_m}(\theta) \hat{A}_t, \text{clip}(\rho_t^{i_m}(\theta), \epsilon) \hat{A}_t \right) \right], \quad (2)$$

$$\rho_t^{i_m}(\theta) = \frac{\pi_{\theta}^{i_m}(a_t^{i_m} | \hat{\mathbf{o}}_t^{i_{1:n}}, \mathbf{a}_t^{i_{1:m-1}})}{\pi_{\theta_{\text{old}}}^{i_m}(a_t^{i_m} | \hat{\mathbf{o}}_t^{i_{1:n}}, \mathbf{a}_t^{i_{1:m-1}})},$$

where $\hat{A}_t$ represents the estimated joint advantage function, applying generalized advantage estimation (GAE) [13] with $\hat{V}_t = \frac{1}{n} \sum_{m=1}^n V(\hat{\mathbf{o}}_t^{i_m})$. During inference, actions are generated auto-regressively; while training leverages full parallelization by utilizing all stored actions $\mathbf{a}^{i_{1:n-1}}$ from the replay buffer to compute $\mathbf{a}^{i_{1:n}}$ simultaneously.

Knowledge Distillation. Knowledge distillation (KD) refers to the process of transferring learned representations or outputs from an often larger model (the teacher) to another, usually smaller one (the student). The original objective, introduced in [14], minimizes the Kullback-Leibler (KL) divergence between their outputs. This approach is particularly intuitive when these predictions take the form of probability distributions, such as an agent's action probability in reinforcement learning (RL). As such, KD techniques have been integrated into RL in unique ways for various purposes, such as unifying task-specific policies for parameter efficiency in [15], creating a bootstrap curriculum over agents for improved training and performance in Mix&Match [16], or MADT [17] which enables a single-agent offline-RL algorithm to function effectively in multi-agent spaces.

III. METHODS

To produce parameter-efficient policies for AVs, we adopt KD without a frozen teacher into the training loop of on-policy MARL algorithms.

A. Knowledge Distillation in on-policy Multi-Agent Reinforcement Learning

Advances in KD strategies for Transformers in sequence-to-sequence tasks across multiple domains have created a viable

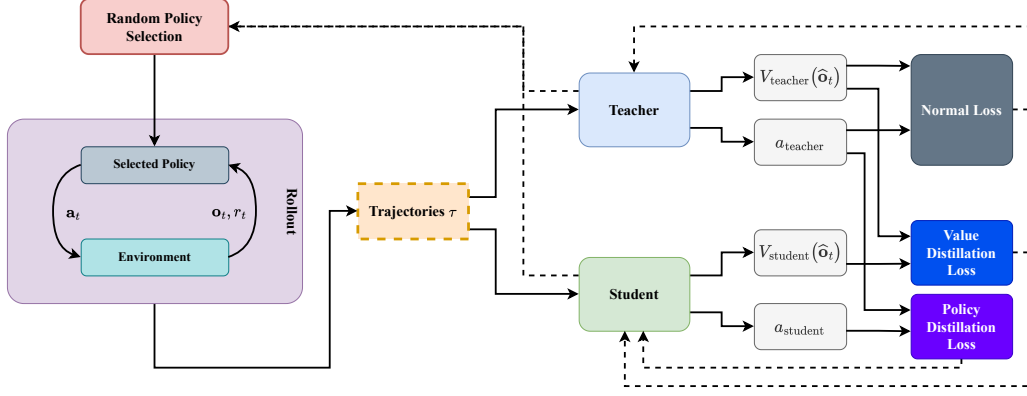


Fig. 2: **The overall framework of the proposed Online-Distilled Multi-Agent Reinforcement Learning paradigm.** During training, the algorithm randomly selects either the teacher or the student model to perform rollouts. The collected trajectories are used to train both the teacher and the student, while the teacher simultaneously imparts its knowledge to the student.

path for adaptation into MARL, despite domain-mismatch nuances. Denoting ‘stu’ for ‘student’ and ‘tea’ for ‘teacher’, Distilled Multi-Agent Reinforcement Learning (D-MARL) has the following value function loss:

$$\mathcal{L}_V^{\text{D-MARL}}(\phi_{\text{stu}}) = \mathcal{L}_V^{\text{MARL}}(\phi_{\text{stu}}) + \beta_V \frac{1}{Tn} \sum_{m=1}^n \sum_{t=0}^{t-1} \mathcal{L}(V_{\phi_{\text{stu}}}(\hat{o}_t^{i_m}), V_{\phi_{\text{tea}}}(\hat{o}_t^{i_m})) \quad (3)$$

where V is the value function approximator parameterized by ϕ , $\mathcal{L}_V^{\text{MARL}}$ is any MARL value function loss approximating the empirical Bellman error, $\mathcal{L}(\cdot, \cdot)$ is any regression loss function (such as Huber or Mean-Squared Error) and β_V is the value-function distillation coefficient. The D-MARL policy loss naturally extends the original distillation loss [14]:

$$\mathcal{L}_\pi^{\text{D-MARL}}(\theta_{\text{stu}}) = \mathcal{L}_\pi^{\text{MARL}}(\theta_{\text{stu}}) + \beta_\pi \frac{1}{Tn} \sum_{m=1}^n \sum_{t=0}^{t-1} \mathcal{D}(\pi_{\theta_{\text{stu},t}}^{i_m}, \pi_{\theta_{\text{tea},t}}^{i_m}) \quad (4)$$

where $\pi_{\theta_{\text{stu/tea},t}}^{i_m}$ is the overloaded notation for $\pi_{\theta_{\text{stu/tea},t}}^{i_m}(a_t^{i_m} | \mathbf{o}_t^{i_{1:n}}, \mathbf{a}_t^{i_{1:m-1}})$ - the student and teacher’s agent i_m ’s policy’s chosen probability for action $a_t^{i_m}$ at timestep t given previous agents’ actions $\mathbf{a}_t^{i_{1:m-1}}$ and the current timestep’s observation $\mathbf{o}_t^{i_{1:n}}$. Here, $\mathcal{L}_\pi^{\text{MARL}}$ can be any MARL policy loss, $\mathcal{D}(\cdot, \cdot)$ is any differentiable distance function comparing two probability distributions (like the KL divergence) and β_π is a hyperparameter controlling the importance of the policy distillation. If the policy shares parameters with the value function approximator, this policy loss will also update the value function’s parameters.

B. Online-Distillation for on-policy Multi-Agent Reinforcement Learning

As the example in Fig. 1 demonstrates, in MARL, the pretrained teacher model, while stable in training, may lose to

Algorithm 1 Online-Distilled on-policy MARL

Input: teacher’s initial policy and value parameters $\langle \theta_{\text{tea},0}, \phi_{\text{tea},0} \rangle$, student’s initial policy and value parameters $\langle \theta_{\text{stu},0}, \phi_{\text{stu},0} \rangle$, normal MARL value and policy losses $\langle \mathcal{L}_\pi^{\text{MARL}}(\cdot), \mathcal{L}_V^{\text{MARL}}(\cdot) \rangle$, learning rate η , number of training iterations T , number of agents n .

for $t = 0, 2, \dots, T-1$ **do**

 // Rollout

 Randomly select rollout policy from teacher and student.

$$\pi_{\text{rollout}} \sim \text{Uniform}(\pi_{\theta_{\text{tea},t}}, \pi_{\theta_{\text{stu},t}})$$

 Run selected policy π_{rollout} in environment

 Collect trajectories $\tau = \langle \tau^1, \tau^2, \dots, \tau^n \rangle$.

 // Training

if importance sampling ratios are used **then**

$$\pi_{\theta_{\text{old}}} \leftarrow \pi_{\text{rollout}}$$

end if

 Update student’s parameter using D-MARL losses

$$\theta_{\text{stu},t+1} \leftarrow \theta_{\text{stu},t} - \eta \nabla_{\theta_{\text{stu},t}} \mathcal{L}_\pi^{\text{D-MARL}}(\theta_{\text{stu},t}) \quad \triangleright \text{eq. (4)}$$

$$\phi_{\text{stu},t+1} \leftarrow \phi_{\text{stu},t} - \eta \nabla_{\phi_{\text{stu},t}} \mathcal{L}_V^{\text{D-MARL}}(\phi_{\text{stu},t}) \quad \triangleright \text{eq. (3)}$$

 Update teacher’s parameters using normal MARL losses

$$\theta_{\text{tea},t+1} \leftarrow \theta_{\text{tea},t} - \eta \nabla_{\theta_{\text{tea},t}} \mathcal{L}_\pi^{\text{MARL}}(\theta_{\text{tea},t})$$

$$\phi_{\text{tea},t+1} \leftarrow \phi_{\text{tea},t} - \eta \nabla_{\phi_{\text{tea},t}} \mathcal{L}_V^{\text{MARL}}(\phi_{\text{tea},t})$$

end for

Output: Student’s trained policy parameters $\theta_{\text{stu},T}$

the performance of the volatile student. This is different from the usual KD settings, where it is assumed that the teacher performs better than the student.

From that intuition, inspired by MADT [17] as well as Momentum-based RL methods [18], [19], the teacher is setting to be trained together with the student. Specifically, in our framework, the algorithm chooses either the teacher or the student, randomly with equal probability, to perform each rollout. Afterwards, the student is trained following D-MARL’s losses

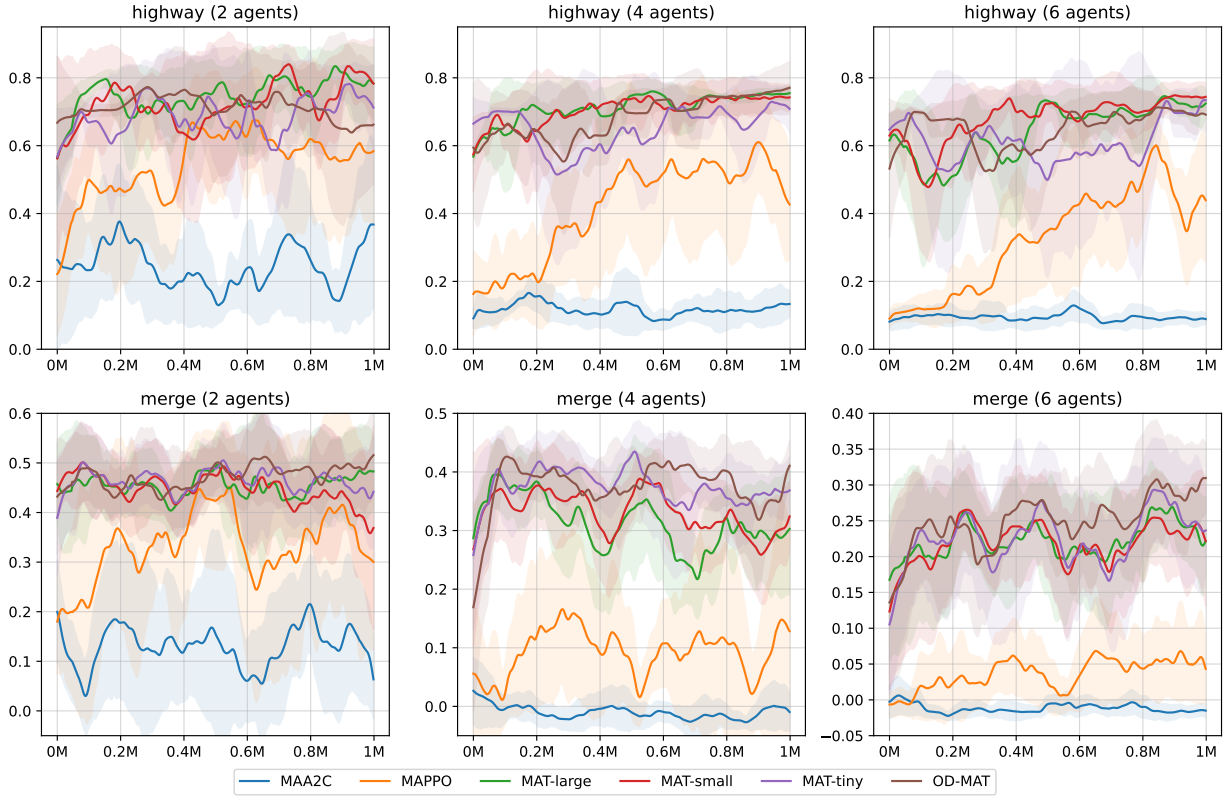


Fig. 3: Results on 6 scenarios averaged over 3 seeds. MAT-based methods achieve superior results over other baselines. OD-MAT combines the stability of MAT-large with the agility and small parameter count of MAT-tiny.

in Eq. 3 and 4, and the teacher with baseline MARL’s losses, $\mathcal{L}_V^{\text{MARL}}$, $\mathcal{L}_\pi^{\text{MARL}}$ using the collected trajectories. The framework of our proposed training modality is illustrated in Fig. 2, and summarized in Algorithm 1.

IV. EXPERIMENTS

A. Experiment Setup

We conduct evaluations in the multi-agent highway environment [20], using two scenarios: a 4-lane straight highway (‘highway’) and a 2-lane on-ramp merging setup (‘merge’). Illustrative screenshots of these scenarios are shown in Fig. 5. The roads are populated with both autonomous vehicles (AVs) and human-driven vehicles (HDVs), with HDV behavior simulated by the environment. To simulate different traffic densities, we vary the number of vehicles (for AVs and HDVs) across three settings: 2, 4, and 6.

Each AV agent, together with its own absolute position and velocity, observes up to an additional 4 vehicles within its surrounding area. For each observed vehicle, the agent perceives its longitudinal and lateral positions, along with its longitudinal and lateral velocities, all measured relative to the agent itself. The feature vector sequence is zero-padded when necessary. AVs then select from a discrete set of 5 actions: *turn left*, *turn right*, *cruise*, *accelerate*, and *decelerate*. The shared reward function imposes heavy penalties for collisions and

tailgating, while incentivizing high velocity. Reward values are normalized to have a maximum value of 1.

We compare our online-distillation method integrated into the Multi-Agent Transformer (OD-MAT) with the baseline MAT variant and other popular on-policy reinforcement learning algorithms: MAPPO [4] and MAA2C (an extension of the A2C [21] algorithm following the CTDE paradigm).

All methods uses $\gamma = 0.99$, learning rate $\eta = 5 \times 10^{-5}$, and trained for $T = 10^6$ steps, averaging rewards over 3 seeds. The PPO loss clip ratio is set to $\epsilon = 0.05$. MAT-based approaches all use a single attention head with standard Softmax attention as introduced in [12]. Otherwise, the actor and critic’s architectures are separate 4-layer fully-connected networks with hidden size 64 and Tanh activation. The critic concludes with a scalar node, while the actor outputs five Tanh-activated units. To contrast MAT’s lack of performance improvement despite an increasing parameter count, we train multiple models with varying Transformer configurations. Table I shows the layer and embedding dimension settings along with the number of parameters for each tested algorithm.

For the configurations of our proposed OD-MAT, the teacher in our model is pretrained for 10^6 steps and follows hyperparameters of ‘MAT-large’ throughout the pretrained process. During the distillation process, $\beta_{\text{Enc}} = \beta_{\text{Dec}} = 0.001$, $\mathcal{L}$ is the Huber loss and $\mathcal{D}$ is the KL divergence. Additionally, following the intuition that the teacher, with its significantly

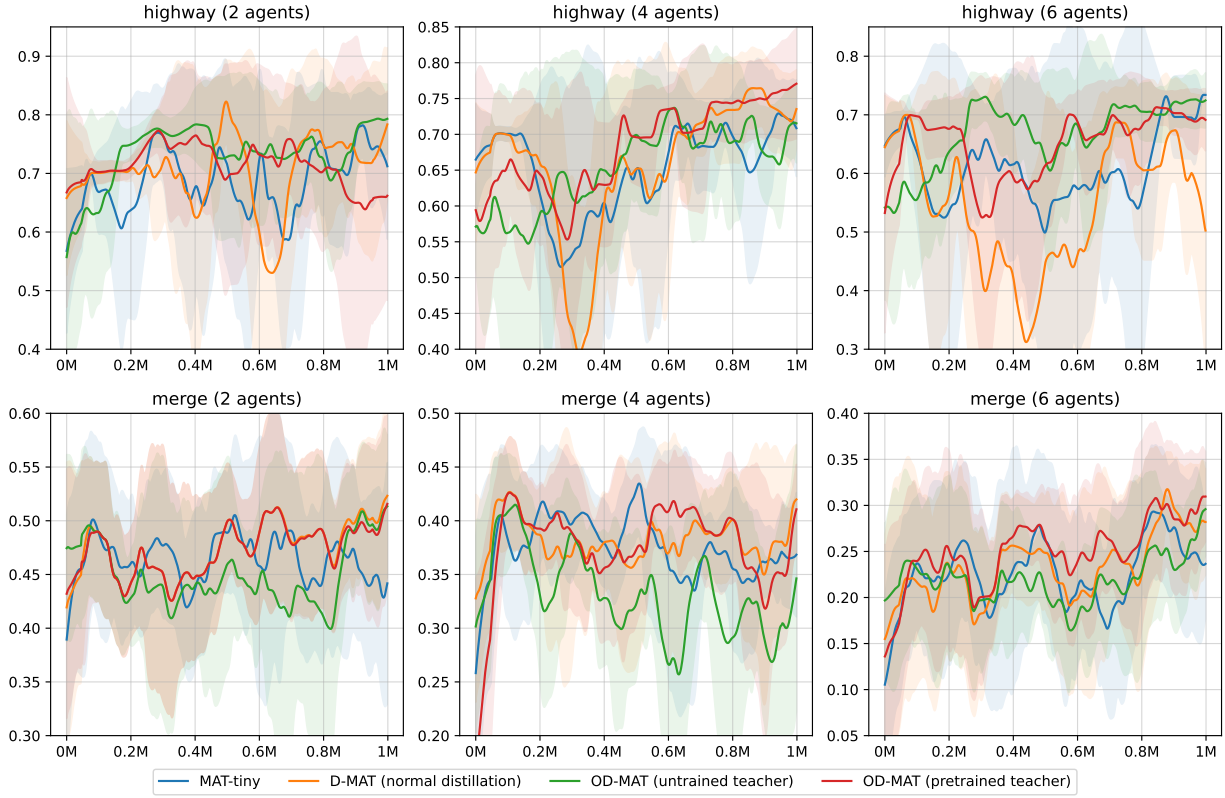
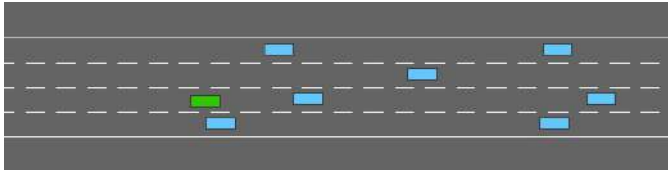
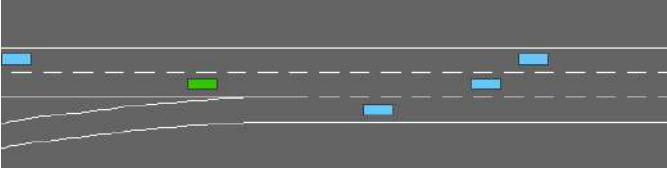


Fig. 4: Performance comparisons between different training modalities, all models have the exact same parameter count and achitecture. Online-Distillation combined with a pretrained teacher produces consistantly comparable or superior results.



(a) ‘highway’ scenario.



(b) ‘merge’ scenario.

Fig. 5: Screenshots of the ‘highway’ and ‘merge’ scenarios.

TABLE I: Algorithms’ model dimensions, their raw parameter counts and relative parameter count ratios.

Algorithm	d_{model}	n_{layer}	Param	Ratio
MAA2C	64	-	20,358	$\times 3.2$
MAPPO	64	-	20,358	$\times 3.2$
MAT-large	64	3	214,634	$\times 34.0$
MAT-small	64	1	80,234	$\times 12.7$
MAT-tiny	16	1	6,314	$\times 1$
OD-MAT	16	1	6,314	$\times 1$

higher capacity, can and wants to learn more from the collected experience, we set the teacher’s PPO clip hyperparameter to 0.1, twice that of the student. Otherwise, the teacher’s training hyperparameters follow those of all other baselines. OD-MAT’s student inherits the embedding dimensionalities, layer counts and training hyperparameters of MAT-tiny.

B. Results

Results are shown in Fig. 3. As illustrated, MAT-based methodologies significantly outperform other methods. The trade-off between performance and parameter efficiency inherent to MAT is also evident. In the easier ‘highway’ scenarios, MAT-tiny struggles to adapt and attain good performance; yet, for the harder ‘merge’ setting, MAT-tiny leverages its agility and successfully discovers policies that match or even surpass its larger-capacity counterparts. OD-MAT results in the lightest policies while simultaneously attaining stable and comparative performances on less complex ‘highway’ tasks, while surpassing traditional MAT methods on the challenging ‘merge’ benchmarks. The ‘online-distillation’ training paradigm allowed for the most parameter-efficient policies, yielding better results than those with 34 times more parameters.

We conduct an ablation study to reveal the importance of each element in the online-distillation training modality. Our method, and the normal MAT training loop, is benchmarked against the normal distillation method (D-MAT), where the

teacher is frozen and not used to collect samples throughout the student's training process, similar to traditional supervised distillation [14]. We also paired the *online-distillation* training paradigm with an *untrained teacher* to emphasize the importance of having a reliable teacher in ensuring meaningful distillation. Fig. 4 showcases results aligning with our intuition. Normal distillation succeeds in guiding the policy through advanced 'merge' environments but displays limited impact on variance reduction in simpler 'highway' scenarios. However, the effectiveness of online-distillation is also intertwined with the teacher's innate and initial competency. The student's learning, when piloted by a teacher with no real experience to offer, exhibits substantially erratic performance spikes, which are clearly evident despite averaging across seeds.

V. CONCLUSION

In this paper, we have proposed the Online-Distillation framework which enhances coordination and safety for AVs operating in complex mixed-traffic conditions. This framework can easily integrate student-teacher learning into any baseline MARL models such as the MAT architecture. Our framework is able to leverage of the knowledge from a large, stable, pretrained policy and the agile adaptability and exploratory capacity of a lightweight model. The approach was benchmarked under diverse traffic settings and densities, achieving higher raw performance in the most challenging scenarios, while using the fewest parameters. The experiment results of our proposed OD-MAT achieves smoother and more parameter-efficient highway control in numerous traffic densities.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-24-1-4012. We would like to thank LT Ha and MT Chau for helpful discussions.

REFERENCES

- [1] D. Chen, M. R. Hajidavalloo, Z. Li, K. Chen, Y. Wang, L. Jiang, and Y. Wang, "Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 11 623–11 638, 2023.
- [2] X. D. Ta, B. G. Le, T. H. Le, and V. C. Ta, "Resolve highway conflict in multi-autonomous vehicle controls with local state attention," in *2024 16th International Conference on Knowledge and System Engineering (KSE)*. IEEE, 2024, pp. 103–108.
- [3] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.
- [4] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 611–24 624, 2022.
- [5] M. Dhuheir, A. Erbad, B. Hamdaoui, S. Belhaouari, M. Guizani, and T. X. Vu, "Multi-agent meta reinforcement learning for reliable and low-latency distributed inference in resource-constrained uav swarms," *IEEE Access*, 2025.
- [6] M. Wen, J. Kuba, R. Lin, W. Zhang, Y. Wen, J. Wang, and Y. Yang, "Multi-agent reinforcement learning is a sequence modeling problem," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 509–16 521, 2022.

- [7] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*, ser. SpringerBriefs in Intelligent Systems. Springer International Publishing, 2016.
- [8] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning," 2017. [Online]. Available: <https://arxiv.org/abs/1706.05296>
- [9] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.
- [10] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2019. [Online]. Available: <https://arxiv.org/abs/1509.02971>
- [11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [13] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2018. [Online]. Available: <https://arxiv.org/abs/1506.02438>
- [14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: <https://arxiv.org/abs/1503.02531>
- [15] W. M. Czarnecki, R. Pascanu, S. Osindero, S. Jayakumar, G. Swirszcz, and M. Jaderberg, "Distilling policy distillation," in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 1331–1340.
- [16] W. Czarnecki, S. Jayakumar, M. Jaderberg, L. Hasenclever, Y. W. Teh, N. Heess, S. Osindero, and R. Pascanu, "Mix & match agent curricula for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1087–1095.
- [17] W.-C. Tseng, T.-H. J. Wang, Y.-C. Lin, and P. Isola, "Offline multi-agent reinforcement learning with knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 226–237, 2022.
- [18] F. Huang, S. Gao, J. Pei, and H. Huang, "Momentum-based policy gradient methods," in *International conference on machine learning*. PMLR, 2020, pp. 4422–4433.
- [19] J. Chen, J. Feng, W. Gao, and K. Wei, "Decentralized natural policy gradient with variance reduction for collaborative multi-agent reinforcement learning," *Journal of Machine Learning Research*, vol. 25, no. 172, pp. 1–49, 2024. [Online]. Available: <http://jmlr.org/papers/v25/22-1036.html>
- [20] E. Leurent, "An environment for autonomous driving decision-making," <https://github.com/eleurent/highway-env>, 2018.
- [21] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

EDGE++: Enhanced Direction Attention via Top-k Refinement for Long-Tail Recommendation

Trung-Hieu Nguyen, Minh-Khoi Le, Duc-Trong Le*

VNU University of Engineering and Technology, Hanoi, Vietnam

{20020136, khoilm, trongld}@vnu.edu.vn

Abstract—Recommender systems often suffer from popularity bias, where a few popular items dominate recommendations while long-tail items remain underrepresented. This bias reduces diversity, fairness, and overall user satisfaction. Recent post-processing methods, such as EDGE, refine item embeddings to improve long-tail recommendations without retraining. However, we identify a key inefficiency in EDGE: its softmax-based attention is highly concentrated on a few head items while still computing over thousands, leading to computational waste and noise. To address this, we propose EDGE++, a lightweight enhancement that introduces top-k refinement for embedding adjustment. By selecting only the most semantically relevant head items, EDGE++ reduces noise, improves directional precision, and lowers computational cost. Extensive experiments on three benchmark datasets, namely Gowalla, Yelp2018, and MovieLens-10M, demonstrate consistent improvements over state-of-the-art baselines. EDGE++ achieves superior performance on tail items while maintaining head item quality, offering a practical and effective solution in mitigating popularity bias in the recommendation task.

Index Terms—Recommender Systems, Popularity Bias, Long-Tail Recommendation, Post-Processing, Embedding Refinement

I. INTRODUCTION

Recommender systems play a central role in modern digital platforms by filtering massive amounts of information and personalizing user experiences. They power applications such as e-commerce, online media, and social networks, where effective recommendation directly influences user engagement and satisfaction. Despite their success, one of the most persistent and challenging issues in recommender systems is *popularity bias*. This bias occurs when models disproportionately recommend popular “head” items, while niche or less popular “tail” items receive little or no exposure [1], [2]. The consequences are significant: users are repeatedly exposed to similar content, recommendation diversity decreases, and creators of long-tail items struggle for fair visibility [3], [4].

The source of popularity bias lies in the nature of user interaction data. A small number of head items typically account for the majority of interactions, while tail items are underrepresented. When trained on such data, models naturally learn to favor head items, reinforcing a “rich-get-richer” cycle [5]. Over time, this feedback loop not only skews user experiences but also limits the discovery of new or personalized content. For platforms aiming to balance commercial goals with fairness and diversity, mitigating this bias is of both practical and ethical importance.

To mitigate popularity bias, prior work falls into two main categories: *in-processing* and *post-processing* strategies. In-processing methods intervene during training through re-weighting, regularization, causal inference, or graph-based adjustments [6]–[9]. While effective, they often require architectural changes and costly retraining, limiting their practicality. In contrast, post-processing methods modify outputs or embeddings after training, making them lightweight and easy to deploy. These include re-ranking [10] and embedding refinement such as TTEN [11], which normalizes embeddings to reduce head-item dominance. A recent post-processing approach, EDGE [12], refines item embeddings by leveraging head-item directions. It is model-agnostic and effective for tail items, but its softmax-based attention is highly concentrated. Over 95% of the weight often goes to just two or three head items, yet all head items are still processed. This causes computational waste and introduces noise, reducing refinement efficiency and precision.

In this work, we address these limitations by introducing **EDGE++**, an enhanced framework for embedding refinement in long-tail recommendation. It retains the model-agnostic and lightweight properties of EDGE but introduces a *top-k refinement strategy*. Instead of attending to all head items, EDGE++ selectively leverages only the most semantically relevant head items for each target. This focused refinement reduces noise, improves the precision of embedding directions, and significantly decreases computational cost. By design, EDGE++ provides a practical and scalable solution that can be seamlessly integrated into existing recommendation pipelines.

Our contributions are summarized as follows:

- *Attention Inefficiency Analysis*: We discover that EDGE’s softmax attention becomes extremely concentrated, with 95% weight on 2-3 head items while still computing over thousands of neighbors, indicating severe computational waste and noise injection.
- *Top-k Refinement*: We propose EDGE++, which enhances EDGE’s diffuse attention with top-k similar items for direction refinement, filtering noise and improving precision with significantly reduced computational cost.
- *Comprehensive Evaluation*: Extensive experiments demonstrate consistent improvements (up to 30.56% NDCG@20 gain) with superior tail item performance while maintaining head item quality, establishing new state-of-the-art for long-tail recommendation.

* Corresponding author.

II. RELATED WORK

This section reviews prior work on mitigating popularity bias in recommender systems, a challenge extensively documented in recent surveys [5], [13]. We first discuss the nature of the problem and then survey two main categories of solutions: in-processing and post-processing methods.

A. Mitigating Popularity Bias in Recommender Systems

Popularity bias is a well-documented phenomenon where recommender systems tend to over-recommend popular "head" items while under-recommending less popular "tail" items [1], [2]. This bias primarily stems from the highly skewed nature of user interaction data, where a few items capture the majority of interactions. Models trained on such data naturally learn to favor popular items, creating a feedback loop that further amplifies their visibility (the "rich-get-richer" effect) [5]. The consequences include reduced recommendation diversity and novelty, lower user satisfaction, and unfair exposure for niche content creators, particularly users with specific interests who are underserved when popular items overshadow personalized preferences [3], [4], [14], [15]. Recommending long-tail items is challenging due to data scarcity and the inherent trade-off between improving tail-item exposure and maintaining overall recommendation accuracy [3].

B. In-Processing Mitigation Strategies

In-processing (or training-phase) methods aim to address popularity bias by modifying the model's training process. Common strategies include:

a) *Re-weighting and Re-sampling*: These techniques adjust the learning process to give more importance to tail items. This can be done by assigning weights to training instances inversely proportional to item popularity or by oversampling interactions with tail items [6]. Wu et al. [7] found that using a sampled softmax loss can also help, as it frequently exposes the model to popular items as negative examples, thus preventing the model from blindly scoring them high.

b) *Regularization and Causal Inference*: Other approaches add regularization terms to the loss function to penalize over-focus on popular items [6]. More advanced methods use causal inference to disentangle an item's intrinsic appeal from its popularity-induced conformity. For instance, the DICE framework learns separate embeddings for "interest" and "conformity" to isolate the effect of popularity [8].

c) *Gradient-Based and Graph-Based Methods*: Some work focuses on the training dynamics, such as rescaling gradients to prevent popular items from dominating the learning process [9] or adapting embedding normalization [16]. In graph-based models like LightGCN [17], which are susceptible to bias since high-degree (popular) nodes dominate message passing, various techniques have been developed. These include adjusting message passing to ensure tail items receive sufficient information [9] and incorporating self-supervised contrastive learning to prevent representation collapse around popular items [18]–[20]. While effective, these methods often

require significant changes to the model architecture and training pipeline.

C. Post-Processing Mitigation Strategies

Post-processing methods offer a more flexible alternative by modifying a pre-trained model's outputs or representations without requiring retraining.

a) *Re-ranking and Score Adjustment*: The most straightforward approach is to re-rank the final recommendation list or adjust item scores based on their popularity. For example, one can penalize the scores of popular items or enforce a cap on their number in the top-K list [10].

b) *Embedding-Based Refinement*: A more sophisticated category of methods directly modifies the learned item embeddings. TTEN (Test-Time Embedding Normalization) [11] addresses bias by partially normalizing embedding norms during inference, thereby balancing between dot product (which favors high-norm popular items) and cosine similarity.

Building on this, EDGE (Embedding Enhancement for Long-Tail Recommendation) [12] proposes a two-step refinement process. It first adjusts the *magnitude* of item embeddings to better reflect their true popularity and then refines their *direction* by steering them towards semantically similar "mentor" head items. EDGE has been shown to significantly outperform prior methods like TTEN. While EDGE demonstrates effectiveness in tail recommendation, our analysis reveals a fundamental limitation: its attention mechanism computes over all head items but utilizes only a tiny fraction effectively. This creates both computational inefficiency and noise injection, motivating our focused top-k approach. Our work, EDGE++, identifies and addresses a key limitation within EDGE's direction refinement step, as detailed in Section III.

III. METHODOLOGY

In this section, we briefly review EDGE [12], a post-processing framework that serves as our primary baseline method. Next, we analyze the core limitation in EDGE's attention mechanism before proposing EDGE++, an enhanced version that utilizes a Top-k refinement for the long-tail recommendation task.

A. Preliminary: The EDGE Framework

EDGE [12] is a model-agnostic post-processing framework designed to improve the representational quality of item embeddings, especially for long-tail items. It operates after training a base recommender model, such as LightGCN, on a user-item interaction matrix $\mathbf{R} \in [0, 1]^{U \times I}$. In this setting, initial embeddings are learned for each user u denoted as $\mathbf{e}_u$ and item i denoted as $\mathbf{e}_i$. After the base model is trained, e.g., LightGCN, EDGE performs a two-stage refinement process to enhance the item embeddings:

a) *Magnitude Refinement*: To better reflect item popularity, the embedding norm is adjusted according to the formula:

$$\|\mathbf{e}_i\| \rightarrow (p_i + 1)^\beta \|\mathbf{e}_i\|^{1-\alpha} \quad (1)$$

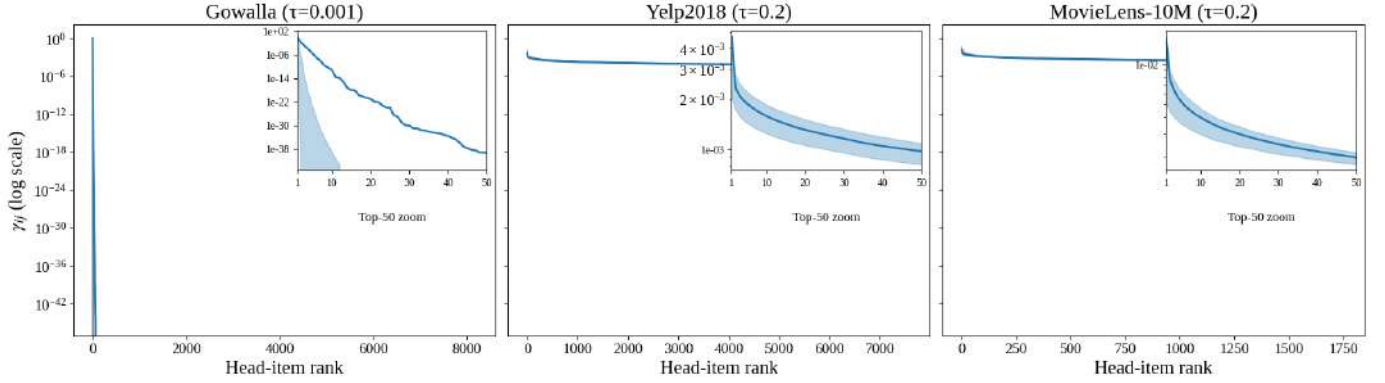


Fig. 1. Aggregate attention decay curves for randomly sampled items. The attention is highly concentrated on the top-ranked head items, while other items have negligible influence.

where p_i is the interaction count of item i , and $\alpha, \beta \in [0, 1]$ are hyper-parameters that balance the learned embedding magnitude with the explicit popularity signal.

b) Direction Refinement: This module refines an item's normalized direction $\mathbf{u}_i$ by combining it with a head-inferred direction $\mathbf{e}_i^H$. The head-inferred direction is computed as a weighted average of the directions of all head items $\mathcal{I}_H$ as:

$$\mathbf{e}_i^H = \sum_{j \in \mathcal{I}_H} \gamma_{ij} \frac{\mathbf{e}_j}{\|\mathbf{e}_j\|} \quad (2)$$

where the attention weights, γ_{ij} , are computed from cosine similarities using a softmax function with a temperature parameter $\tau \in \mathbb{R}^+$:

$$\gamma_{ij} = \frac{\exp(\cos(\mathbf{e}_i, \mathbf{e}_j)/\tau)}{\sum_{h \in \mathcal{I}_H} \exp(\cos(\mathbf{e}_i, \mathbf{e}_h)/\tau)} \quad (3)$$

The direction refinement of the embedding $\mathbf{e}_i$ is computed by combining the original and head-inferred vectors, controlled by a non-negative hyperparameter λ as:

$$\tilde{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|} + \lambda \frac{\mathbf{e}_i^H}{\|\mathbf{e}_i^H\|}, \quad (4)$$

c) Joint Refinement: The final item embedding is enhanced by jointly applying both magnitude and direction refinement as follows:

$$\mathbf{e}_i \rightarrow (p_i + 1)^\beta \cdot \|\mathbf{e}_i\|^{1-\alpha} \cdot \frac{\tilde{\mathbf{e}}_i}{\|\tilde{\mathbf{e}}_i\|} \quad (5)$$

The refined representation is later used to generate long-tail item recommendations.

B. EDGE++: Top-k Attention Refinement

a) Intuition: Despite its effectiveness in improving long-tail recommendations, EDGE's direction refinement mechanism has a critical limitation. Empirical analysis shows that under optimal hyperparameter settings—particularly with a low temperature value τ —the softmax attention weights γ_{ij} become highly concentrated on only a few top head items. As shown in Figure 1, most head items receive near-zero attention weights but are still involved in the computation.

This causes two main issues namely: i) *Computational waste*, due to unnecessary processing of the full head item set; ii) *Potential noise*, introduced by irrelevant head items, which may degrade the quality of the refined embedding direction.

b) Top-k Attention Refinement: To address these limitations, we propose **EDGE++**, a method that integrates a top-k selection mechanism into the direction refinement step. Instead of aggregating information from all head items, we only consider the k most similar ones.

To achieve this goal, we select the set of k head items, denoted as $\mathcal{I}_{\text{top-}k}$, that have the highest attention weights γ_{ij} for a given item i as:

$$\mathcal{I}_{\text{top-}k} = \underset{\mathcal{I}' \subset \mathcal{I}_H, |\mathcal{I}'|=k}{\operatorname{argmax}} \sum_{j \in \mathcal{I}'} \gamma_{ij} \quad (6)$$

Then, the direction refinement in Eq 7 is computed only on this subset $\mathcal{I}_{\text{top-}k}$ via:

$$\tilde{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|} + \lambda \frac{\mathbf{e}_i^{H(k)}}{\|\mathbf{e}_i^{H(k)}\|}, \quad (7)$$

$$\mathbf{e}_i^{H(k)} = \sum_{j \in \mathcal{I}_{\text{top-}k}} \gamma_{ij} \frac{\mathbf{e}_j}{\|\mathbf{e}_j\|} \quad (8)$$

This approach ensures that the refinement is both computationally efficient and robust to noise from irrelevant head items, leading to more precise embedding adjustments, especially for less-popular items.

c) Computational Complexity: **EDGE++** reduces the attention computation from $\mathcal{O}(|\mathcal{I}_H|)$ to $\mathcal{O}(k)$ where $k \ll |\mathcal{I}_H|$. For example, it may achieve up to $10\times$ speed-up when $k = 10$ and $|\mathcal{I}_H| = 1000$, while maintaining or improving accuracy.

IV. EXPERIMENTS AND ANALYSIS

We evaluate **EDGE++** through a detailed analysis guided by the following research questions:

- **RQ1:** Does **EDGE++** achieve higher overall recommendation accuracy compared to state-of-the-art baselines?
- **RQ2:** How effective is **EDGE++** in mitigating popularity bias, particularly for tail item recommendations?
- **RQ3:** How sensitive is **EDGE++** to key hyper-parameters?

TABLE I
DATASET STATISTICS.

Dataset	# Users	# Items	# Interactions	Density
Gowalla (GW)	29,858	40,981	822,465	0.00084
Yelp2018 (YL)	31,668	38,048	1,371,166	0.00130
MovieLens (ML)	69,166	8,790	4,916,999	0.00823

TABLE II
OVERALL PERFORMANCE COMPARISON IN TERMS OF RECALL@20 AND NDCG@20. THE BOLD VALUE DENOTES THE BEST PERFORMANCE WHILE THE UNDERLINE VALUE REPRESENTS THE SECOND-BEST ONE.

Dataset	Model	Recall@20	NDCG@20
Gowalla	LightGCN	0.2040	0.1314
	TTEN	0.2055	0.1330
	EDGE	<u>0.2160</u>	<u>0.1409</u>
	EDGE++	0.2176	0.1423
Yelp2018	LightGCN	0.1081	0.0614
	TTEN	0.1081	0.0614
	EDGE	<u>0.1129</u>	<u>0.0647</u>
	EDGE++	0.1156	0.0665
ML-10M	LightGCN	0.2654	0.1250
	TTEN	0.2731	0.1285
	EDGE	<u>0.3170</u>	<u>0.1613</u>
	EDGE++	0.3181	0.1632

A. Experimental Settings

1) *Datasets*: Following Lee et al. [12], we assess the proposed method on three widely used benchmark datasets: Gowalla (GW) [21], a location-based social network; Yelp2018 (YL)¹, containing user reviews of businesses; and MovieLens-10M (ML)², consisting of movie ratings. All datasets follow a long-tail distribution, which reflects real-world scenarios. Their statistics are presented in Table I.

2) *Evaluation Protocol*: Following the biased data protocol from [12], we split each dataset into training, validation, and test sets (70:10:20), preserving the natural long-tail distribution. We use two typical top-K ranking metrics: *Recall@20* and *Normalized Discounted Cumulative Gain (NDCG@20)*. All results are averaged over five runs with different random seeds for robustness.

3) *Baselines*: We compare EDGE++ against several strong baselines:

- **LightGCN** [17]: An advanced graph-based model that serves as our base model.
- **TTEN** [11]: A post-processing method that mitigates bias by partially normalizing item embedding norms at test time.
- **EDGE** [12]: The cutting-edge post-processing framework that EDGE++ is built upon. It refines both magnitude and direction using all head items.

4) *Implementation Details*: We extend the official EDGE framework. LightGCN is trained with Sampled Softmax loss,

¹<https://business.yelp.com/data/resources/open-dataset/>

²<https://grouplens.org/datasets/movielens/10m/>

TABLE III
THE NDCG@20 PERFORMANCE FOR HEAD AND TAIL ITEMS. THE BOLD VALUE DENOTES THE BEST PERFORMANCE WHILE THE UNDERLINE VALUE REPRESENTS THE SECOND-BEST ONE.

Dataset	Model	Head NDCG@20	Tail NDCG@20
Gowalla	LightGCN	0.2106	0.0850
	TTEN	0.2126	0.0875
	EDGE	<u>0.2203</u>	0.0976
	EDGE++	0.2236	<u>0.0970</u>
Yelp2018	LightGCN	0.0935	0.0455
	TTEN	0.0935	0.0455
	EDGE	<u>0.0974</u>	<u>0.0487</u>
	EDGE++	0.1002	0.0504
ML-10M	LightGCN	0.1371	0.0788
	TTEN	0.1425	0.0833
	EDGE	<u>0.1764</u>	<u>0.1116</u>
	EDGE++	0.1785	0.1120

an embedding size of 64, and the Adam optimizer. The key hyperparameters for EDGE++, τ and k , are tuned via grid search on the validation set to maximize Recall@20. Other parameters (α, β, λ) adopt the optimal values from the original EDGE paper [12] for a fair comparison. All experiments were run on an NVIDIA Tesla T4 GPU.

B. Main Results

1) *Performance Comparison to Baselines (RQ1)*: As shown in Table II, the EDGE++ model consistently establishes itself as the state-of-the-art solution for recommendation tasks across all three datasets. It systematically outperforms the other evaluated models including LightGCN, TTEN, and its direct predecessor, EDGE—in both Recall@20 and NDCG@20 metrics. This indicates that the enhancements introduced in EDGE++ provide a tangible and generalizable performance boost. On Gowalla, EDGE++ achieves the highest scores with a Recall@20 of 0.2176 and an NDCG@20 of 0.1423. A similar pattern of dominance is observed on the Yelp2018 dataset, where it scores 0.1156 and 0.0665 for Recall@20 and NDCG@20 respectively, once again leading the comparison. The model’s superiority is further solidified on the larger and denser ML-10M dataset, where it obtains its most impressive results: a Recall@20 of 0.3181 and an NDCG@20 of 0.1632. The consistent improvement of EDGE++ over the already strong EDGE model underscores the effectiveness of its refined architecture. The results consistently position EDGE++ as the top-performing model, demonstrating its robustness and superior accuracy in capturing user preferences across diverse data domains.

2) *Analysis of Popularity Bias Mitigation (RQ2)*: To analyze the impact on popularity bias, we evaluate NDCG@20 separately for head, i.e., top 20% popular items, and tail items, i.e., the remaining ones. Table III shows that EDGE++ consistently improves performance for head items. More importantly, for tail items, it achieves notable gains on Yelp2018 and ML-10M, highlighting its strength in mitigating popu-

larity bias. For example, on Yelp2018, EDGE++ improves Tail NDCG@20 by 3.5% over EDGE. These results confirm that focused refinement is particularly effective when base embeddings provide strong directional signals. Additionally, our analysis further reveals that embeddings with higher initial directional variance benefit more from the selective top- k refinement. This finding suggests that improving embedding quality during training can amplify the effectiveness of post-processing methods such as EDGE++.

C. Ablation Studies

1) *Impact of top- k* : We conduct an ablation study to analyze the sensitivity of EDGE++ to its key hyperparameters, k (number of neighbors) and τ (attention temperature). The optimal settings vary across datasets, correlating with data density. For the sparsest dataset, Gowalla, a small $k = 3$ is optimal. For the denser Yelp2018 and ML-10M datasets, larger values of $k = 30$ and $k = 110$ are preferred, respectively. This indicates that as data density increases, leveraging a broader set of relevant head items becomes more beneficial for direction refinement. Figure 2 illustrates this sensitivity on the Gowalla dataset (the figure is illustrative and may not use τ).

2) *Impact of τ* : The temperature τ controls the sharpness of the softmax attention: smaller τ produces a spikier distribution that concentrates weights on the most similar neighbors, whereas larger τ spreads attention more evenly. We tune τ on the validation split over the common grid $\{0.5, 0.6, 0.8, 1.0\}$ and select the value that maximizes Recall@20 (ties are broken in favor of a smaller τ). The resulting optima are dataset-dependent: Gowalla performs best with $\tau = 0.8$, while Yelp2018 and ML-10M both achieve their best results with $\tau = 0.5$. This pattern reflects a practical trade-off: under sparser conditions, a larger τ yields a gentler, more stable aggregation of directional signals; in denser regimes, a smaller τ maintains selectivity among highly similar candidates. Note that attention is computed only over the retrieved top- k neighbors (see Methodology), so τ primarily shapes the weight distribution within this subset without increasing computational overhead.

3) *Impact of refined embeddings*: Figure 3 reveals a strong positive correlation ($R^2 = 0.89$) between base embedding directional variance and EDGE++’s improvement over EDGE. This indicates that EDGE++ is most effective when base embeddings possess distinct directional signals, allowing our top- k selection to identify relevant mentor items for refinement. This insight explains performance variations across datasets: dense datasets like ML-10M naturally produce embeddings with higher directional variance, yielding greater benefits from focused refinement. Conversely, sparse datasets with limited directional diversity show modest gains. Practically, this suggests that improving base embedding quality during training can amplify post-processing effectiveness, and provides guidance for practitioners on when EDGE++ deployment will be most beneficial.

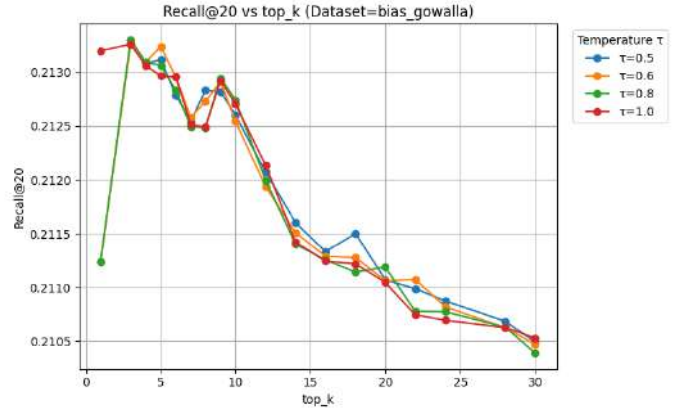


Fig. 2. Parameter sensitivity on Gowalla: Recall@20 versus k for various τ values in EDGE++.

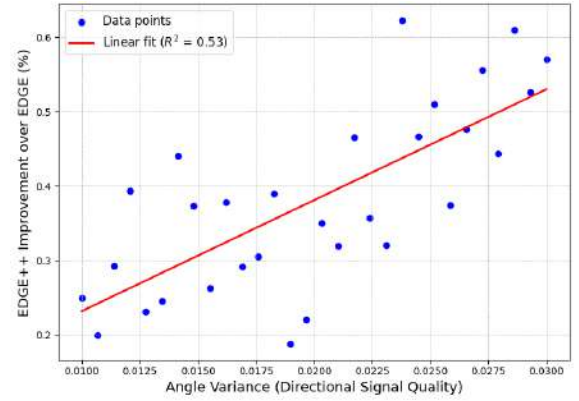


Fig. 3. Correlation between base embedding directional signal quality (angle variance) and the improvement provided by EDGE++ over EDGE on ML-10M dataset. A clear positive correlation ($R^2 = 0.89$) indicates that EDGE++ is particularly effective when base embeddings possess strong directional signals.

4) *Computational Efficiency*: As illustrated in Figure 4, our proposed method EDGE++ offers a substantial improvement in computational efficiency compared to the original EDGE baseline. This efficiency gain is achieved by strategically limiting the similarity computations to a small neighborhood of top- k items, rather than the entire set of head items. The computational efficiency gains are substantial: EDGE++ delivers speedups of $3.2\times$, $7.5\times$, and $7.4\times$ on Gowalla, Yelp2018, and ML-10M datasets, respectively. This considerable reduction in computational overhead not only accelerates the inference process but also enhances scalability and practicality for large-scale industrial applications.

V. CONCLUSION AND FUTURE WORK

This work demonstrates that popularity bias mitigation through post-processing can be significantly enhanced by principled attention design. Our proposed method, EDGE++, introduces a targeted top- k refinement strategy that simultaneously reduces computational overhead and substantially improves recommendation accuracy, especially benefiting underrepresented tail items. EDGE++ is practical and easy to deploy,

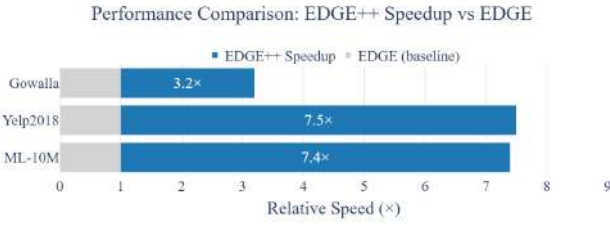


Fig. 4. Computational efficiency comparison of EDGE and EDGE++, with speedup highlighted.

requiring minimal implementation overhead, making it highly suitable for immediate integration into existing recommendation systems aiming to improve fairness and diversity. A key insight from our analysis is that the effectiveness of EDGE++ strongly correlates with the quality of the base embeddings’ directional signals. This finding provides valuable guidance for practitioners on when and how to most effectively deploy EDGE++, and suggests exciting future research directions to further enhance recommendation quality and fairness.

Limitations. As a post-hoc enhancement of EDGE, EDGE++ leverages head-item mentor directions and thus depends on the directional signal already encoded by the base model. When that signal is weak (e.g., high angular dispersion), improvements may diminish; when it is strong, the selective top- k attention amplifies it. This clarifies when EDGE++ will be most effective in practice and motivates strengthening base directionality during training.

Future Directions. To further amplify EDGE++ by strengthening base embeddings during training: (i) use self-supervised/contrastive objectives with graph/sequence augmentations; (ii) apply angle-aware regularization or angular-margin losses; (iii) adopt debiased negative sampling or re-weighting to curb head dominance; (iv) calibrate training selectivity via temperature design and τ -annealing; and (v) track lightweight directionality diagnostics (alignment–uniformity, angle variance) for early stopping. These are complementary to our post-hoc refinement and require no inference changes.

Collectively, these directions promise to broaden bias mitigation beyond conventional collaborative filtering towards emerging recommendation paradigms, ensuring improved user experience and fairness while preserving efficiency and practical deployability.

REFERENCES

- [1] Y.-J. Park and A. Tuzhilin, “The long tail of recommender systems and how to leverage it,” in *Proceedings of the 2008 ACM Conference on Recommender Systems*, ser. RecSys ’08. New York, NY, USA: Association for Computing Machinery, 2008, p. 11–18. [Online]. Available: <https://doi.org/10.1145/1454008.1454012>
- [2] H. Abdollahpour, M. Mansoury, R. Burke, and B. Mobasher, “The unfairness of popularity bias in recommendation,” *arXiv preprint arXiv:1907.13286*, 2019.
- [3] H. Steck, “Item popularity and recommendation accuracy,” in *Proceedings of the fifth ACM conference on Recommender systems*, 2011, pp. 125–132.
- [4] C. Musto, P. Lops, G. Semeraro *et al.*, “Fairness and popularity bias in recommender systems: an empirical evaluation,” in *DP@ AI* IA*, 2021, pp. 77–91.
- [5] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, “Bias and debias in recommender system: A survey and future directions,” *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–39, 2023.
- [6] H. Abdollahpour, R. Burke, and B. Mobasher, “Managing popularity bias in recommender systems with personalized re-ranking,” *arXiv preprint arXiv:1901.07555*, 2019.
- [7] J. Wu, X. Wang, X. Gao, J. Chen, H. Fu, and T. Qiu, “On the effectiveness of sampled softmax loss for item recommendation,” *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1–26, 2024.
- [8] Y. Zheng, C. Gao, X. Li, X. He, Y. Li, and D. Jin, “Disentangling user interest and conformity for recommendation with causal embedding,” in *Proceedings of the Web Conference 2021*, 2021, pp. 2980–2991.
- [9] W. Ren, L. Wang, K. Liu, R. Guo, L. E. Peng, and Y. Fu, “Mitigating popularity bias in recommendation with unbalanced interactions: A gradient perspective,” in *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022, pp. 438–447.
- [10] H. Abdollahpour, M. Mansoury, R. Burke, and B. Mobasher, “The connection between popularity bias, calibration, and fairness in recommendation,” in *Proceedings of the 14th ACM conference on recommender systems*, 2020, pp. 726–731.
- [11] D. Kim, J. Park, and D. Kim, “Test-Time Embedding Normalization for Alleviating Popularity Bias in Recommendation,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2023.
- [12] G. Lee, K. Kim, and K. Shin, “Post-Training Embedding Enhancement for Long-Tail Recommendation,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2024.
- [13] A. Klimashevskaya, D. Jannach, M. Elahi, and C. Trattner, “A survey on popularity bias in recommender systems,” *User Modeling and User-Adapted Interaction*, vol. 34, no. 5, pp. 1777–1834, 2024.
- [14] F. Carnovalini, A. Rodà, and G. A. Wiggins, “Popularity bias in recommender systems: The search for fairness in the long tail,” *Information*, vol. 16, no. 2, p. 151, 2025.
- [15] S. Channamsetty and M. D. Ekstrand, “Recommender response to diversity and popularity bias in user profiles,” in *FLAIRS*, 2017, pp. 657–660.
- [16] J. Chen, J. Wu, J. Wu, X. Cao, S. Zhou, and X. He, “Adap- τ : Adaptively modulating embedding magnitude for recommendation,” in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 1085–1096.
- [17] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, “Lightgcn: Simplifying and powering graph convolution network for recommendation,” in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 639–648.
- [18] Y. Zhang, H. Zhu, Y. Chen, D. Jin, Y. Li *et al.*, “Mitigating the Popularity Bias of Graph Collaborative Filtering: A Dimensional Collapse Perspective,” in *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.
- [19] L. Li, C. Zhang, S. Li, Y. Zhong, Q. Liu, and X. Zeng, “Graph contrastive learning with multi-objective for personalized product retrieval in taobao search,” *arXiv preprint arXiv:2307.04322*, 2023.
- [20] Y. Bai, L. Xiao, C. Xia, K. Zeng, and X. Shi, “Tca4rec: contrastive learning with popularity-aware asymmetric augmentation for robust sequential recommendation,” *Journal of Big Data*, vol. 12, no. 1, p. 137, 2025.
- [21] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: user movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1082–1090.

EmbedRank-LM: Natural Language-Driven Web Element Identification for Automated UI Testing

Trinh Le-Khanh, Quyen Hoang-Van and PHAM Ngoc Hung
VNU University of Engineering and Technology, Hanoi, Vietnam
{trinhlk, quyenhv, hungpn}@vnu.edu.vn

Abstract—User Interface (UI) testing is crucial for ensuring the functionality and user experience of software applications. Traditional manual UI testing approaches are time-consuming and error-prone, particularly as web applications scale. While automated testing has emerged as a solution, existing methods rely heavily on brittle element locators such as CSS selectors and XPath expressions, which are vulnerable to structural changes in web applications. This paper presents EmbedRank-LM, a novel approach for web element identification using natural language descriptions to reduce dependency on conventional element locators in automated UI testing. The proposed method employs a two-stage process: first, a sentence embedding model selects the top-k candidate elements with the highest semantic similarity to the input description; second, a large language model performs final element selection from the candidate set. An experimental evaluation on a dataset of 300 samples derived from Pasupat et al.’s benchmark demonstrates that EmbedRank-LM achieves 67% accuracy, significantly outperforming baseline methods, including Trinh et al. (42.67%) and AgentQL (59.67%). While EmbedRank-LM exhibits a longer execution time compared to baseline approaches, it demonstrates a substantially lower execution cost, being 18 times more cost-effective than AgentQL, and establishes an advantageous trade-off between accuracy, processing time, and computational cost for practical UI testing applications.

Index Terms—Element Locators, Large Language Model, Natural Language Processing, User Interface Testing, Web Element Identification

I. INTRODUCTION

The proliferation of web applications has fundamentally transformed how users interact with digital services, spanning diverse domains, including e-commerce, education, enterprise systems, and entertainment platforms [1]. As these applications become increasingly sophisticated and integral to daily operations, ensuring the quality and reliability of user interfaces (UIs) has emerged as a critical concern for software developers and quality assurance professionals [2]. UI Testing represents a cornerstone of software quality assurance, directly impacting both functional correctness and user experience satisfaction [3], [4].

Traditional manual testing approaches, while comprehensive in their coverage, exhibit inherent limitations that become increasingly pronounced as application complexity scales. Manual testing processes are characteristically labor-intensive, time-consuming, and susceptible to human error due to subjective interpretation and cognitive fatigue. These limitations are particularly acute in modern agile development environments

where rapid iteration cycles demand efficient and reliable testing methodologies.

The transition toward automated UI testing has addressed many efficiency concerns associated with manual approaches. However, automated testing faces technical challenges, particularly in the domain of web element identification and interaction. Contemporary automated testing frameworks rely heavily on element locators—structured expressions such as XPath, CSS selectors, and HTML attributes to precisely identify and interact with specific UI components within the DOM hierarchy. These locators serve as the fundamental mechanism through which automated tests perform actions such as clicking buttons, entering text, and verifying content.

Despite their widespread adoption, traditional locator-based approaches suffer from a critical vulnerability: *locator fragility*. Web applications undergo continuous evolution during development and maintenance cycles, resulting in frequent modifications to the underlying HTML structure. Such changes often invalidate existing locators, leading to test failures and necessitating extensive maintenance efforts. This fragility not only increases the total cost of test ownership but also reduces the overall reliability of automated testing suites.

A. Related Work

Several studies address the limitations of traditional element locators. Leotta et al. cite robula+ introduced ROBULA+, an algorithmic framework designed to generate robust XPath expressions that maintain stability across different versions of web applications, even when the underlying HTML structures undergo modification. While ROBULA+ demonstrates improved resilience compared to conventional XPath generation techniques, it remains dependent on the structural characteristics of HTML, which limits its effectiveness for applications with complex or frequently changing layouts.

Nass et al. [5] proposed Similo and its enhanced variant, VON Similo, both targeting element identification across updated application versions. Similo employs similarity scoring mechanisms to establish correspondences between elements across different versions, while VON Similo extends this foundation by incorporating visual overlap analysis between UI components. VON Similo achieved notable accuracy improvements, reaching 94.7% in element identification tasks compared to Similo’s 89% performance. However, both methodologies require access to historical application versions,

presenting practical limitations in scenarios involving new applications or when historical data is unavailable.

Kirinuki et al. [6] explored Natural Language Processing (NLP) integration with heuristic search strategies to enable element identification based on natural language descriptions. Their approach utilizes a Domain-Specific Language (DSL) to represent user actions, facilitating element identification without strict dependence on HTML structural characteristics. While this method demonstrates improved flexibility and maintainability, it requires users to adhere to specific DSL syntax constraints and supports only a limited action vocabulary. Additionally, the computational overhead associated with NLP processing and heuristic search may result in extended execution times for complex web applications.

More recently, Trinh et al. [7] developed a hybrid approach combining string matching with heuristic search algorithms to identify elements from natural language descriptions. This method achieves high accuracy rates, but requires at least one word that exactly matches either visible text or element attribute values. This constraint limits the method's capability to handle descriptions based on visual properties such as color, position, or contextual relationships. Furthermore, the reliance on exact string matching impedes the recognition of synonymous or semantically equivalent expressions.

AgentQL [8] represents a commercial solution that enables natural language-based element identification, eliminating the need for traditional XPath or CSS selectors. However, its closed-source nature limits the potential for customization and may present cost barriers for organizations with constrained budgets or specialized requirements.

B. Contributions

This research introduces EmbedRank-LM, a novel method combining sentence embedding models with Large Language Models (LLMs) for robust web element identification from natural language descriptions. EmbedRank-LM effectively handles diverse natural language descriptions, including synonymous expressions and contextually equivalent terms, without requiring exact string matches, thereby significantly improving robustness across varied linguistic inputs. Our approach reduces dependence on fragile HTML structural characteristics while maintaining identification accuracy, addressing fundamental limitations of traditional locator-based methods, and improving test maintainability.

Experimental assessment demonstrates superior performance compared to existing methodologies across dimensions of accuracy, processing efficiency, and cost-effectiveness.

The remainder of this paper is structured as follows. Section II presents a detailed exposition of the proposed EmbedRank-LM methodology, including architectural design and implementation considerations. Section III describes the experimental framework employed to evaluate method effectiveness, including comparative analysis with existing approaches. Section IV discusses current limitations and potential directions for future research enhancement. Finally,

Section V summarizes the principal contributions and outlines implications for future work in automated UI testing.

II. METHODOLOGY

This section presents EmbedRank-LM, a novel method for identifying web elements using natural language descriptions by integrating sentence embedding models and large language models. Our proposed method addresses the limitations of the locator fragility by leveraging semantic understanding and contextual reasoning capabilities.

The methodology implements a four-stage processing pipeline, as illustrated in Fig. 1. Each stage serves a specific purpose in progressively refining the search space and improving identification accuracy. The overall workflow comprises four main stages: (i) filtering non-meaningful elements, (ii) extracting semantic representations, (iii) selecting top-k candidate elements, and (iv) identifying the final target element.

A. Filtering Non-meaningful Elements

Modern web applications contain extensive HTML structures with numerous elements that serve different purposes. Processing complete HTML documents without preprocessing would significantly increase computational overhead and introduce noise, thereby reducing the accuracy of target element identification. The filtering stage addresses this challenge by implementing systematic criteria to identify meaningful elements for user interaction.

The filtering process applies two primary criteria: (i) ignore CSS properties that render them invisible; (ii) the element's tag belongs to commonly interactive categories (`<a>`, `<button>`, `<input>`, `<select>`, `<textarea>`, `<label>`, `<option>`, `<div>`, `<span>`) or includes an `onclick` attribute; or (iii) the element contains visible or displayable text or possesses at least one of the following attributes: `href`, `src`, `alt`, `title`, `value`, `aria-label`, or `placeholder`. This filtering approach ensures that only visible, meaningful, and naturally describable elements proceed to the identification pipeline, thereby enhancing both computational efficiency and identification accuracy.

B. Extracting Element Information

Following web element identification, our method extracts information to construct semantic representations for effective comparison with natural language descriptions. The extraction process captures visual, content-related, and semantic aspects as demonstrated in Fig. 2.

The system collects HTML attributes as key-value pairs, including `href`, `src`, `alt`, `title`, `aria-label`, `placeholder`, and custom `data-*` attributes, which provide essential functionality and content information. Visual features are extracted selectively, retaining only `width`, `height`, `top`, and `left` properties to minimize noise while preserving functionally meaningful characteristics. Text content undergoes normalization through the removal of whitespace to facilitate subsequent processing. Additional features include HTML tag names, class lists, and complete raw HTML snippets to maintain structural context.

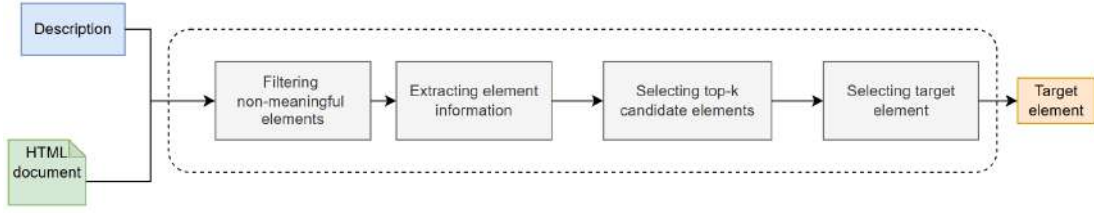


Fig. 1. Overview of the EmbedRank-LM method.

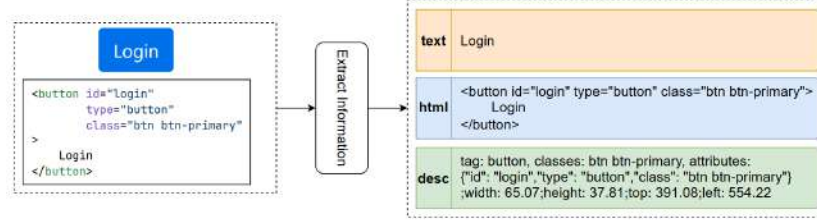


Fig. 2. An example of a web element and the extracted information.

Based on the extracted information, a descriptive sentence is generated by concatenating key attributes to provide an overview of the element’s characteristics.

The final semantic representation comprises three components: visible text content, raw HTML snippet, and generated descriptive sentence. These elements are aggregated into a unified structure representing each webpage element for downstream processing.

C. Selecting Top-k Candidate Elements

After obtaining semantic representations, EmbedRank-LM identifies the top- k most likely to match user intent for subsequent processing. Fig. 3 illustrates the similarity computation process. Each semantic representation component is encoded into an embedding vector using a sentence embedding model. These vectors are subsequently compared with the input description embedding using cosine similarity. The three resulting similarity scores are combined into a final score using Equation 1:

$$elm_{sim} = w_{txt} \cdot sim_{txt} + w_{desc} \cdot sim_{desc} + w_{html} \cdot sim_{html} \quad (1)$$

where sim_{txt} , sim_{desc} , and sim_{html} represent cosine similarity scores between the input description embedding and the embeddings of the element’s text, description, and raw HTML snippet, respectively. The weights w_{txt} , w_{desc} , and w_{html} are assigned values of 0.5, 0.2, and 0.3, respectively, based on empirical evaluation. After calculating similarity scores for all elements, the methodology sorts them in descending order of similarity and selects the top- k candidates for final processing.

Sentence embedding models typically undergo pretraining on general natural language data. The EmbedRank-LM methodology utilizes these models to encode visible text elements, generated descriptions, and raw HTML snippets. Without domain-specific fine-tuning, these models may not achieve optimal performance for web element identification tasks. To enhance performance, the methodology utilizes the

dataset published by Pasupat et al. [9] for fine-tuning purposes. This publicly available dataset¹ contains 51,664 examples distributed across training, development, and test subsets. Each HTML file includes uniquely identified elements through data-xid attributes.

Following the removal of automatic duplicates and invalid samples, the fine-tuning dataset contains 17,160 examples. The model undergoes fine-tuning using Multiple Negatives Ranking Loss (MNRL) [10], which proves suitable in the absence of explicit negative samples.

D. Selecting target element

The final stage identifies the single most appropriate element from the top- k candidates. This stage corresponds to the *Rerank* phase in *Retrieve-and-Rerank* [11] techniques. While traditional approaches employ Cross-Encoder models, the EmbedRank-LM utilizes large language models (LLMs) to leverage their powerful reasoning and contextual understanding capabilities. The LLM approach enables comprehensive element evaluation based on provided descriptions, displayed text, and contextual information represented through raw HTML code. Given the input description and list of k candidate elements, the system constructs a clearly structured prompt as illustrated in Listing 1. The prompt consists of two components: the system prompt specifies the model’s role as an assistant for selecting the best match for a query from a list of elements, and the user prompt provides the input description and candidate elements in JSON array format.

Listing 1. Prompt structure for *Selecting target element* step

```
//System prompt
You are an assistant that helps to select the best match for
a query from a list of elements.
//User prompt
Given these elements:
[ {
  "index": index, "description": "element description",
```

¹<https://nlp.stanford.edu/projects/phrasenode/>

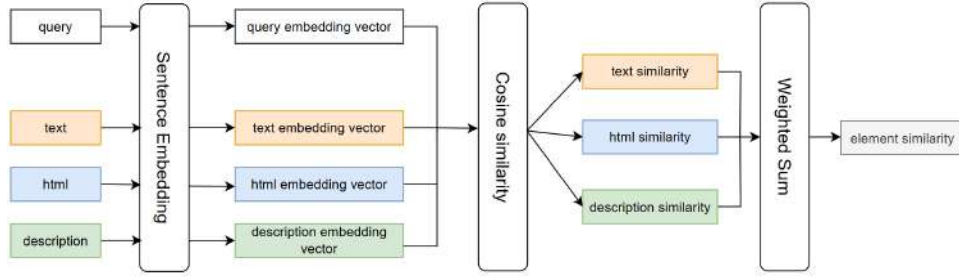


Fig. 3. The similarity between the input description and a web element computation process.

```

"text": "element text", "html": "element rawHTML",
}, ...]
Select the best match for the: "input query". If no element
match the query return the index -1.
Important: Only output a JSON object, do not include any
explanations, commentary, or surrounding text.
The response format must be a JSON object with fields:
{ "index": index, "reason": "REASON_STRING" }

```

Our approach combines the efficiency of semantic similarity computation with the reasoning capabilities of LLMs, providing a solution for natural language-driven web element identification in automated UI testing scenarios.

III. EXPERIMENTS AND RESULTS

This section presents an experimental evaluation of the proposed EmbedRank-LM method. The evaluation encompasses four distinct experiments designed to assess the method's effectiveness, efficiency, and practical applicability. First, we compare various sentence embedding models to identify the optimal choice for the candidate element selection stage. Second, we evaluate multiple LLMs to determine the most suitable model for final element selection. Third, we assess Cross Encoder models as potential alternatives to LLMs. Finally, we benchmark EmbedRank-LM against existing state-of-the-art methods to demonstrate its superiority.

A. Evaluation of Sentence Embedding Models

The evaluation was conducted on Google Colab² with the following specifications: Intel Xeon CPU @ 2.20GHz (12 vCPUs), 53GB RAM, and NVIDIA L4 GPU with 22.5GB VRAM. The experimental dataset was derived from Pasupat et al. [9], utilizing the first 500 samples from the original 10,252-sample test set. Each sample comprises a natural language description paired with web elements identified by unique *data-xid* attributes. Table I presents the comparative performance of state-of-the-art sentence embedding models (selected from the Hugging Face³). Model performance was assessed using *Recall@k* (with $k = 1, 3, 5, 10$)—the proportion of samples where the target element appears within the top- k retrieved elements; *MRR* (*Mean Reciprocal Rank*)—The average reciprocal rank of the correct element in the similarity-ranked list; and *Average Time*—the average computation time per sample in seconds.

²<https://colab.google.com/>

³<https://huggingface.co>

The *all-roberta-large-v1* model achieved superior performance across all **Recall@k** metrics, recording values of 0.566, 0.792, 0.82, and 0.852 for $k = 1, 3, 5$, and 10, respectively, with an MRR of 0.606. However, this performance comes at the cost of increased computational overhead, requiring 9.3 seconds per sample due to its 355 million parameters. The substantial performance gap between **Recall@1** and higher- k values (minimum gap of 0.19) indicates that expanding the candidate set significantly improves target element coverage. Conversely, the modest differences among **Recall@3**, **Recall@5**, and **Recall@10** (0.02-0.05) suggest diminishing returns beyond $k = 3$. Based on these findings, *all-roberta-large-v1* was selected for the candidate element selection stage.

B. Evaluation of Large Language Models

Due to computational constraints associated with local LLM deployment, experiments were conducted using the OpenRouter⁴ platform, which provides API-based access to diverse large language models. The evaluation dataset comprised samples where the target element appeared within the top- k candidates selected by *all-roberta-large-v1*, resulting in 396, 410, and 426 samples for $k = 3, 5$, and 10, respectively. Table II presents the performance comparison across different k values by the *Accuracy*, *Improved Accuracy* (the percentage of samples where the LLM correctly identified the target element despite it not being ranked first by the sentence embedding model), *Latency*, and *Cost*.

The *gemini-2.5-flash* model demonstrated superior performance, achieving accuracies of 88.38%, 83.90%, and 84.27% for $k = 3, 5$, and 10, respectively. It also recorded the highest improved accuracy values, indicating exceptional reasoning capabilities beyond initial similarity rankings. The model's latency of approximately 1.9 seconds remains acceptable for real-time applications. The improved accuracy metric, ranging from 5.81% to 24.41% across all models, demonstrates the significant value added by incorporating LLMs in the final selection stage. The *gemini-2.5-flash* model's maximum improvement of 22.68% ($k = 5$) underscores the substantial enhancement over similarity-based approaches alone. Considering accuracy as the primary criterion for the final selection stage, *gemini-2.5-flash* emerged as the optimal choice, deliver-

⁴<https://openrouter.ai/>

TABLE I
COMPARISON RESULTS OF SENTENCE EMBEDDING MODELS

Model	Recall@1	Recall@3	Recall@5	Recall@10	MRR	Time (s)
<i>all-mpnet-base-v2</i>	0.568	0.784	0.810	0.840	0.590	7.462
<i>all-roberta-large-v1</i>	0.566	0.792	0.820	0.852	0.606	9.302
<i>LaBSE</i>	0.536	0.778	0.810	0.846	0.590	3.175
<i>paraphrase-multilingual-MiniLM-L12-v2</i>	0.508	0.752	0.788	0.828	0.559	1.188
<i>use-cmlm-multilingual</i>	0.568	0.772	0.812	0.838	0.590	3.192

TABLE II
COMPARISON OF LARGE LANGUAGE MODELS WITH DIFFERENT VALUES OF k

Model	Accuracy (%)			Improved Accuracy (%)			Time (s)			Cost (USD)		
	k=3	k=5	k=10	k=3	k=5	k=10	k=3	k=5	k=10	k=3	k=5	k=10
gemini-2.5-flash	88.4	83.9	84.3	22.0	22.7	24.4	1.8	1.9	2.0	6.6	11.3	29.4
gemma-3-4b	56.6	51.5	44.8	9.9	10.0	8.0	1.4	1.5	1.6	0.6	1.2	3.2
llama-3.1-8b	59.3	58.5	51.6	5.3	7.1	9.2	3.6	3.1	5.0	0.8	1.3	2.8
phi-4-multimodal	42.7	54.1	44.8	5.8	9.8	8.2	6.1	1.3	1.7	1.8	2.6	6.8

ing consistent high performance while maintaining reasonable latency and cost characteristics.

C. Cross Encoder Model Evaluation

This experiment compared Cross Encoder models against the best-performing LLM (*gemini-2.5-flash*) using identical datasets and evaluation conditions. Three Cross Encoder models were evaluated: *ms-marco-TinyBERT-L2-v2*, *ms-marco-MiniLM-L6-v2*, and *ms-marco-MiniLM-L12-v2*. These models, pre-trained on the MS MARCO dataset [12], range from 4.39 million to 109 million parameters. The models are evaluated based on Accuracy, Improved Accuracy (The proportion of samples where the target element is not ranked first by the sentence embedding model but is correctly selected by the Cross Encoder), and Time.

Table III presents the Cross Encoder model performance comparison. The *ms-marco-MiniLM-L6-v2* model achieved the highest accuracy among Cross Encoders (72.22% for $k = 3$), while *ms-marco-TinyBERT-L2-v2* provided the fastest inference (0.02 seconds for $k = 3$). However, even the best-performing Cross Encoder significantly underperformed compared to *gemini-2.5-flash* (88.38% for $k = 3$), justifying the selection of LLMs for the final selection stage.

D. Comparative Analysis with Existing Methods

The final experiment compared EmbedRank-LM with two baseline methods: Trinh [7] and AgentQL [8]. The evaluation employed 300 samples from Pasupat’s dataset [9], selected based on compatibility with the structured input requirements of the Trinh et al. method. Performance was evaluated using accuracy, processing time, and execution cost metrics.

Table IV presents the comparative performance of all evaluated methods. EmbedRank-LM with $k = 10$ achieved the highest accuracy (67.67%), representing improvements of 7.33% over AgentQL and 24.33% over Trinh et al. The $k = 5$

configuration provides an optimal balance, achieving 67.00% accuracy while maintaining cost efficiency at 0.53 USD per 1,000 samples—approximately 18 times more cost-effective than AgentQL. While EmbedRank-LM exhibits higher processing time than baseline methods, this overhead is justified by its superior accuracy and cost-effectiveness. The method’s ability to accurately interpret natural language descriptions for web element identification, combined with its economic efficiency, positions it as a practical solution for large-scale automated UI testing applications.

The experimental evaluation demonstrates that EmbedRank-LM significantly outperforms existing methods in accuracy while maintaining cost-effectiveness. The optimal configuration employs *all-roberta-large-v1* for candidate selection and *gemini-2.5-flash* for final element selection with $k = 5$. This configuration achieves 67% accuracy, representing a substantial improvement over baseline methods while providing an economical solution for practical deployment in automated UI testing scenarios. **Note** that the cost in Table II only includes the LLM API (using the OpenRouter platform), while the cost in Table IV is the integration cost of the entire process, normalised to 1,000 samples.

IV. THREATS TO VALIDITY

Several factors may limit the validity and generalizability of our experimental results. The effectiveness of EmbedRank-LM depends critically on the top- k candidate filtering stage, as the LLM cannot recover target elements excluded from the initial candidate set, creating a fundamental bottleneck that may affect system reliability. Our experimental scope presents additional limitations. The evaluation was conducted exclusively on English-language queries and desktop web pages using Pasupat et al.’s benchmark dataset, leaving performance on multilingual inputs and mobile interfaces unvalidated.

TABLE III
COMPARISON OF CROSS ENCODER MODELS WITH DIFFERENT VALUES OF k

Model	Accuracy (%)			Improved Accuracy (%)			Time (s)		
	k=3	k=5	k=10	k=3	k=5	k=10	k=3	k=5	k=10
TinyBERT-L2-v2	70.0	62.2	58.5	0.0	0.0	0.1	9.8	9.3	9.6
MiniLM-L6-v2	72.2	64.4	60.6	0.1	0.2	0.4	12.1	10.2	10.1
MiniLM-L12-v2	70.5	63.4	60.1	0.2	0.3	0.8	12.4	11.5	11.5

TABLE IV
COMPARATIVE PERFORMANCE OF EMBEDRANK-LM AND BASELINE METHODS

Method	Accuracy (%)	Time (s)	Cost (USD)
Trinh et al.	42.67	19.47	0.00
AgentQL	59.67	12.24	9.9
Our method $k = 3$	63.00	23.70	0.23
Our method $k = 5$	<i>67.00</i>	<i>23.46</i>	<i>0.53</i>
Our method $k = 10$	67.67	25.17	1.66

The current implementation processes only static DOM structures and cannot handle dynamic elements generated through user interactions or JavaScript execution, limiting effectiveness in modern interactive web applications. Additionally, although cost-effective compared to existing methods, the 355-million-parameter embedding model introduces considerable processing time, which may restrict its real-time applicability. Finally, like many machine learning approaches, EmbedRank-LM exhibits limited interpretability in element selection, which may hinder its adoption in production environments that require explainable decisions for maintenance and debugging.

V. CONCLUSION

This research introduced EmbedRank-LM, a novel two-stage methodology designed for web element identification using natural language descriptions, specifically addressing the critical fragility associated with traditional element locators in automated UI testing. The approach utilizes a two-stage process: a fine-tuned sentence embedding model is employed for initial candidate filtering, followed by a Large Language Model (LLM) to perform the final, precise selection. EmbedRank-LM demonstrates superior performance, achieving an accuracy of 67% on a dataset of 300 samples, significantly outperforming baseline methods such as Trinh et al. (42.67%) and AgentQL (59.67%). Furthermore, the method achieves an optimal balance, being approximately 18 times more cost-effective than AgentQL while maintaining superior accuracy, thereby establishing an advantageous trade-off for practical UI testing applications.

Based on these promising results, future research will focus on enhancing the existing framework and expanding its generalizability. A primary limitation lies in the initial candidate filtering stage, as the LLM cannot recover target elements excluded early on. Future work, therefore, focuses on developing learning-based ranking models to enhance this filtering stage

and minimize the initial bottleneck. Furthermore, research will expand the method's applicability by developing capabilities to handle dynamic elements generated through user interactions or JavaScript execution, as the current implementation is limited to static DOM structures. Finally, to improve scalability and real-time applicability, computational optimization techniques such as model distillation and quantization would be implemented, particularly given the considerable processing time introduced by the *all-roberta-large-v1*. Efforts to enhance explainability mechanisms and address the current limitation of single-element selection are also vital for practical adoption in production environments.

REFERENCES

- [1] S. Sotnik, T. Shakurova, and V. Lyashenko, "Development features web-applications," 2023.
- [2] A. Abdul-Aziz, N. A. Emran, N. Abdullah, and M. S. Sulong, "The need for web development quality to develop high-quality web applications," *International Journal of Research and Innovation in Social Science*, vol. 8, no. 12, pp. 4427–4434, 2024.
- [3] N. Junior, H. Costa, L. Karita, I. Machado, and L. Soares, "Experiences and practices in gui functional testing: A software practitioners' view," in *Proceedings of the XXXV Brazilian Symposium on Software Engineering*, 2021, pp. 195–204.
- [4] J. S. Alqurni, "Evaluating the user interface and usability approaches for e-learning systems," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 18, no. 1, pp. 1–25, 2023.
- [5] M. Nass, E. Alégroth, R. Feldt, and R. Coppola, "Robust web element identification for evolving applications by considering visual overlaps," in *2023 IEEE Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 2023, pp. 258–268.
- [6] H. Kirinuki, S. Matsumoto, Y. Higo, and S. Kusumoto, "Nlp-assisted web element identification toward script-free testing," in *2021 IEEE International Conference on Software Maintenance and Evolution (IC-SME)*. IEEE, 2021, pp. 639–643.
- [7] Le-Khanh Trinh, Bui-The Cong, Pham-Hoang An, Hoang-Van Quyen, Cao-Thi-Minh Tam, Ngoc Hung PHAM, "A novel method for high-fidelity web element identification," Unpublished manuscript, 2024, accepted at KSE 2024.
- [8] AgentQL, "Make the web ai-ready," <https://www.agentql.com/>, 2024, accessed: 2025-30-05.
- [9] P. Pasupat, T.-S. Jiang, E. Z. Liu, K. Guu, and P. Liang, "Mapping natural language commands to web elements," 2018. [Online]. Available: <https://arxiv.org/abs/1808.09132>
- [10] N. Reimers and I. Gurevych, "Multiple negatives ranking loss - sentence transformers documentation," accessed: 2025-05-08. [Online]. Available: https://www.sbert.net/docs/package_reference/sentence_transformer/losses.html#multiplenegativesrankingloss
- [11] —, "Retrieve & re-rank," https://sbnet.net/examples/sentence_transformer/applications/retrieve_rerank/README.html, accessed: 2025-06-05.
- [12] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human-generated machine reading comprehension dataset," *arXiv preprint arXiv:1611.09268*, 2016.

Enhancing Personalized Learning With Real-Time Data Warehousing and Linked Open Data Integration For Open Knowledge Domain

1st Pham Ngoc Thuy Trang

*Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
trang.pham@jaist.ac.jp*

3rd Mau Nguyen

*Data Solutions Dept.
FPT AI and Data Integration Center, FPT Software
Tokyo, Japan
maund@fpt.com*

2th Xianwen Zheng

*Center for Innovative Distance Education and Research
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
zheng-x@jaist.ac.jp*

4rd Shinobu Hasegawa*

*Center for Innovative Distance Education and Research
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
hasegawa@jaist.ac.jp*

Abstract—Modern education is undergoing a dramatic transformation in the era of Artificial Intelligence (AI) and the Industrial Revolution 4.0. This shift is pushing education systems toward more flexible, data-driven, and learner-centric approaches. In this context, personalized learning has emerged to address individual differences and overcome the limitations of the traditional ‘one size fits all’ teaching approach, with numerous research studies exploring knowledge management and the challenges associated with this educational paradigm. However, current work has not yet resolved the issue of data, particularly the difficulty of continuously capturing learners’ experiences in real-time within open-domain learning environments - unconstrained educational settings where learners freely access diverse resources and engage in various activities without predetermined boundaries. This challenge is compounded by the unrestricted nature of such learning contexts. To address this challenge, this paper proposes a novel architecture that dynamically integrates learners’ data stored in a real-time data warehouse (DWH) with Linked Open Data (LOD). This integration aims to enhance personalized learning systems in an open-domain context and develop an educational temporal concept map. Our goal is to combine the DWH with a knowledge graph to create a robust foundation for personalized learning. This unified framework offers valuable resources for learners while generating rich data for researchers in the field of smart education. The code is available at: <https://github.com/PikarryPham/PL-OpenDomain>.

Index Terms—Personalized Learning, Open Domain Learning, Dynamic Knowledge Graph, Knowledge Integration, Data Warehouse, Knowledge Management

I. INTRODUCTION & RELATED WORK

Education has evolved into a digital, data-driven paradigm where learner-centered personalization addresses traditional teaching limitations, adapting learning paths to improve outcomes, motivation, autonomy, and equity [1]. In addition to technologies, data play an important role in this context, and it is now fueling sentiment analysis, performance prediction, learning analytics, and individualized assessment. This makes effective knowledge management and seamless integration of

technology and data decisive for next-generation personalized learning systems. Knowledge integration in personalized learning extends beyond merely offering content and requires the ability to dynamically adapt to the context and needs of learners in real-time [2]. A truly open-domain learning system can engage with and learn from various topics, transcending fixed domains or specific platforms. This system must assimilate and connect knowledge from multiple sources. Previous research often relies on synthetic/simulated datasets that lack realism and complexity, leading to poor real-world performance and potential biases [3], [4]. Models trained on such data become inefficient, unreliable, and inflexible, complicating research comparisons and negatively impacting under-resourced communities. Additionally, personalized learning research is constrained by single-platform data usage, causing “high course dependence” and limiting broader behavior analysis [5]. Datasets from ASSISTment and Junyi Academy [6] focus solely on mathematics, whereas other studies utilize self-collected data from specific courses [7] or Massive Open Online Course (MOOC) platforms [4], which limits broader pattern identification. With the emergence of Large Language Models (LLMs) and Big Data, there is a critical need for systems capable of real-time retrieval across diverse domains, moving beyond fixed platforms. In the educational realm, Linked Open Data (LOD) has emerged as a mechanism for sharing across different systems, consequently supporting knowledge source integration and providing access to the latest information [8]. For this reason, we utilize DBpedia as a freely accessible LOD hub [9]. Furthermore, current approaches using traditional graph representations cannot effectively process real-time streaming data, capturing only static snapshots and thus hindering dynamic, personalized learning experiences [7]. Despite advances, current studies connecting input data with ontology knowledge graphs still face significant limitations resulting in domain-specific solutions [10]. Therefore, we propose combining traditional methods for integrating streaming learning data with LOD

*Corresponding Author

by strategically narrowing search spaces according to automatically defined learner topics, using recursive queries for related concepts, and generating embeddings to compare with browsing data. Additionally, traditional data warehouses struggle with processing unstructured data [11], whereas Data Lakes face persistent challenges including data staleness and lack of robust management features [12]. In response, we propose adopting Data Lakehouse architecture, which effectively combines low-cost storage with management capabilities for real-time access [11]. Having recognized static knowledge graphs' inherent limitations in capturing dynamic data [5], [7], we thoughtfully employ Educational Temporal Knowledge Graphs (ETKG), incorporating temporal data as timestamps, hence transforming triples into quadruples to comprehensively account for data evolution over time [13].

II. METHODOLOGY

Our approach consists of three core components corresponding to three stages as follows. In Fig. 1, the red flow illustrates the data processing pipeline associated with the learner's learning data, while the purple flow outlines the corresponding processing pipeline of DBpedia.

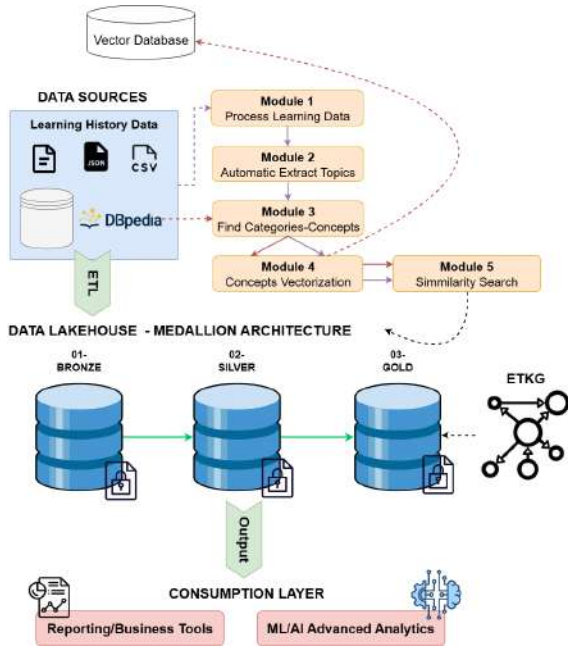


Fig. 1. An overview of the proposed system architecture

A. Learning Data Collection

To address data-related challenges, such as synthetic data, limited platform, and specific-domain issues, we developed a real-time web extension that collects and analyzes real user learning behaviors based on their browsing history within an open-domain learning context. This approach allows learners to study on any platform and at their convenience. Specifically, we develop tracking and streaming data processing algorithms using a streaming data platform architecture RedPanda [14]

and a column store distributed database ClickHouse [15]. This setup is optimized for real-time analysis at a petabyte scale, ensuring high performance and low latency during data collection due to their wire-protocol compatibility, balanced resource usage, and the ability to achieve operational simplicity at scale [15], [16]. The system also extracts web page content and metadata that track page transition events, automatically storing relevant information without user interaction. Despite its simplicity, this schema effectively supports scalable and efficient streaming learners' learning data collection. The database diagram depicts tables related to users, questions, options, and browsing learning history. The relationships between tables are primarily one-to-many, helping to link data about users' learning and browsing behavior.

Our approach utilizes the extension not only as a tool for gathering learning behavior data from learners but also as a means to enhance user experience. The system architecture implements a sophisticated data flow mechanism designed to capture and process learner information and learning history in real-time. User interaction data follows two distinct pathways based on its nature: non-streaming data is transmitted directly from the server to ClickHouse for persistent storage. In contrast, streaming data generated during web browsing is first pushed from the server to RedPanda, a streaming event processor. After event processing and data ingestion, RedPanda forwards the processed streaming data to ClickHouse for long-term storage. This bifurcated approach optimizes data handling efficiency. Additionally, the system facilitates bidirectional communication, allowing the "Extension" to request user history from the server, which retrieves the required historical data from the ClickHouse.

B. ETL Processing On Data Lakehouse

In addition to developing a web extension for the data collection phase, to support real-time personalized learning, we implement a modern DWH based on the Medallion architecture, integrated with LOD for semantic enrichment and deeper analysis. Data is processed on cloud infrastructure using Airbyte, Airflow, Polars and Azure Data Lake for secure, scalable storage of structured and unstructured data. The ETL process follows a three-layered Medallion design [17]. The Bronze Layer stores raw batch and streaming data while preserving its original structure and lineage using Change Data Capture (CDC) for historical tracking and reprocessing. The Silver Layer transforms and cleans data, applying ACID-compliant open table formats, such as Parquet, to enable schema evolution and time travel. This layer also integrates external knowledge from DBpedia for enriched analytics. In the Gold Layer, data is denormalized, combined, organized by domain, and optimized for consumption via Business Intelligence (BI) tools and AI/Machine Learning (ML) workflows. As shown in Fig. 2, the entire ETL process in layers will be optimized to handle large amounts of data efficiently within a short timeframe. The same happens with the other two layers.

In more detail, each layer in the Data Warehouse (DWH) is organized into two main task groups: batch sources and

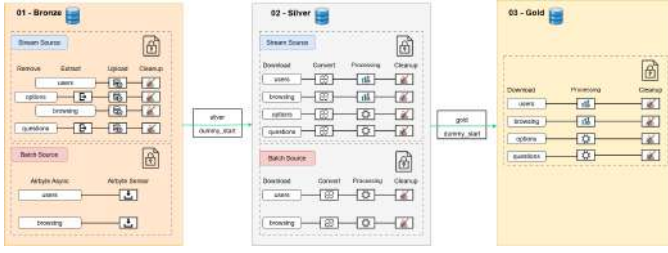


Fig. 2. DWH ETL flow: batch and stream pipelines

stream sources. Each task group is further divided into smaller tasks, which correspond to individual tables in the data source, allowing for easier management. Security is ensured through fine-grained authentication, authorization, and Role-Based Access Control (RBAC), enabling safe data sharing and enhancing the DWH's scalability. Data within the Data Warehouse (DWH) will be organized in a hierarchical structure to facilitate efficient storage, management, and processing. This structure will also support data backup at various levels. The ETL pipeline uses Fernet symmetric encryption with Azure Key Vault for secure data processing. During Bronze-to-Silver transformation, sensitive User table fields (username, email) are encrypted before pre-processing. At the Gold layer, selective decryption creates two data versions: encrypted for general access and decrypted for authorized personnel only. This encryption strategy ensures data privacy compliance while maintaining operational flexibility through role-based access controls.

C. Learning Data With Knowledge From LOD Integration And ETKG Construction

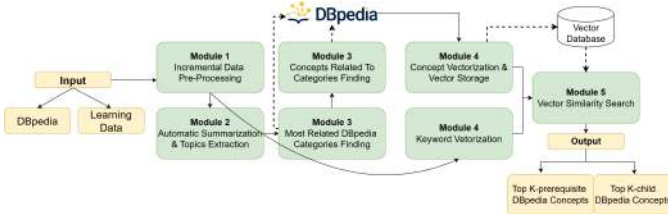


Fig. 3. DBpedia-learner activity integration workflow

1) *Content Summarization & Topic Extraction*: Module 1 is designed to pre-process the learner's historical data collected. It merges page visits into sessions with auxiliary metrics for enhanced behavioral analytics. Then, Module 2 utilizes DistilBERT-CNN [18] to generate lightweight topics on demand, thereby avoiding the costly limitations of DBpedia embedding. DistilBERT-CNN summariser provides bidirectional attention for faithful abstraction while delivering 300ms inference for real-time personalisation with reduced computational costs [19]. Content summarization involves two phases: model initialization with DistilBERT-CNN optimization through quantization for efficiency [20], and processing that includes document detection, pre-processing, length-adjusted parameters, quality evaluation, and post-processing to

deliver concise, representative summaries. Important features include adaptation to various document types, multi-layered redundancy systems ensuring continuous output, and post-processing steps that remove repetition and verify similarity with the original text, delivering concise and representative summaries. Our system utilizes summarization results to automatically extract the most relevant topics from the content of each entry in a learner's study history. Rather than relying on a single method, we propose a hybrid approach that combines the strengths of three distinct techniques: Zero-shot classification with DistilBERT, Latent Dirichlet Allocation (LDA), and NLP-based keyword analysis. LDA compensates for DistilBERT's requirement of predefined topics by discovering latent topics without labeled training data [21], while keyword extraction addresses LDA's weaknesses with single-topic texts and sparse documents [22]. The implementation involves three steps: weighting and combining multi-method results, filtering by minimum thresholds, and providing meta-data with method attribution. Additional techniques enhance quality through embeddings and domain mapping, yielding top-5 scored topics.

2) *From Vectorization, Similarity Searching to ETKG Construction*: The remaining modules focus on vectorizing text, storing the results in a vector database, and comparing the similarities between vectors to find information related to "Topics", "Categories," and "Related Concepts" tied to each entry in the learner's learning history data. More specifically, the system leverages the "Topics" identified in module 2 to discover relevant "Categories" and "Related Concepts" found in DBpedia. By intelligently determining topics in advance, the process enhances information retrieval within a vast knowledge base, conserving space, time, resources, and costs. This approach also enhances the flexibility and efficiency of information retrieval, resulting in improved classification accuracy and reduced processing errors. Topics refer to broad central concepts, "Categories" are detailed subgroups organizing information under "Topics", and "Related Concepts" represent specific ideas within categories, forming a many-to-many relationship. This interconnected structure creates potentially infinite depth when retrieving information from DBpedia, requiring optimization strategies.

For vectorization, we enhance the PageRank algorithm to optimize search processes while allowing the specification of maximum DBpedia traversal levels. We experiment and utilize various embedding models to find the most suitable embedding models, including BM25 and BERT models such as DistilBERT, XLM-RoBERTa, and RoBERTa, along with hybrid methods between BM25, TF-IDF with a traditional BERT model to vectorize Related Concept information based on abstracts and comments, fine-tuning it on learning history data to improve extraction accuracy before quantization. For similarity searching, we store vectors in Qdrant, selected for its high search performance that enhances information recognition accuracy and supports complex applications like semantic search in short timeframes [23]. Moreover, we also employ the relational database MySQL to store recorded topic,

category, and child page information using indexing, retrieval, and augmentation techniques. Learning history data is vectorized by processing keywords into the `exact_keyword` field using the same embedding model. Finally, in the Module 5, we compare learning history vectors with DBpedia vectors using Qdrant’s similarity search, which searches by stored vectors without embedding model inference to identify knowledge most relevant to learners. Fig. 3 illustrates the workflow of Stage 2, spanning Modules 1 through 5.

The outcome of the knowledge integration process is a list of entry IDs, with each entry linked to relevant parent and child concepts along with descriptive information, forming a dynamic knowledge network that reflects users’ learning behavior in real-time. We implement ETKG stored in Neo4j Graph Database for processing rather than using simple graph manipulation libraries. Our ETKG consists of seven primary node types (User (purple color), Question (pink color), Option (brown color), Entry (orange color), and Concept Nodes, including Topic (yellow color), Category (red color), and Related Concept (green color)) with various relationships such as User ”CHOOSE” Option or User ”HAS” Entry. The concept hierarchy is structured in three levels: Topics (Level 1, Root Nodes), Categories (Level 2), and Related Concepts (Level 3, Leaf Nodes). Each edge includes a timestamp indicating when relationships take effect, with temporal labels varying across different relationships. For example, User-Option timestamps indicate when learners select survey answers and User-Entry timestamps mark first access to learning pages. Additionally, entry-topic relationships contain `weightDurationPerSession` and `weightView` values, which measure session duration and page visit frequency, respectively. Higher values indicate topics that are more relevant to learners. The detailed graph representation is illustrated in Fig. 4, where each relationship is annotated with its respective properties in addition to a timestamp attribute. For conciseness, auxiliary node types present in the full schema—such as Question and Option—are omitted from the figure.

III. EVALUATION

The evaluation aims to assess the quality, performance, and flexibility of the developed Data Warehouses (DWH) in storing and processing large volumes of data in real-time. Additionally, it seeks to compare these methods with one another to identify the best approach for Linked Open Data (LOD) integration, as outlined in the proposed method. Our experimental process begins with a comprehensive dataset preparation, detailed extensively in Section 3.1. Most of the learning behavior data was collected from learners aged 18 to 34, including undergraduate and graduate students from various disciplines and backgrounds. These participants had either basic or advanced English reading comprehension skills and came from different countries around the world. Although we collected nearly 10,000 records in total, we implemented a validation methodology to ensure the accuracy of our integration results. This involved two reputable reviewers who

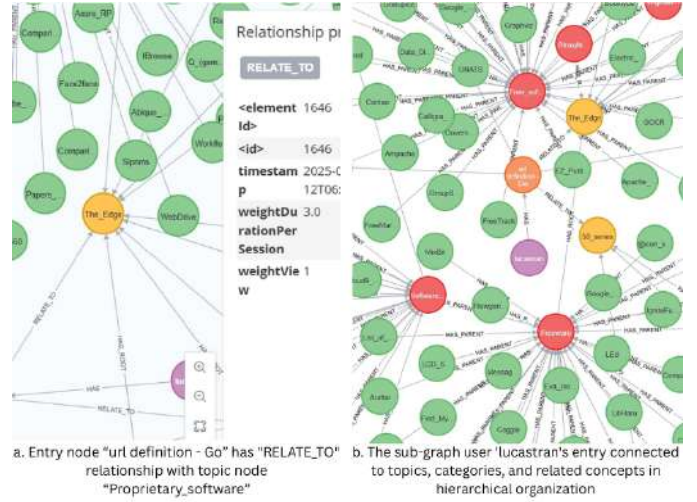


Fig. 4. Several relationship types supported in ETKG

independently annotated and assessed 200 randomly selected learning history records from diverse domains. The reviewers were two co-authors of this paper. In addition to evaluating the performance of DWH in processing and integrating large volumes of data from multiple sources, we also utilized the Junyi dataset, which contains a total of 25,925,992 student logs [24].

A. ETL Processing On Data Lakehouse

About configuration: To conduct our experiments, we deployed the DWH system on a cloud server optimized for handling large data volumes. The configuration features 24 virtual CPUs (vCPUs) for efficient multi-threaded processing, 96 GB of RAM for effective caching, and a 100 GB SSD that ensures high read and write speeds, as well as reduced I/O latency. The server runs on Ubuntu 22.04 LTS, providing stability and user-friendly management. This setup offers exceptional computing performance, robust parallel processing, flexible scalability, and straightforward upgrade options for CPU, RAM, or additional drives, ensuring high availability and reliability.

About data retrieval performance: The data retrieval performance involves multi-layered data collection and extraction through the Bronze layer, which extracts data from two sources: the Junyi dataset and Clickhouse. Streaming data from Clickhouse is retrieved hourly throughout the day, while batch data from the Junyi dataset is processed once daily. Despite an elevated average response time of 574.73 seconds due to outlier processes, the DWH demonstrates excellent operational efficiency with 66.7% of jobs completing within 0-2 seconds and a median execution time of just 1.20 seconds. The seven-day performance trend shows consistent improvement, with success rates—defined as the daily successful execution rate of ETL jobs—maintaining above 97.8% and response times stabilizing. This bimodal performance distribution indicates robust handling of both lightweight operations and complex data processing tasks, positioning the system as highly reliable

for production environments with strong availability guarantees. Fig. 5 provides an overview of the collected stream data, including the total page count of all websites for each timeframe on each day of the week and the top keywords searched by learners over the last seven days, etc.

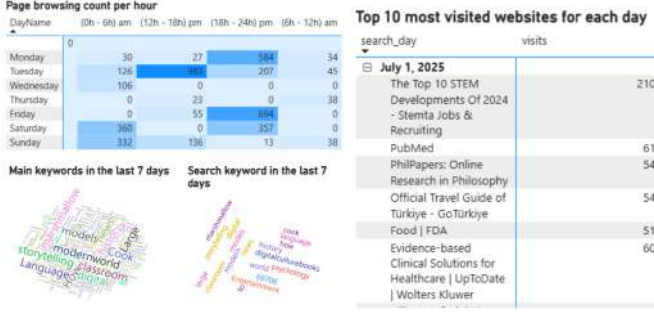


Fig. 5. Learner engagement metrics dashboard (7-day period)

About system extensibility: The DWH extensibility has achieved an overall score of 0.72/1.0. This score is calculated by averaging the modularity score, achieving a perfect score of 1.0, the configurability score at 0.67, and the schema flexibility score at 0.5. More precisely, while the Modularity Score measures task organization across six categories, validated by Ernst Pisch’s M-score metric [25], the Configurability Score evaluates configuration flexibility through configurable tables, queries, and connections, serving as an extensibility predictor. Besides, the Schema Flexibility Score measures schema adaptability through schema awareness and column flexibility, grounded in the HECATAEUS framework and VLDB schema evolution research [26]. This demonstrates that a balanced capability across architectural organization, operational adaptability, and data structure evolution allows for seamless extensions without the need for major restructuring. As a result, extensions can be made with minimal impact, while maintaining high task modularity across all operational categories. Schema evolution capabilities are robust, requiring only 30-45 minutes for fundamental schema changes and 0.5-1 hour for adding new schemas. The configuration-driven approach minimizes code modifications, with 75% of extension scenarios (including new data sources, schema evolution, and transformations) completing within 0.5 to 1 hour. This well-structured pipeline design facilitates rapid integration of new data sources and supports agile development practices, positioning the system as highly adaptable to evolving business requirements.

B. Learning Data With Knowledge From LOD Integration

The agreement between these reviewers served as a benchmark for assessing the accuracy of our integration approach. We utilize Overlap Coefficient (Szymkiewicz–Simpson) metrics to quantify the inter-reviewer agreement, as this measure is particularly suitable for arbitrary keyword extraction tasks where the potential keyword space is vast, unconstrained, and subject to variations between annotators. The statistics reveal strong similarity patterns between reviewers, with the Overlap

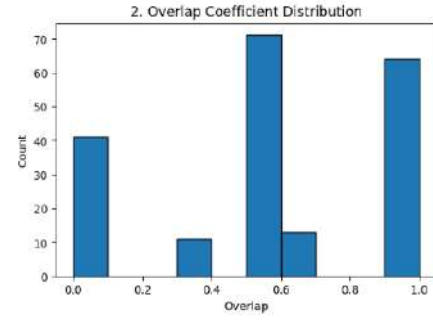


Fig. 6. Inter-reviewer concordance percentages for annotation outcomes

Coefficient Distribution exhibits a distinct bimodal pattern, with the majority of values concentrated at the two extremes: approximately 60% of cases show complete overlap (coefficient = 1.0), while about 40% show no overlap (coefficient = 0.0). The intermediate values (0.2, 0.4, 0.6, 0.8) demonstrate markedly lower frequencies, with each representing fewer than 15% of cases. These complementary distributions indicate substantial reviewer consensus, with overlap measures revealing excellent agreement despite more conservative Jaccard similarity scores. Moreover, to evaluate the accuracy of the proposed method for integrating knowledge with DBpedia, we use AvgBERTScore by calculating and measuring the cosine similarity index between the integration results of the system with two gold standards from two reviewers. BERTScore is an index created using the BERT (Bidirectional Encoder Representations from Transformers) model to assess the semantic similarity between words, sentences, or paragraphs. It employs cosine similarity between embeddings—vectors that represent the semantics of words—to measure the similarity between each keyword from the system and the corresponding ground-truth set.

AvgBERTScore is the average of the two BERTScore results calculated when comparing with the two gold standards from the two reviewers. Using AvgBERTScore enhances the reliability of the evaluation results, as comparisons with multiple benchmarks provide a more comprehensive view of quality. In this case, if BERTScore is equal to or greater than the threshold (here we take the number as 0.7), we consider the integration results between knowledge with DBpedia to have “matched” both in meaning and matched with the learner’s learning information. We choose a threshold of 0.7 because it is commonly used to distinguish similar pairs, has a good balance between coverage and accuracy, and has been verified in previous experiments [27]. The results show that the method using Module 2 with the embedding model XLM-RoBERTa has the highest AvgBERTScore. Other embedding models, such as RoBERTa or hybrid methods, also give equally good results. Using the same model but with Module 2 to summarize and extract topics from learning pages automatically, the experiment shows that the integrated results are always higher and more accurate. The result of the integration between historical learning data and knowledge sourced from DBpedia,

which yields the highest AvgBERTScore, has been employed to construct ETKG in Fig. 4.

In our data integration method with DBpedia, we identify three primary error types affecting result accuracy. First, ambiguity errors encompass issues such as synonyms or near-synonyms (e.g., "car" versus "automobile"), same concepts with different specificity levels (e.g., "vehicle" versus "car"), and morphological differences. Second, incomplete match errors occur when keywords are semantically different or refer to distinct topic scopes. Third, DBpedia errors include cases where correct keywords are not found in DBpedia, keywords require adjustment to match DBpedia's structure, or equivalent concepts do not exist in DBpedia despite existing conceptually. Embedding models like TF-IDF and BM25, which rely solely on term frequency and simple text statistics without capturing deeper semantic context, are more susceptible to these errors, resulting in lower performance than other models.

TABLE I
AVGBERTSCORE COMPARISON WITH/WITHOUT MODULE 2

Model	With	Without	Improvement (%)
TF-IDF	0.42	0.38	+10.5
BM25	0.55	0.44	+25.0
DistilBERT	0.63	0.56	+12.5
RoBERTa	0.72	0.63	+14.3
XLM-RoBERTa	0.84	<u>0.78</u>	+7.7
TF-IDF+BERT	0.71	0.64	+10.9
BM25+BERT	0.73	0.66	+10.6

IV. CONCLUSION

In this work, we propose a novel architecture that enhances personalized learning in open knowledge domains by integrating real-time learning history data, which is captured across multiple platforms in real-time and stored in DWH with LOD. Moreover, our ETKG marks an important innovation, enabling personalized systems that reflect the evolving nature of learner behavior and knowledge state over time. We address the main challenges of collecting, enriching learner data, and enabling powerful semantic query capabilities. The implementation and testing demonstrate the practicality and scalability of our approach, showing strong potential in supporting next-generation personalized education systems and data-driven educational research. Future directions include refining knowledge graph embedding techniques and integrating with intelligent personalized learning recommendation systems.

REFERENCES

- [1] O. H. L. Huang, Anna YQ and S. J. Yang, "Effects of artificial intelligence-enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom," *Computers and Education*, 2023.
- [2] S. Hu and X. Wang, "Foke: A personalized and explainable education framework integrating foundation models, knowledge graphs, and prompt engineering," *China National Conference on Big Data and Social Computing*, 2024.
- [3] A. Scarlatos and A. Lan, "Exploring knowledge tracing in tutor-student dialogues," *arXiv preprint arXiv:2409.16490*, 2024.
- [4] H. e. a. Zhu, "Personalized learning path recommendations: Fusing knowledge graph embedding, sequence mining, and collaborative filtering," *2024 IEEE International Conference on Big Data (BigData)*, 2024.
- [5] H. J. et al., "Kpcr: Knowledge graph enhanced personalized course recommendation," *AI 2021: Advances in Artificial Intelligence - 34th Australasian Joint Conference, AI 2021, Sydney, NSW, Australia, February 2-4, 2022, Proceedings*, 2022.
- [6] W. et al., "A comprehensive exploration of personalized learning in smart education: From student modeling to personalized recommendations," *arXiv preprint arXiv:2402.01666*, 2024.
- [7] Q. Wei and X. Yao, "Personalized recommendation of learning resources based on knowledge graph," *2022 11th International Conference on Educational and Information Technology (ICEIT)*, 2022.
- [8] S. B. Abbes, Ferial and A. Maalel, "Generative ai and gamification for personalized learning: Literature review and future challenges," *SN Computer Science*, 2024.
- [9] M. Farber, B. Ell, C. Menne, and A. Rettinger, "A comparative survey of dbpedia, freebase, openycyc, wikidata, and yago," *Semantic Web Journal*, 2015.
- [10] A. MATSUURA and Y. HAYASHI, "Facilitating thinking like a historian in open-ended learning space: A white box approach," in *International Conference on Computers in Education*, 2024.
- [11] S. et al., "The lakehouse: State of the art on concepts and technologies," *SN Computer Science*, p. 449, 2024.
- [12] V. K. Ravi and A. Ayyagari, "Data lake implementation in enterprise environments," *International Journal of Progressive Research in Engineering Management and Science (IJPREAMS)*, 2023.
- [13] Y. S. Liu, Yihe and Y. Dai, "Enhancing temporal knowledge graph representation with curriculum learning," *Electronics* 13, 2024.
- [14] S. Bhaskaran, "Resilient real-time data delivery for ai summarization in conversational platforms: Ensuring low latency, high availability, and disaster recovery," *Journal of Intelligent Connectivity and Emerging Technologies*, 2023.
- [15] S. et al., "Clickhouse-lightning fast analytics for everyone," *Proceedings of the VLDB Endowment*, 2024.
- [16] K. Jaspin, Y. S. S. JS, et al., "Real time network threat detection using machine learning and kafka in middleware," in *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, IEEE, 2024.
- [17] S. J. Shabu, M. Maheswari, J. Refonaa, et al., "Analysis of semiconductor chip performance using medallion architecture," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, 2024.
- [18] S. et al., "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [19] A. Venkataramana, K. Srividya, and R. Cristin, "Abstractive text summarization using bart," in *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, IEEE, 2022.
- [20] B. Rokh, A. Azarpeyvand, and A. Khanteymoori, "A comprehensive survey on model quantization for deep neural networks in image classification," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 6, pp. 1–50, 2023.
- [21] A. Kaur and J. R. Wallace, "Moving beyond lda: A comparison of unsupervised topic modelling techniques for qualitative data analysis of online communities," *arXiv preprint arXiv:2412.14486*, 2024.
- [22] T. Trinh, T. Quan, and T. Mai, "Nested variational autoencoder for topic modelling on microtexts with word vectors," *Expert Systems*, 2021.
- [23] P. N. Singh, S. Talasila, and S. V. Banakar, "Analyzing embedding models for embedding vectors in vector databases," in *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 2023.
- [24] R. Schmucker, J. Wang, S. Hu, and T. M. Mitchell, "Assessing the performance of online students—new data, new approaches, improved accuracy," *arXiv preprint arXiv:2109.01753*, 2021.
- [25] J. F. Roddick, "Schema evolution in database systems - an annotated bibliography," *SIGMOD record*, vol. 21, no. 4, pp. 35–40, 1992.
- [26] E. Pisch, Y. Cai, R. Kazman, J. Lefever, and H. Fang, "M-score: An empirically derived software modularity metric," in *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pp. 382–392, 2024.
- [27] T. J. Cann, B. Dennes, T. Coan, S. O'Neill, and H. T. Williams, "Using semantic similarity to measure the echo of strategic communications," *EPJ Data Science*, vol. 14, no. 1, p. 20, 2025.

Finetuning LLMs for Automatic Form Interaction on Web-Browser in Selenium Testing Framework

Nguyen-Khang Le¹, Hiep Nguyen¹, Minh Ngoc Nguyen¹, Son T. Luu^{1,3}, Trung Vo¹,
Quan Minh Bui², Shoshin Nomura², Le-Minh Nguyen¹

¹*Japan Advanced Institute of Science and Technology, Ishikawa, Japan*

²*Amifiable Inc., Tokyo, Japan*

³*VNU University of Information Technology, Ho Chi Minh City, Vietnam*

Abstract—Automated web application testing is a critical component of modern software development, with frameworks like Selenium widely adopted for validating functionality through browser automation. Among the essential aspects of such testing is the ability to interact with and validate web forms, a task that requires syntactically correct, executable scripts with high coverage of input fields. Despite its importance, this task remains underexplored in the context of large language models (LLMs), and no public benchmark or dataset exists to evaluate LLMs on form interaction generation systematically. This paper introduces a novel method for training LLMs to generate high-quality test cases in Selenium, specifically targeting form interaction testing. We curate both synthetic and human-annotated datasets for training and evaluation, covering diverse real-world forms and testing scenarios. We define clear metrics for syntax correctness, script executability, and input field coverage. Our empirical study demonstrates that our approach significantly outperforms strong baselines, including GPT-4o and other popular LLMs, across all evaluation metrics. Our work lays the groundwork for future research on LLM-based web testing and provides resources to support ongoing progress in this area.

Index Terms—Web application testing, Large language models, Form interaction automation

I. INTRODUCTION

Web applications are at the core of modern digital infrastructure, enabling critical services across domains such as finance, healthcare, education, and e-commerce. Ensuring their correctness and robustness is essential, as even small defects in user-facing components—especially web forms—can lead to serious usability issues or functional failures. Among the wide range of web functionalities, form interaction is particularly important and ubiquitous, facilitating tasks such as user registration, login, checkout, and data submission.

While frameworks like Selenium have become standard tools for automating web testing, they typically rely on manually authored test scripts. Writing these scripts is labor-intensive and error-prone, especially when dealing with modern, dynamic web forms that include conditional inputs, client-side validation, and real-time feedback. Despite the critical role of form testing, there exists no public dataset or benchmark that focuses on this specific task, limiting the development and evaluation of automated solutions. Moreover, existing large language models (LLMs), although capable of generating code, often struggle with producing syntactically correct,

executable, and semantically accurate test cases for real-world web forms.

The emergence of large language models (LLMs), such as ChatGPT and GPT-4 [1], has enabled new capabilities in automating complex tasks across domains. In web-based environments, LLM-powered agents have been developed to perform operations such as navigation and interaction by leveraging instruction-following and tool-use capabilities [2]–[4]. Several studies have investigated LLMs in web automation settings, introducing text-based browsing interfaces and strategies for managing verbose HTML structures through simplification and restructuring [5]–[8]. WebVoyager [9] further advances this line of work by incorporating multimodal LLMs into end-to-end automated web testing pipelines.

Despite these advances, most existing approaches primarily target high-level navigation or general-purpose web exploration. In contrast, fine-grained form interaction—such as accurately filling out input fields and generating executable scripts for testing frameworks like Selenium—remains underexplored. Existing systems often lack robustness in generating test scripts with comprehensive input coverage and execution reliability. Addressing this gap is essential for advancing LLM-driven web testing toward practical integration into software development pipelines.

In this paper, we introduce a novel approach for training large language models (LLMs) to generate high-quality form interaction test cases in Selenium. Our method emphasizes three key criteria: syntax correctness, executability, and input field coverage. To support this goal, we construct and release the first form interaction testing dataset, which includes both human-written and LLM-generated examples covering diverse real-world web forms. The contributions of this paper are as follows.

- We propose a training approach for LLMs to generate executable Selenium test scripts that ensure syntax correctness and achieve high input field coverage.
- We develop a dataset for form interaction testing, combining human-annotated and synthetic web data.
- Extensive experiments demonstrate that our approach significantly outperforms strong baselines, including GPT-4o, across multiple evaluation metrics.

* Corresponding author: lnkhang@jaist.ac.jp.

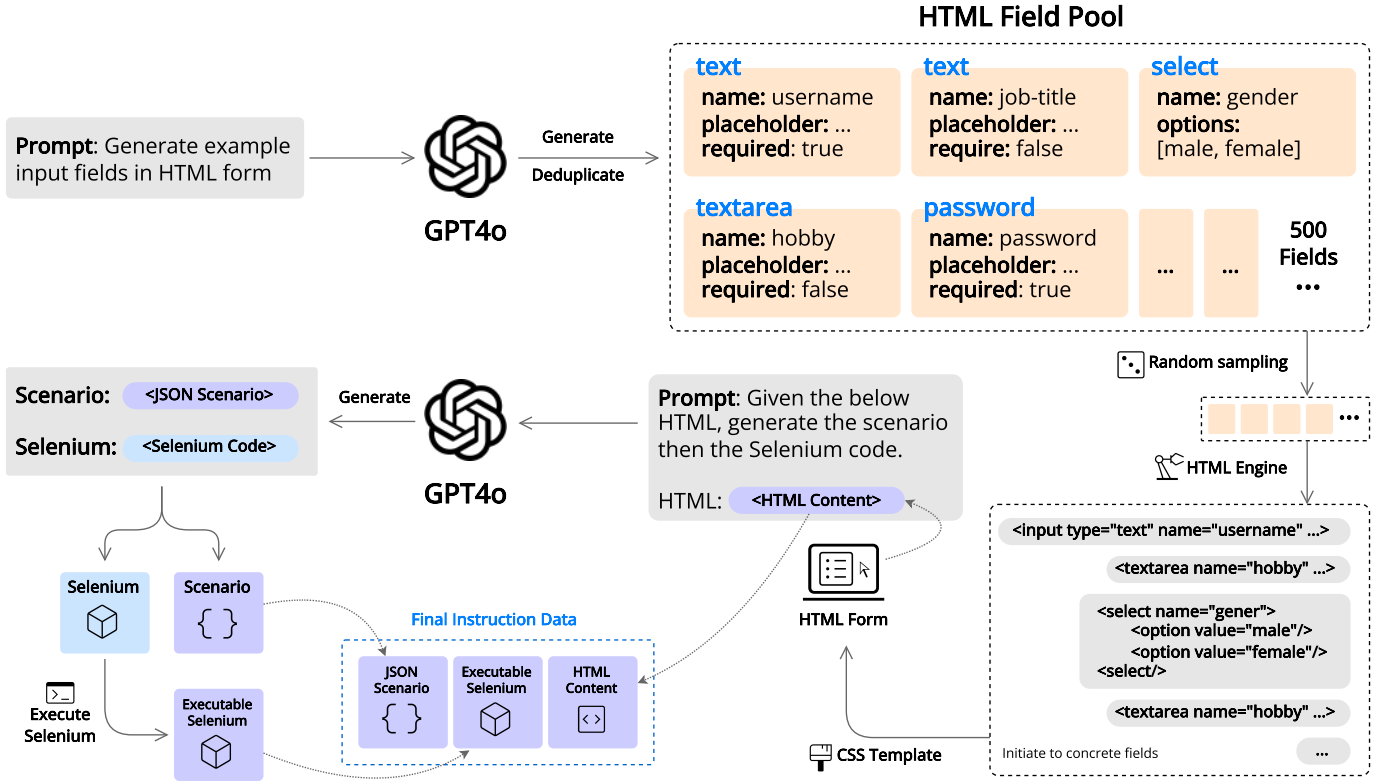


Fig. 1: Overview of the data creation process.

II. RELATED WORK

Web application testing is a fundamental yet resource-intensive task in modern software engineering. Traditional research has focused on both static and dynamic analysis techniques to ensure the functionality and security of web systems. Early efforts, such as [10] and [11], laid the groundwork by exploring systematic approaches for improving test coverage and detecting faults through formal analysis and test case generation.

With the growing complexity of web applications, automation frameworks like Selenium have become essential tools for simulating user interactions and validating UI behavior. Despite their effectiveness, these tools typically rely on hand-crafted test scripts, which are time-consuming to author and maintain. Studies such as [12] have surveyed the inherent challenges in automating web application testing, emphasizing the need for scalable and low-maintenance solutions. More recent frameworks attempt to alleviate manual effort by recording user actions and translating them into test cases automatically [13]. While such techniques reduce scripting overhead, they still depend on access to human demonstrations and lack adaptability to unseen page structures or form behaviors.

In parallel, researchers have begun exploring the use of large language models (LLMs) to automatically generate test scripts. For instance, [14] proposed using generative models that adapt to changes in the application under test, enabling more resilient and dynamic testing workflows. WebVoyager [9]

further advances this direction by deploying multimodal LLM agents to explore and interact with live websites end-to-end. However, existing LLM-based methods often struggle with key aspects of form interaction, including syntax correctness of the generated scripts, execution reliability, and comprehensive input field coverage—especially in forms with conditional logic or validation constraints.

Our work addresses these challenges by introducing a dataset-driven approach to training LLMs for form interaction testing using Selenium. Unlike prior methods that treat form interaction as an auxiliary task, we treat it as a central objective, developing human-annotated and LLM-generated datasets specifically designed for evaluating and training models on script generation quality. By measuring and optimizing for syntax correctness, script executability, and input field coverage, our method achieves robust test case generation across diverse web forms. This approach complements and extends existing research by focusing on the intersection of LLMs and practical web testing needs—particularly in the domain of automated, reliable form filling.

III. METHODOLOGY

A. Training Data Creation

To generate high-quality instruction data for training a large language model (LLM) to produce test scenarios and corresponding Selenium code, we design a multi-stage pipeline centered around GPT-4o, as illustrated in Figure 1.

TABLE I: Prompts used in GPT4o to generate the scenario and Selenium code.

Generate a test case scenario for filling in a given HTML form, providing step-by-step instructions, dummy data, the expected outcome. The test case scenario must be JSON-formatted.
The goal is to ensure that each form element is filled appropriately, actions are conducted in the correct order. # Steps
1. **Form Analysis** : Break down the HTML form into its major fields (e.g., text inputs, radio buttons, dropdowns) and requirements.
2. **Generate JSON Test Case** : - Identify each field by its name or identifier.
- Identify if the field is required and only include required fields in the test case.
- Include the html snippet for each form field.
- Include appropriate dummy data for each form field and make sure the date is valid and follows the field's requirements. For entering date, use format 'mm/dd/yyyy'.
- Specify the instructions for filling out each field in sequential order.
- Mention the expected outcome (e.g., successful submission, error messages indicating missing required fields).
- Exclude the submit button and use the 'form.submit()' method for submitting forms instead of click button.
Output Format
Test Case JSON : - Formatted in a structured JSON that describes the form fields, the data to be filled, step-by-step instructions, and the expected outcome.
JSON Structure : [Json structure]

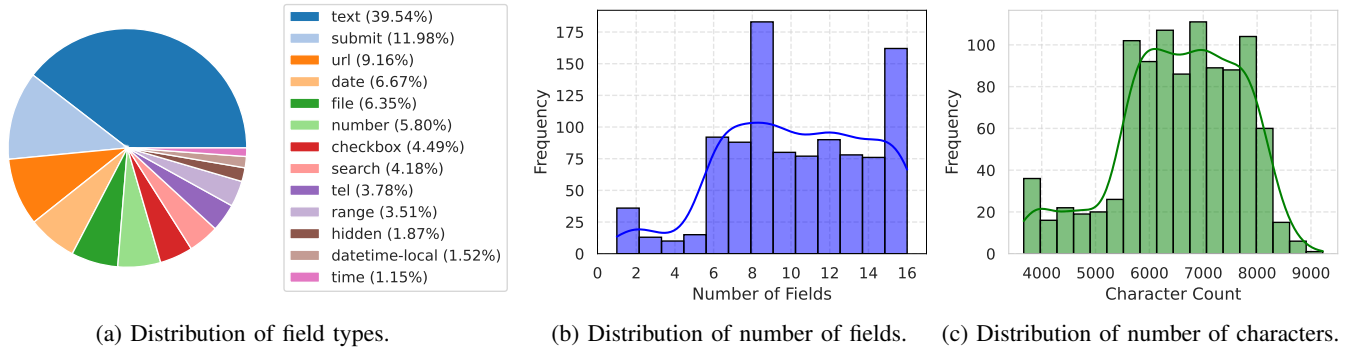


Fig. 2: Statistics of the synthetic HTML in training data.

The process begins by prompting GPT-4o to generate diverse input fields commonly found in HTML forms, such as `<input>`, `<textarea>`, `<select>`, and `<password>`. Each field is annotated with attributes such as `name`, `placeholder`, `required`, and, in the case of selection fields, a list of `options`. Redundant entries are removed to create a clean and diverse HTML Field Pool containing 500 unique form field configurations.

A random sampling strategy is then applied to select a subset of fields from this pool. These fields are assembled into a concrete HTML form using an HTML engine and styled with a predefined CSS template to enhance realism. Figure 2 shows the statistics of the generated HTML. This generated HTML form serves as the input to a second GPT-4o prompt.

B. Test Scenario and Selenium Generation

We prompt GPT-4o to simultaneously generate both the test scenario and the corresponding Selenium code from a given HTML form. Table I shows the prompts used in GTP-4o. GPT-4o then generates two outputs: (1) a natural language scenario represented in JSON format and (2) the corresponding Selenium code that automates interaction with the form. The scenario is structured in JSON format and includes a detailed description of the user action, the relevant input fields, their identifiers (e.g., by `id`, by `name`), the interaction method, the instruction, and the expected outcome. Based on this structured

scenario, GPT-4o proceeds to generate the Selenium script. The generated code is then executed using a web engine; scripts that fail to execute are discarded, while only executable ones are retained for further use.

The final instruction data consists of three components: the HTML content, the scenario in JSON, and the executable Selenium code. This structured and realistic dataset enables the LLM to learn to generate both test scenarios and automation scripts grounded in actual web content. The key advantage of this process lies in its ability to synthesize high-quality, executable, and semantically consistent data by leveraging LLMs and controlled sampling mechanisms.

IV. EXPERIMENTS

We evaluate our approach on the task of generating Selenium scripts for form interaction testing. Our experiments assess the ability of large language models (LLMs) to produce scripts that are syntactically correct, executable, and achieve high coverage of input fields. Below, we describe the models, datasets, and evaluation metrics used in our study.

A. Models

GPT Family: We compare our approach with the GPT family models, including *GPT4o* and *GPT-OSS* (20B and 120B versions), which serve as a strong baseline for code generation.

State-of-the-art LLMs: We use the families of *Qwen2.5* [15], *Qwen3* [16], and *Llama3.1* [17], which are the recent state-of-the-art open-source language models with strong performance on code generation tasks. We employ their instruction-tuned variants and fine-tune them on our synthetic data of form interaction. We compare the performance of the original model with the model finetuned using our approach.

Code generation LLMs: We use the model *Qwen2.5 Coder Instruct* [18]. We fine-tune them on our synthetic data of form interaction. We compare the performance of the original model with the model finetuned using our approach.

B. Datasets

We evaluate the models on two test sets:

- **Synthetic Forms:** A curated collection of 500 HTML forms programmatically generated to cover a wide range of input types (e.g., text, email, password, checkbox, radio, select), layout patterns, and interaction complexity (e.g., required vs. optional fields, conditional logic).
- **Real-World Forms:** 133 web forms collected from publicly available websites, including login, registration, and contact forms. These forms reflect real-world HTML structure diversity and domain-specific constraints.

C. Evaluation Metrics

To measure the quality of generated scripts, we use the following three metrics:

- **Syntax Correctness:** Whether the generated script is syntactically valid. We validate Selenium scripts using language-specific syntax checkers.
- **Executability:** Whether the script runs without errors in a headless browser environment (e.g., Chromium). This checks runtime correctness and functional validity.
- **Input Coverage:** The proportion of input fields in the form that are correctly identified and filled by the script. This metric captures the completeness of the test case in terms of form interaction.

V. RESULTS

A. Main Results

Table II presents the performance comparison on real-world form interaction testing. Across different model families, our approach consistently improves executability and input coverage compared to the original baselines, while maintaining high syntax correctness. The filtering strategy proves effective in eliminating noisy or unexecutable samples, leading to more reliable script generation that runs successfully and covers a broader range of input fields. Importantly, the improvements are observed across diverse backbones, including Qwen, Llama, and code generation models, demonstrating that our method generalizes well beyond a single model family. These results confirm the effectiveness of our approach in enhancing LLM-based form interaction testing in realistic web environments.

Table III reports the performance on synthetic form interaction testing. Across all models, our approach consistently improves executability and input field coverage compared to the original settings, while maintaining perfect syntax correctness. This indicates that our approach leads to more reliable script generation that not only runs successfully but also achieves broader coverage of input fields. Notably, the improvements hold across different model backbones, demonstrating the robustness and generality of our approach in enhancing LLM-based form interaction testing.

B. Analysis of Error rate in different field types

Figure 3 shows the percentage of form field errors across different HTML field types for the GPT-4o model, and baseline LLMs supervised fine-tuning with our approach (Filtered SFT), and fine-tuning without filtering, where all Selenium code is used (Full SFT). In Qwen2.5-14B-Instruct (Figure 3a), all three models exhibit relatively high error rates in complex field types such as text, dropdown, and radio, while achieving low error rates on simpler or infrequently used fields like tel, textarea, and hidden. The Filtered SFT variant displays a more even distribution of errors, reducing failure in common fields such as submit and radio compared to the Full SFT baseline. This suggests that data filtering may help reduce overfitting to noisy or rare patterns. GPT-4o shows overall lower error percentages and maintains consistency across field types, serving as a reference point for generalization performance.

In the Qwen2.5-7B-Instruct (Figure 3b) and Qwen3-14B-Instruct (Figure 3c), similar trends emerge, but with sharper contrasts between field types. Notably, Full SFT and Filtered SFT both struggle significantly with the other and text categories, indicating that smaller models may be more sensitive to ambiguous field semantics. The Filtered SFT variant again shows improved robustness in fields such as submit and radio, although its error rate on dropdown slightly increases compared to the Full SFT model. Interestingly, GPT-4o’s error pattern remains stable across both scales, reinforcing the observation that model size and fine-tuning strategy can introduce distinct biases in field-specific performance. Comparing the two model scales, Qwen2.5-14B exhibits more stable performance across field types, while Qwen2.5-7B shows more pronounced spikes in error for specific categories.

C. Analysis of Causes of Errors

Across both model scales and all training regimes, the dominant failure mode underlying the per-field-type error profiles in Figure 3 is *inconsistent identification of the intended target field* within the active interaction context (typically, the currently focused form). Concretely, the model often predicts a correct *semantic intent* (e.g., “fill email”), yet binds that intent to an *incorrect DOM element*. This manifests as: (i) selecting a visually adjacent but non-interactable wrapper (e.g., a <div> around an <input>); (ii) selecting a hidden or disabled control; (iii) picking the wrong member of a group (radio/checkbox); (iv) confusing native <select> with

TABLE II: Performance comparison on real-world form interaction testing. OUR APPROACH is evaluated under two settings: with filtered training data that excludes unexecutable Selenium code; and without filtering (w/o filter), where all Selenium code is used for training.

Model	Run	Syntax Correctness (%)	Executability (%)	Input Coverage (%)
<i>GPT Family (Baseline)</i>				
GPT-4o	Original	97.67	50.38	39.28
GPT-OSS (120B)	Original	100	45.74	25.07
GPT-OSS (20B)	Original	91.47	46.51	23.74
<i>Qwen2.5 Family of models</i>				
Qwen2.5-14B (Instruct)	Original	99.22	49.61	39.35
	OUR APPROACH (w/o filter)	100.00	58.91	39.79
	OUR APPROACH	99.22	59.69	40.53
Qwen2.5-7B (Instruct)	Original	99.22	48.84	4.96
	OUR APPROACH (w/o filter)	59.38	23.26	16.26
	OUR APPROACH	100.00	48.84	23.68
<i>Qwen3 Family of models</i>				
Qwen3-14B (Instruct)	Original	97.67	48.06	37.83
	OUR APPROACH (w/o filter)	98.44	60.47	33.15
	OUR APPROACH	100.00	60.47	35.96
Qwen3-8B (Instruct)	Original	100.00	17.24	19.12
	OUR APPROACH (w/o filter)	98.41	28.57	27.38
	OUR APPROACH	100.00	25.20	21.44
<i>Llama Family of models</i>				
Llama3.1-8B (Instruct)	Original	94.57	34.11	0.45
	OUR APPROACH (w/o filter)	99.22	39.53	34.24
	OUR APPROACH	98.45	45.74	36.80
<i>Code Generation Models</i>				
Qwen2.5-7B-Coder (Instruct)	Original	96.80	56.80	10.89
	OUR APPROACH (w/o filter)	100.00	57.85	20.12
	OUR APPROACH	98.31	66.10	15.40

TABLE III: Performance comparison on Synthetic Form interaction testing, measured on syntax correctness (Syntax), executability (Execute), and input coverage (Cover).

Model	Setting	Syntax	Execute	Cover
GPT-4o	API	99.60	74.68	76.40
Qwen2.5-14B-Inst	Original	96.40	77.40	77.97
	OUR APPROACH	100.00	82.00	86.38
Qwen3-14B	Original	100.00	65.20	85.30
	OUR APPROACH	100.00	81.60	87.18

a custom dropdown built from `<div><li>`; or (v) leaving the form scope and hitting a similarly named field elsewhere on the page. Typical Mis-Bind patterns include:

- **Text inputs** (text, email, number): Choosing a placeholder -only field, a hidden clone used for masking, or a sibling used for validation.
- **Dropdowns**: Treating a custom combobox as a native `<select>` (or vice versa).
- **Radio/Checkbox groups**: Clicking the first matching input by `@type` or `@name`, ignoring the label/value.
- **Submit**: Clicking a non-submit button (e.g., a reset or a tab switch) or a disabled button before client-side validation completes.

- **Hidden/Offscreen**: Selecting the elements with attributes `display:none`, `visibility:hidden`, `zero-size`, or `off-canvas`.
- **Context leakage**: Leaving the current form and interacting with a homonymous field elsewhere (e.g., header search).

D. Examples of Faulty Selenium Snippets and Implications

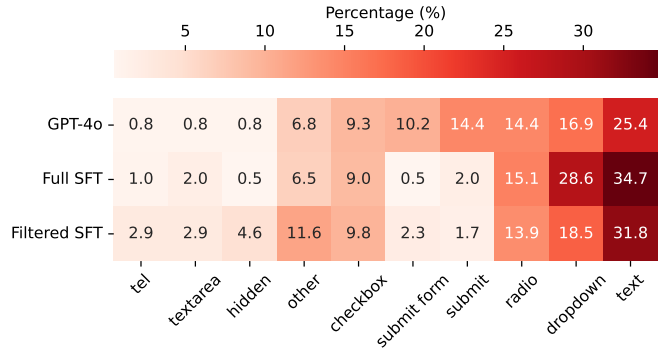
To better illustrate how incorrect field identification translates into faulty automation scripts, we analyze common failure patterns observed in generated Selenium code. The following examples reflect real mis-bindings found in the evaluation set, highlighting the subtle but critical mismatches between semantic intent and DOM binding.

- **Example 1: Misbinding to Wrapper instead of Input**

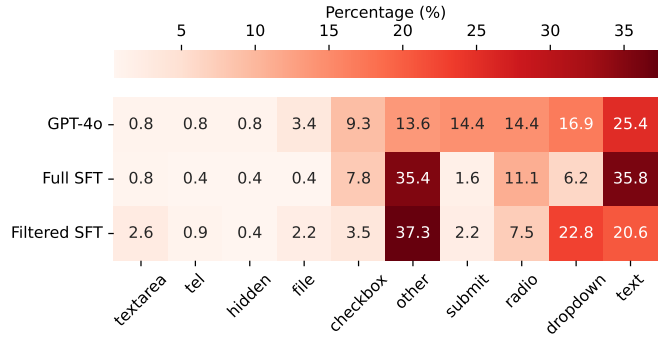
```
driver.find_element(By.XPATH,
    "//div[@id='mail']").send_keys("a@a")
```

This line targets a `<div>` that visually wraps the input field (e.g., for styling or layout). Since `<div>` is not an interactable input element, this action raises a `ElementNotInteractableException`.
- **Example 2: Selecting Disabled or Hidden Field**

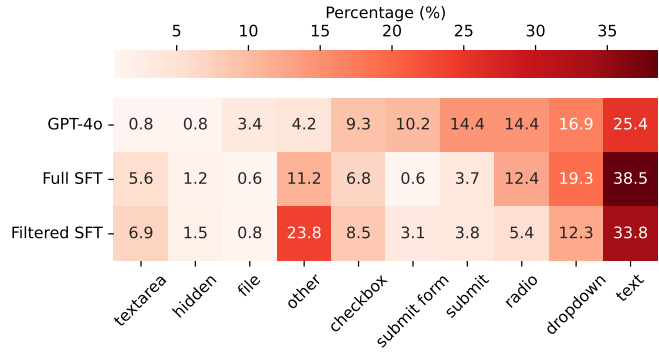
```
driver.find_element(By.CSS_SELECTOR,
    "input[name='tel']").send_keys("123")
```



(a) **Qwen2.5-14B-Instruct's** Percentage of error in each field type.



(b) **Qwen2.5-7B-Instruct's** Percentage of error in each field type.



(c) **Qwen3-14B-Instruct's** Percentage of error in each field type.

Fig. 3: Percentage of errors across field types for various LLMs. "Filtered SFT" refers to models fine-tuned with our filtering step, where unexecutable Selenium code is excluded, while "Full SFT" denotes fine-tuning without filtering, using all Selenium code for training.

The matched element exists in the DOM but is either disabled or hidden (e.g., `display:none`). No visible feedback or state change occurs on the page.

• Example 3: Context Leakage

```
driver.find_element(By.NAME,
    "search").send_keys("query")
```

If multiple fields share the same name (e.g., one in the form, one in the navbar), the global selector may match the wrong instance.

VI. CONCLUSIONS

This work presents a comprehensive framework for form-centric web automation using large language models, addressing a critical yet underexplored area in automated web testing. By introducing both a novel data generation pipeline and robust evaluation metrics, we enable systematic training and assessment of LLMs on form interaction tasks. Our curated datasets and empirical results demonstrate that LLMs can be effectively adapted to generate syntactically correct, executable, and high-coverage Selenium scripts. The significant performance improvements over strong baselines highlight the potential of LLMs in this domain. We hope our contributions—benchmarks, datasets, and methodology—serve as a foundation for further research into LLM-based software testing and broader applications in web automation.

REFERENCES

- [1] OpenAI, "Gpt-4 technical report," *arXiv preprint*, 2023.
- [2] AutoGPT, "Autogpt," 2022.
- [3] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, and e. a. Bill Qian, "Toolllm: Facilitating large language models to master 16000+ real-world apis," *arXiv preprint*, 2023.
- [4] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *arXiv preprint*, 2023.
- [5] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, and e. a. William Saunders, "Webgpt: Browser-assisted question-answering with human feedback," *arXiv preprint*, 2021.
- [6] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, and e. a. Uri Alon, "Webarena: A realistic web environment for building autonomous agents," *arXiv preprint*, 2023.
- [7] I. Gur, H. Furuta, A. Huang, M. Safdari, Y. Matsuo, D. Eck, and A. Faust, "A real-world webagent with planning, long context understanding, and program synthesis," *arXiv preprint*, 2023.
- [8] X. Deng, Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun, and Y. Su, "Mind2web: Towards a generalist agent for the web," *arXiv preprint*, 2023.
- [9] H. He, W. Yao, K. Ma, W. Yu, Y. Dai, H. Zhang, Z. Lan, and D. Yu, "WebVoyager: Building an end-to-end web agent with large multimodal models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 6864–6890.
- [10] F. Ricca and P. Tonella, "Analysis and testing of web applications," in *Proceedings of the International Conference on Software Engineering*, vol. 25, no. 3, 2001, pp. 25–34.
- [11] A. Andrews, J. Offutt, and R. Alexander, "Test generation for web applications," *IEEE Transactions on Software Engineering*, vol. 31, no. 3, pp. 187–202, 2005.
- [12] A. Marchetto, P. Tonella, and F. Ricca, "A survey on web application testing," *ACM Computing Surveys*, vol. 44, no. 3, pp. 1–36, 2012.
- [13] J. Smith and R. Taylor, "Automated frameworks for dynamic web testing," *Software Testing Journal*, vol. 37, no. 1, pp. 45–67, 2022.
- [14] K. Lee and S. Johnson, "Leveraging generative ai for automated test case creation," in *Proceedings of ICSE*, 2022, pp. 198–207.
- [15] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
- [16] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [17] A. G. et al., "The llama 3 herd of models," 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [18] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, K. Dang, Y. Fan, Y. Zhang, A. Yang, R. Men, F. Huang, B. Zheng, Y. Miao, S. Quan, Y. Feng, X. Ren, X. Ren, J. Zhou, and J. Lin, "Qwen2.5-coder technical report," 2024. [Online]. Available: <https://arxiv.org/abs/2409.12186>

Improving Adaptive Learning by Incorporating Question Attributes: A System for Special Needs Education in Vietnam

Thi Quynh Hoa Nguyen*

Department of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
hoantq@hnue.edu.vn

Duy Hai Nguyen

Department of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
ndhai@hnue.edu.vn

Tuan Long Tran

Department of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
longtt@longit.id.vn

Tan Hai Nguyen

Department of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
nguyentanhai1610@gmail.com

Thanh Trung Dang

Department of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
trungdt@hnue.edu.vn

Xuan Hien Le

Department of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
hienlx@hnue.edu.vn

Abstract—Supporting children with learning difficulties in Vietnam requires technological solutions that are capable of deep personalization. However, current adaptive learning systems often only adjust the difficulty based on the learner’s performance (correct/incorrect), ignoring the intrinsic properties of the questions themselves. This shortcoming limits the system’s ability to accurately approach the Zone of Proximal Development (ZPD). To address this limitation, our research presents the design and development of a web-based adaptive learning system with an improved adaptive algorithm. The novelty of this algorithm is the ability to make decisions to adjust the difficulty not only based on the learner’s performance, but also based on the properties of the questions, including context (e.g., mathematics topic) and format (e.g., multiple choice, fill-in-the-blank). The system can provide more sophisticated intervention options to reinforce foundational academic skills by modelling both the learner and the question set. A preliminary evaluation study on a group of children with diverse learning difficulties showed that this multi-attribute approach not only improved targeted academic indicators but also showed potential in maintaining learner engagement. The initial data show promising results, affirming the possibility of building next-generation personalized intervention tools in Vietnam.

Index Terms—adaptive learning, support, learning difficulties, dyslexia, dyscalculia, language disorder

I. INTRODUCTION

Inclusive education plays a crucial role in ensuring that all children have access to learning opportunities that align with their abilities and individual needs [1] [2]. In this context, learning disabilities (LDs) are considered a group of developmental disorders that affect the ability to acquire, process, and retain information, often manifesting as difficulties in reading (dyslexia), writing (dysgraphia), or mathematics (dyscalculia) [3]. According to UNICEF (2023), more than 317 million children and adolescents worldwide are living with conditions that affect their learning functions, with LD rates ranging from 5% to 15%, depending on the country and assessment methods [4]. In Vietnam, official statistics are limited, but recent surveys on elementary school pupils show a significant proportion of children with difficulties

in reading, writing, or math, alongside limitations in foundational cognitive skills [5]. However, practical implementation of support faces many challenges: large class sizes, a shortage of specialized teachers, and a one-size-fits-all curriculum that is difficult to personalize.

Worldwide, Adaptive Learning Systems (ALS) have been proven to improve learning outcomes and learners’ ability to self-regulate by adjusting content, pace, difficulty, and feedback based on real-time data [6]. A typical ALS usually integrates three components: the learner model, the knowledge domain model, and the instructional model [7]. For children with LDs, many studies have shown that ALS helps optimize foundational skills, such as adjusting text complexity for children with reading difficulties or gamifying basic arithmetic for children with math difficulties [8]. However, actual effectiveness depends on the pedagogical design, data quality, and level of interaction [9]. In Vietnam, most research and applications of ALS are limited to higher education or online learning; there is no specialized system for elementary LD pupils that fully integrates the three core components and has been rigorously evaluated in the local context [10].

To address this gap, this study presents the design and development of a specialized adaptive learning system for elementary LD children in Vietnam. The novelty of the system lies in the improved adaptive algorithm, which can adjust difficulty based not only on the learner’s performance (correct/incorrect), but also by analyzing the intrinsic properties of the questions, such as context (mathematical topic) and format (multiple-choice, fill-in-the-blank). This multi-attribute approach allows the algorithm to adjust the learning level, creating appropriate challenges that not only enable learners to succeed but also maximize their potential. This enables the system to accurately identify the learner’s true developmental level, focusing on core aspects of academic skills, rather than merely reducing difficulty. As a result, learners can achieve genuine progress, not just along an easy path but also maximizing their personal potential within the Zone of Proximal Development (ZPD). The system is realized as a web-based prototype and deployed for trial testing in a controlled environment to evaluate its feasibility, user acceptance, and initial

*Corresponding author: Thi Quynh Hoa Nguyen (hoantq@hnue.edu.vn)

ORCID of authors are 0009-0004-4331-6539, 0009-0003-6001-1326, 0009-0002-6090-9120, 0009-0003-3324-4749, 0009-0000-4403-7719, 0009-0008-9890-2518 respectively.

effectiveness.

II. RELATED WORKS

ALS in specialized education are designed to adjust content, pace, challenge level, and feedback based on real-time learner data [11] [12]. This adjustment process typically integrates three core components: the learner model, the knowledge domain model, and the instructional model, with the goal of keeping tasks within the “just-right” level of difficulty [13] [7]. From a technological perspective, the adjustment focuses on two main approaches. The first is rule-based systems that use IF-THEN conditions to change difficulty or provide hints, which have the advantage of being easy to interpret but are limited in flexibility when handling complex errors. The second is systems using artificial intelligence and machine learning (AI/machine learning), including models like IRT, BKT/DKT, or bandit algorithms, which allow for deeper personalization by estimating skill states but require large datasets and face challenges in model interpretation [14]. In parallel, the effectiveness of ALS is closely tied to digitized pedagogical methods, with three common practices: explicit and systematic guidance through skill breakdown; game-based learning to maintain attention; and multisensory learning to enhance memory [15]. The level of adaptation of the system thus depends on the match between the learner’s profile (e.g., phonemic profile) and the content metadata (difficulty, prerequisites) [16]. Nevertheless, international studies have reported inconsistent effectiveness of ALS in skill domains such as accuracy and mastery, with results varying significantly depending on the study design, content quality, and implementation methods [17].

However, these systems still have significant limitations. In terms of pedagogy, some models may deviate from the ZPD or fail to distinguish the cause of learner errors, such as lack of knowledge, guesswork, or cognitive fatigue, while excessive gamification may blur learning objectives [18]. In terms of technology and systems, the content authoring process and metadata tagging remain significant obstacles, with a lack of high-quality data leading to cold-start issues, and the “black box” nature of AI reducing the interpretability of pedagogical decisions [19]. Additionally, high demands for digital literacy among teachers and parents, along with ethical concerns and data privacy issues, continue to pose challenges [20]. In Vietnam, these general barriers are further impacted by various specific factors. Infrastructure disparities between regions, the high cost of international platforms, and the limitations of domestic products complicate uniform deployment. Furthermore, the characteristics of the Vietnamese language (tones, syllabic structure), along with the limited availability of digital content aligned with the current curriculum, make model training and optimization more complex. The readiness in digital literacy, the ability to integrate with “support rooms,” and early screening processes are not uniform; legal requirements regarding parental consent, child data protection, and the standardization of digital content still need to be clarified [21]. Therefore, scientific publications on ALS specifically designed for elementary school children with learning difficulties, fully integrating core models and rigorously evaluated in the Vietnamese context, are still very limited [22].

III. PROPOSED SYSTEM

A. Overview of the proposed method

The proposed method in this study is built on a loop intervention model, including three core stages: (1) Assessment

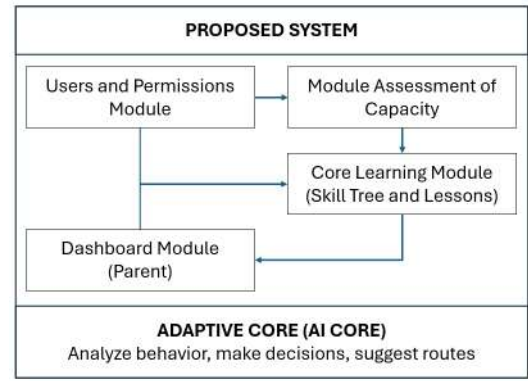


Figure 1. Overview of proposed system.

and Screening, (2) Personalized Intervention, and (3) Continuous Adaptation. The focus of this study is that the adaptation always follows the learner’s ZPD, thanks to the improvement in the adaptive algorithm. At the same time, the learning platform that the study provides is not just digital documents but a flexible learning platform that always listens to users’ needs and is maximally met by the personalization of the user experience.

The Assessment and Screening phase aims to build each learner’s competency profile. The system starts by implementing diagnostic tests, designed based on scientific assessment tools, through the Competency Assessment Module. This input data is used not only to screen for risk signs related to learning disorders (literacy, numeracy, language) but also to quantify the learner’s level of difficulty. Based on the identified competency profile, the system’s Adaptive Core (AI Core) will automatically structure a Personalized Intervention Path. This pathway consists of a series of micro-interventions that focus on learners’ core skill gaps rather than a general curriculum.

The Continuous Adaptation phase is a dynamic operating phase, ensuring that the learning environment is always maintained in the learner’s ZPD. As learners interact with the Learning Core Module, the AI Core continuously analyzes the data stream in real time to adjust the challenge level of the activities: increasing the difficulty when learners demonstrate mastery, or decreasing the difficulty and providing reinforcement exercises when signs of difficulty are detected.

This Assessment-Intervention-Adaptation loop forms a dynamic cycle, continuously optimising the learning pathway based on the learner’s progress.

The architecture of the proposed method is designed according to a modular model, including five main components that interact synergistically, with an AI Core acting as the central processing unit (Figure 1). The operation starts at the User Management & Authorization Module, an access control gateway. This module is responsible for authenticating user identities through a digital “pass” mechanism (authentication token) and assigning access rights based on the defined roles (parents, pupils), ensuring that each role is navigated to the appropriate functional interface. After authentication, learners are given access to the Competency Assessment Module, where initial diagnostic tests are performed. This module will request and display a set of questions selected from the data bank, then collect raw performance data (e.g., {question_id, result, time}, ...) and forward it to the AI Core for analysis.

The main interaction environment for learners is the Learning Core Module. This module does not decide on content itself, but

operates in a loop of instructions from the AI Core: it requests an intervention, receives instructions (e.g., display a specific exercise), executes the instructions, and finally, records all the learner interaction data in real time and sends it back to the AI Core. The entire intelligence of the system is concentrated in the AI Core. This component processes input data using two mechanisms: (1) Real-time processing to adjust micro-interventions during the session, and (2) Batch processing after a session ends to update the user model and adjust the long-term learning path. Finally, the loop is closed by the Dashboard Module, the parent/teacher interface. This module translates complex data that has been aggregated and analyzed by the AI Core into easy-to-understand visualizations (charts, progress reports) and actionable intervention recommendations. The continuous and close interaction between these modules creates a dynamic learning ecosystem capable of self-adjusting and personalizing the experience for each learner.

B. Design of improved adaptive algorithm for the proposed method

1) *Algorithm description:* To realize the principle of Continuous Adaptation, the proposed system integrates a real-time difficulty adjustment algorithm to keep the learner in the Zone of Proximal Development (ZPD). The operating mechanism of this algorithm is based on a state model and a set of predefined state transition rules. Specifically, the algorithm continuously monitors the learner's state through three main variables: the current difficulty level of the task (N), the number of consecutive correct answers (S_c), and the result history of the last k interactions (H_r , with $k = 3$). The variable S_c measures proficiency and will be reset to 0 once an incorrect answer occurs. Based on these variables, two main rules are applied to adjust the difficulty of the following task (N'). The Difficulty Escalation Rule is triggered, updating $N' = N + 1$, when the learner achieves consecutive correct answers exceeding a certain threshold ($S_c \geq 3$). Conversely, the Consolidation Rule is applied, updating $N' = N - 1$, when the system detects signs of recurring difficulty, namely when the number of errors in the most recent interaction history H_r is two or more. The difficulty level is maintained if no rule is satisfied ($N' = N$). By continuously applying these state transition rules after each user interaction, the algorithm creates a dynamic feedback loop, allowing the system to automatically adjust and personalize the learning path, ensuring that the level of challenge is always appropriate to the individual's ability and progress, thereby maximizing learning effectiveness and maintaining learner motivation.

2) *Variables and Core Concepts:* To work, the algorithm needs to track a few key parameters for each skill the child is learning:

- *current_difficulty:* An integer representing the difficulty level of the exercise (e.g. 1 = Easy, 2 = Medium, 3 = Hard).
- *correct_streak:* Counts the number of questions the child answers correctly in a row at the current difficulty level. This variable is reset to 0 each time the child answers incorrectly.
- *recent_attempts* (Recent answer history): A list of the results of the last 3 answers (e.g. [1, 0, 1] with 1 being correct, 0 being incorrect). This list helps the system detect when a child is having difficulty repeating.

New Question Attributes:

- *error_type:* Specific tag for each micro-skill question it tests (e.g., ghep_van_oan, so_sanh_so_2_chu_so).

- *context:* The context of the question (e.g., truong_hoc_gia_dinh, dong_vat).
- *type:* How the question is presented (e.g., trac_nghiem, keo_tha, dien_khuyet).
- New User state variables (per skill): *current_error_type:* Records the type of error the child just got wrong so the system can focus on it.
- *failed_contexts:* A list of contexts and question types where the child has repeatedly answered wrong for *current_error_type*.

3) *Flowchart:* The flowchart in figure 2 describes the system's processing after each child answers a question.

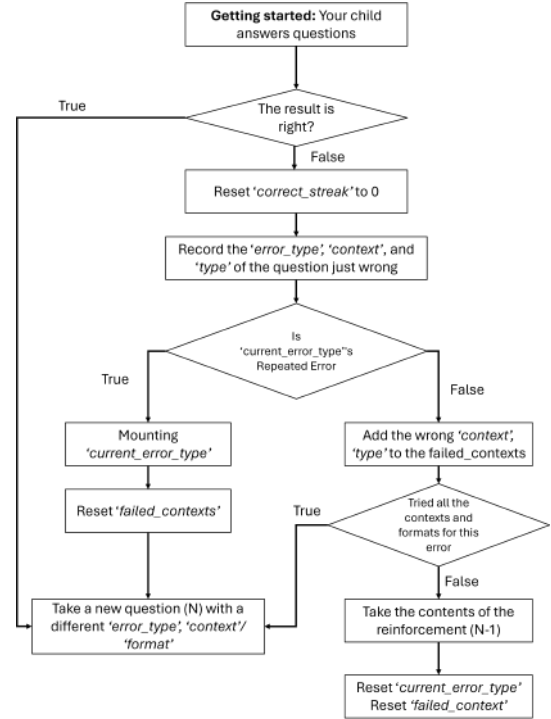


Figure 2. Block diagram of the improved adaptive algorithm for the proposed method

4) *Scenario illustration:* Situation: An is learning the skill "Rhyming 'oan'" at Medium difficulty (N).

Adaptive Algorithm (Basic)

- 1st mistake: An gets a multiple-choice question wrong about the rhyme "oan".
System: Records 1 mistake. recent_attempts is [0].
- 2nd mistake: An continues to get another multiple choice question wrong about the rhyme "oan".
System: Records 2nd mistake. recent_attempts is [0, 0]
- Analysis & Decision:
AI Analysis: The system checks and finds that An got 2/2 of the most recent questions wrong, satisfying the condition "wrong $\geq$ 2/3 questions".
AI Decision: Immediately triggers the action to consolidate knowledge.
- Result:
The system will reduce the difficulty to Easy ($N-1$) and provide a more intuitive exercise, such as matching the letters "o-a-n".

=> Comment: This algorithm quickly and effectively prevents the child from getting discouraged. However, it may have lowered the difficulty too early if An had trouble with the multiple choice question format and not the rhyme "oan" itself.

Adaptive Algorithm (Advanced)

- 1st mistake: An gets the question "Choose the word that rhymes 'oan'" wrong (multiple choice, school setting).
System: Recorded that this is the first time An made a mistake in the rhyme_oan_matching error type. The system will find another question with the same Medium difficulty and the same error type to check again.
- 2nd mistake: An continues to make a mistake in the question "Which of the following words rhymes with 'oan'?" (still multiple choice, object context).
System: Recognizes that An has made 2 consecutive mistakes in the same type of rhyming error.
- Analysis & Decision (Difference):
AI Analyzes: The system that An has failed 2 times in the multiple-choice form.
AI Decides: Does NOT reduce the difficulty. Instead, it decides to change the question type to test further.
- Result:
The system keeps the Medium difficulty (N), but gives a drag-and-drop exercise: "Drag the letters o, a, n into the blank boxes to form the correct word: kho _ _ _".

From here, the algorithm will have two directions: If An does it CORRECTLY: AI will understand that An has a good grasp of the rhyme "oan" but has difficulty with the multiple choice form. The system will continue letting An practice at Medium difficulty with other exercises. If An continues to get it WRONG in all other types (drag and drop, fill in the blanks...): Only after trying all approaches and An is still wrong, will the AI conclude that this knowledge is really too complex and reduce the difficulty to Easy (N-1).

IV. EXPERIMENT AND RESULT

The experiment was designed using a pre-test/post-test model to evaluate the effectiveness of the improved adaptive algorithm. The degree of improvement in learner performance was measured, and the effectiveness of the baseline and improved algorithms was compared to validate the value of context analysis and question formatting in the learning process.

A. Scenarios

1) *Participants & Stimulus input*: The experiment was conducted with 5 first-grade pupils (average age 6) whose native language is Vietnamese, selected from the same primary school and similar educational centers. All of them were identified as having learning difficulties at varying levels. As the proposed method targets children with LDs, the study carefully addressed ethical considerations and data privacy protection, including parental consent, anonymization of pupil information, and secure data management, to ensure research reliability and responsible deployment.pupil

The experimental subjects included 5 first graders (mean age = 6) whose mother tongue is Vietnamese, identified as having different levels of learning difficulties. Based on the individual profiles and error analysis results, the adaptive algorithm automatically retrieves and organizes exercise sequences from an online bank. The practice content is labeled according to various attributes such as target reading skills, context, and presentation format, to ensure that the level of challenge is always within each pupil's zone of proximal development.

2) *Procedure*: The experiment was conducted over 4 weeks (2–3 sessions/week) in a controlled classroom. Each pupil used a laptop and headphones to interact with the system. An intervention session (20–30 minutes) consisted of three phases: (1) Error analysis through recorded readings; (2) Prioritization of intervention goals; and (3) Targeted practice with exercises selected by the algorithm within the pupil's zone of proximal development. Data were saved for analysis and adjustment for subsequent sessions.

B. Result

1) *Overall Effectiveness Evaluation through Pre-test and Post-test Comparison*: The data collected from the pre-tests and post-tests of 5 pupils (N=5) were used to evaluate the effectiveness of the proposed method. The analysis of the results shows an increase in accuracy across all pupils after the training process, as shown in Table I.

Table I
COMPARE THE PRE-TEST AND POST-TEST SCORES OF 5 PUPILS

ID	Pre-test (/100)	Post-test (/100)	Improve
HS1	49	87	38
HS2	45	77	32
HS3	28	77	49
HS4	62	80	18
HS5	71	79	8

The quantitative results indicate a positive impact of the training process on the reading performance of the pupils involved in the experiment.

2) *Analyze the learning process and the adaptation of the system.*: The training process is carried out through a closed-loop data processing cycle, designed to transition from initial diagnosis to real-time adaptive training. This mechanism is divided into logical stages.

a) Input Assessment and Error Analysis Stage

The process begins with the pupils taking the pre-test assessment. The system automatically analyzes the results to identify the limitations and specific error types *error_type* characteristic of each pupil. Based on this analysis, an initial "competency profile" is created, serving as the foundation for personalizing the learning path.



Figure 3. The pre-test assessment interface

b) Personalized Training Path Development Stage

Based on the competency profile, the system creates an initial personalized training path for each pupil. This path is not fixed but consists of a sequence of suggested exercises targeting the specific errors identified. This ensures that each pupil begins the session with content most relevant to their needs.

c) Training and Adaptive Adjustment Stage

Content-based adaptive adjustment

The algorithm dynamically adjusts the learning path by analyzing real-time performance data, including results, time, and specific error types.



Figure 4. Learning path based on the pupil's 'competency profile

The core innovation of the algorithm is its response to incorrect answers: instead of immediately simplifying the task, the algorithm probes the nature of the error by presenting a challenge in a new context. This approach helps distinguish between short-term errors and systemic skill gaps. Interventions at the easier level (N-1) are only applied after weaknesses have been confirmed across multiple contexts, ensuring that the learning path is always dynamically adjusted based on evidence of the pupil's actual ability.

Interface customization for reading support

In conjunction with content adjustment, the system provides a control panel that allows learners to actively personalize their display experience (Figure 5). This feature is designed to support pupils with specific reading difficulties (dyslexia). Learners can customize various elements.



Figure 5. Learner's display customization interface

Allowing pupils to adjust these presentation attributes helps reduce visual stress and perceptual barriers, thereby enabling pupils to focus better on the training content.

d) Effectiveness Evaluation Stage

After completing the training path, pupils take the post-test. The data from this assessment is used to measure the improvement and overall effectiveness of the personalized and adaptive training process.

3) *Case Study Analysis*: In a reading exercise, the system recorded a characteristic interaction sequence with pupil HS03. This process clearly illustrates the micro-diagnostic mechanism of the adaptive algorithm, including the following steps:

- 1) **Initial Error Recording**: When asked to read the sentence "...*Ảnh nắng chiếu khắp mọi nơi...*", the system recorded two pronunciation errors from HS03 related to similar sounds:
 - Target word: **nắng** → Mispronounced as: **năng** (error with the sound /ă/).
 - Target word: **khắp** → Mispronounced as: **khắc** (error with the rhyme /ăp/).
- 2) **Adaptive Mechanism Deployment**: Instead of immediately concluding that HS03 lacked these skills and lowering the exercise level, the system implemented the '**context probing**' mechanism. Specifically, the system temporarily

navigated to a sentence with a completely different topic: "*Bạn Hằng sắp đọc bài trước lớp.*"

- 3) **Analysis and Conclusion**: This choice was intentional because the new sentence contained the words "*Hằng*" and "*sắp*", which directly tested the two initial errors. When HS03 correctly read this entire sentence, the system had enough data to conclude that the initial errors were **temporary performance errors**, not **systemic competency gaps**.

As a result, the pupil's competency level remained unchanged, and the learning path continued without any unnecessary reinforcement loop. This demonstrates that the system accurately assessed and optimized the intervention process.

V. CONCLUSIONS

The proposed system architecture is designed in a modular model, consisting of five main components that interact synergistically, with an AI Core acting as the central processing unit. The operation starts at the User Management & Authorization Module, an access control gateway. This module is responsible for authenticating user identities through a digital "pass" mechanism (authentication token) and assigning access rights based on the defined roles (parents, pupils), ensuring that each role is navigated to the appropriate functional interface. After authentication, learners are given access to the Competency Assessment Module, where initial diagnostic tests are performed. This module will request and display a set of questions selected from the data bank, then collect raw performance data (e.g., {question_id, result, time}, ...) and forward it to the AI Core for analysis. The main interaction environment for learners is the Learning Core Module. This module does not decide on content itself, but operates in a loop of instructions from the AI Core: it requests an intervention, receives instructions (e.g., display a specific exercise), executes the instructions, and finally, records all the learner interaction data in real time and sends it back to the AI Core. The entire intelligence of the system is concentrated in the AI Core. This component processes input data using two mechanisms: (1) Real-time processing to adjust micro-interventions during the session, and (2) Batch processing after a session ends to update the user model and adjust the long-term learning path. Finally, the loop is closed by the Dashboard Module, the parent/teacher interface. This module translates complex data aggregated and analyzed by the AI Core into easy-to-understand visualizations (charts, progress reports) and actionable intervention recommendations. The continuous and close interaction between these modules creates a dynamic learning ecosystem capable of self-adjusting and personalizing the experience for each learner.

REFERENCES

- [1] Mayonel J Jardinez and Lexter R Natividad. The the advantages and challenges of inclusive education: Striving for equity in the classroom. *Shanlax International Journal of Education*, March 2024.
- [2] UNICEF. Inclusive education is the most effective way to give all children a fair chance to go to school, learn and develop the skills they need to thrive. <https://www.unicef.org/education/inclusive-education>, 2023. Accessed: 2025-08-15.
- [3] Priya Srikanth Rao, Manoj K. Pandey, Prabha Mishra, Seema Deshmukh, Masroor Jahan, and Shivananda Manohar J. Is training working memory in children with learning disabilities a viable solution? a systematic review. *Annals of Neurosciences*, October 2023.
- [4] Menkeoma Laura Okoli, Chukwuemeka E. Ogbu, Chioma O. Enyi, Ibuchim C. Okoli, Ronee E. Wilson, and Russell S. Kirby. Sociodemographic and socioeconomic correlates of learning disability in preterm children in the united states. *BMC Public Health*, February 2022.

- [5] Hai Nguyen Xuan, Thao Do Thi, Ngoc Tran Thi Bich, and Hang Le Thi Thuy. Developing inclusive education support models for children with disabilities in vietnam. *International Journal of Elementary Education*, July 2023.
- [6] Elvira G. Rincon-Flores, Leticia Castano, Sadie Lissette Guerrero Solis, Omar Olmos Lopez, Carlos Felipe Rodríguez Hernández, Laura Angélica Castillo Lara, and Laura Patricia Aldape Valdés. Improving the learning-teaching process through adaptive learning strategy. *Smart Learning Environments*, June 2024.
- [7] Nur Baiti Afini Normadhi, Liyana Shuib, Hairul Nizam Md Nasir, Andrew Bimba, Norisma Idris, and Vimala Balakrishnan. Identification of personal traits in adaptive learning environment: Systematic literature review. *Computers & Education*, 130:168–190, March 2019.
- [8] Penelope J. Standen, David J. Brown, Mohammad Taheri, Maria J. Galvez Trigo, Helen Boulton, Andrew Burton, Madeline J. Hallowell, James G. Lathe, Nicholas Shopland, Maria A. Blanco Gonzalez, Gosia M. Kwiatkowska, Elena M. Stefano C, Annaleda M, Marco Traversi, and E H. An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities. *British Journal of Educational Technology*, 2020.
- [9] Tumaini Kabudi, Ilias Pappas, and Dag Håkon Olsen. Ai-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, 2021.
- [10] Hong-Van Thi Dinh, Quynh Anh Thi Nguyen, Ngoc Hai-Tran, Le-Hang Thi-Do, Hung Thanh Nguyen, Long Minh Dau, and Bao-Phan Phung-Dinh. Professional learning communities in vietnamese primary schools in the educational reform context: Forms and challenges. *European Journal of Educational Research*, January 2023.
- [11] Husam Yaseen, Abdelaziz Saleh Mohammad, Najwa Ashal, Hesham Abu-saimeh, Ahmad Ali, and Abdel-Aziz Ahmad Sharabati. The impact of adaptive learning technologies, personalized feedback, and interactive ai tools on student engagement: The moderating role of digital literacy. *Sustainability*, January 2025.
- [12] Xiyu Wang, Yukiko Maeda, and Hua-Hua Chang. Development and techniques in learner model in adaptive e-learning system: A systematic review. *Computers & Education*, 225:105184, February 2025.
- [13] Deborah L. Taylor, Michelle Yeung, and A. Z. Bashet. *Personalized and Adaptive Learning*. Springer International Publishing, 2021.
- [14] Xiaoman Wang, Rui “Tammy” Huang, Max Sommer, Bo Pei, Poorya Shidfar, Muhammad Shahroze Rehman, Albert D. Ritzhaupt, and Florence Martin. The efficacy of artificial intelligence-enabled adaptive learning systems from 2010 to 2022 on learner outcomes: A meta-analysis. *Journal of Educational Computing Research*, May 2024.
- [15] Ahmed Tlili, Mouna Denden, Anqi Duan, Natalia Padilla-Zea, Ronghuai Huang, Tianyue Sun, and Daniel Burgos. Game-based learning for learners with disabilities—what is next? a systematic literature review from the activity theory perspective. *Frontiers in Psychology*, 2022.
- [16] Nisha S. Raj and V. G. Renumol. An improved adaptive learning path recommendation model driven by real-time learning analytics. *Journal of Computers in Education*, 11(1):121–148, December 2022.
- [17] Xiaoman Wang, Rui “Tammy” Huang, Max Sommer, Bo Pei, Poorya Shidfar, Muhammad Shahroze Rehman, Albert D. Ritzhaupt, and Florence Martin. The efficacy of artificial intelligence-enabled adaptive learning systems from 2010 to 2022 on learner outcomes: A meta-analysis. *Journal of Educational Computing Research*, 2024.
- [18] Tingting Wang, Shan Li, Xiaoshan Huang, Zexuan Pan, and Susanne P. Lajoie. Examining students’ cognitive load in the context of self-regulated learning with an intelligent tutoring system. *Education and Information Technologies*, 2022.
- [19] Nabeel Gillani, Rebecca Eynon, Catherine Chiabaut, and Kelsey Finkel. Unpacking the “black box” of ai in education. *Educational Technology & Society*, 2023.
- [20] Hien Nguyen, Van DT Thanh, and VP Dinh. Bringing ai into teaching: Understanding vietnamese teachers’ perspectives and pedagogical challenges. *International Journal of Educational Methodology*, 2025.
- [21] T. Pham. Bring your own device (byod) in teaching the vietnamese context, 2023. Victoria University Research Repository.
- [22] Thi Thu Huyen Duong, Thi To Hoa Nguyen, and Tuyet Minh Hoang. Exploring teachers’ perspectives on adaptive learning in undergraduate programs. *International Journal of TESOL & Education*, 2024.

iTestLayout: An Automatic Pagination Algorithm for Print-Ready Examination Papers

Thanh Le Dinh

VNU University of Engineering and Technology

Hanoi, Vietnam

thanhd@vnu.edu.vn

Abstract—In educational institutions, especially during high-stakes assessments, preparing well-formatted printed exam papers is a critical yet time-consuming task. Traditional methods of exam layout creation require significant manual effort and are prone to formatting inconsistencies. This paper proposes an efficient algorithm for automatic pagination and layout generation of exam papers, designed to meet practical constraints such as spacing, question grouping, and optimal page utilization. The algorithm is implemented in the iTest system, which is currently deployed across multiple universities and testing centers. Through experimental validation on real-world datasets, the proposed method demonstrates high layout quality, correctness, and runtime efficiency. A public deployment is made available for reproducibility and live demonstration.

Index Terms—Exam paper layout, automatic pagination, iTest system.

I. INTRODUCTION

Assessment remains a cornerstone of educational systems worldwide, serving not only to evaluate student performance but also to guide instruction and curriculum development. In many educational institutions—especially in regions where digital infrastructure is still evolving—paper-based examinations are a standard and reliable modality. Despite the surge in online testing platforms, there is an enduring need for tools that support the efficient generation and printing of high-quality exam papers.

A key challenge in preparing paper-based assessments lies in the layout process. Educators frequently struggle with formatting exams that contain diverse elements such as diagrams, mathematical formulas, multi-part questions, and custom headers or footers. Manual formatting is time-consuming, error-prone, and does not scale well for institutions managing large numbers of students or subjects.

While several Learning Management Systems (LMS) and test generation software support the creation and delivery of online quizzes, few offer robust support for optimizing the layout of printable exams. Moreover, current systems often employ rigid templates or basic export mechanisms that do not adapt well to varying content structures. In practice, this results in inefficient page usage, awkward page breaks, and inconsistent formatting that can hinder student comprehension.

In this paper, we propose a novel algorithm for automatic pagination and layout optimization of exam papers, as implemented in the *iTest* system. Our solution dynamically arranges exam content—considering text length, embedded visuals,

and mathematical notation—to produce a clean and readable layout optimized for print. The approach has been deployed in numerous Vietnamese institutions and has demonstrated both technical efficiency and practical usability.

We begin by surveying existing platforms and related academic work in Section II. Section III introduces the proposed algorithm. Sections IV presents experimental results. Finally, Section V summarizes our contributions and outlines future directions for research.

II. RELATED WORK

A. Academic Research

While exam-specific layout algorithms are scarce in academic literature, related work in document formatting, pagination, and layout generation provides valuable foundations:

- **Knuth-Plass Line Breaking Algorithm** [13]: A dynamic programming method used in $\text{\LaTeX}$ for line breaking. Not directly applicable to complex documents like exams with figures and math.
- **Document Layout Generation with Deep Learning**: Techniques like LayoutGAN and DocLayNet apply GANs and transformers to generate aesthetically pleasing document layouts [14], [15].

B. Platforms and Assessment Tools

Most learning management systems (LMS) and assessment platforms focus primarily on online examination, item banking, and quiz delivery. However, the functionality for generating optimally paginated and printable paper-based exams remains underdeveloped, i.e. automatic print layout remains a largely unaddressed challenge.

- **Moodle**: An open-source LMS with a robust "Quiz" module for online testing. Moodle allows export of quizzes in HTML format, but lacks print layout optimization [1].
- **Canvas LMS**: Supports various question formats in online mode. However, it does not offer automatic pagination or printing features [2].
- **Blackboard Learn**: Focused on e-assessment with no support for optimized exam paper generation [3].
- **ExamView** and **Respondus**: Commercial tools that support test creation and export to Word or PDF with basic formatting. These systems use template-based rendering and do not implement optimization algorithms [4], [5].

- **LaTeX exam class:** A LaTeX package designed for manual exam layout. While precise and powerful, it requires technical knowledge and does not support automatic layout or adaptive pagination [6].
- **SpeedExam:** A cloud-based online examination software offering test authoring and delivery. While it includes question banking and reporting tools, it lacks specialized features for printable exam layout optimization [7].
- **Creatrix Campus:** An enterprise academic management solution supporting assessments and grading. Exam generation is included, but layout for print exams is handled via fixed templates without dynamic pagination [8].
- **McMIX** [9]: Supports randomizing multiple-choice questions, but exam layout follows a fixed template.
- **SmartTest** [10]: A commercial system supporting exam formatting using predefined Word templates.
- **AZTest** and **In.Test** [11], [12]: Tools for generating basic printable tests. Layout is rigid and manually specified.

C. Overview of iTest

iTest is a comprehensive examination management system developed to support the end-to-end workflow of computer-based assessments [16]. It offers advanced features, effectively addressing test management issues. The platform supports importing entire question banks directly from Microsoft Word, making it highly accessible to teachers without technical backgrounds. iTest has been widely adopted for its practical feature set, user-friendly interface, and high customization capabilities.

III. AUTOMATIC PAGINATION ALGORITHM

A. Definitions

Definitions Related to Exam Content:

- **Question:** A question can be either multiple-choice or essay-based. Multiple-choice questions fall into several types: single correct answer, multiple correct answers, true/false, matching, categorization, and ordering.
- **Question Group:** A group of related questions that refer to a common context (e.g., a paragraph, diagram, or image). This shared context is called the *stimulus* or *context* of the group, and the questions in the group are logically linked to it.
- **Exam Structure:** An exam comprises one or more *sections*, each containing one or more *parts*. Each part includes one or more questions.
- **Header and Footer:** The header appears once at the beginning of the exam, and the footer appears once at the end.

The overall logical structure of an exam is as follows:

$$\text{Exam} = \text{Header} \left(\text{Section} \left(\text{Part} \text{ Question}^+ \right)^+ \right)^+ \text{Footer}$$

Definitions Related to Page Layout:

- **Print Page (Page):** A rectangular printable area, typically corresponding to an A4 sheet minus margins. Margins (top, bottom, left, right) are predefined.

- **Print Block (Block):** A rectangular unit representing a complete rendering of a question, context, paragraph, table, image, etc. A block spans the full page width and stacks vertically within the page.
- **Print Compartment (Compartment):** A rectangular unit inside a block that contains either the question stem or one of its answer options. A block may include one or more compartments, laid out either vertically or horizontally.
- **Print Box (Box):** A generic term referring to either a block, a compartment, or a group of compartments.
- **Rendering Line:** A horizontal layout line containing one or more boxes.
- **Block Display:** A layout mode where a box is rendered on its own line, stacked vertically. Inspired by the CSS `display: block` model.
- **Flex Display:** A layout mode where boxes are aligned horizontally, each occupying an equal fraction of the line width. Inspired by the CSS `display: flex` model.

From a layout perspective, the printable content of an exam can be abstracted as:

$$\text{Exam Layout} = \left(\text{Page} \left(\text{Block} \text{ Compartment}^+ \right)^+ \right)^+$$

B. Constraints

Content-Related Constraints:

- An exam may contain one or more sections.
- Each section may contain one or more parts.
- Each part must contain at least one question.
- Exams may include both standalone and grouped questions, as well as both multiple-choice and essay questions.

Layout-Related Constraints:

- The width of a print block must equal the width of the print page. Its height must not exceed the page height.
- The width and height of a print compartment must not exceed the corresponding dimensions of its containing block.
- Each print block must fit entirely on a single page. A block cannot be split across multiple pages to avoid disrupting the readability of questions.
- Each compartment must be fully contained within its associated block.
- Section and part headers must not appear at the bottom of a page; they must be placed at the beginning of the block containing the first related question.
- The footer must not appear alone on a separate page.
- Excess white space at the bottom of a page must be minimized to ensure efficient space usage.

C. The Algorithm

The overall algorithm (Fig. 1) performs automatic pagination for exam papers. It first initializes necessary data structures and then proceeds through nine structured steps. These include transforming hierarchical exam content into layout-ready print blocks, optimizing visual structure for clarity, and assigning content to pages while respecting layout constraints

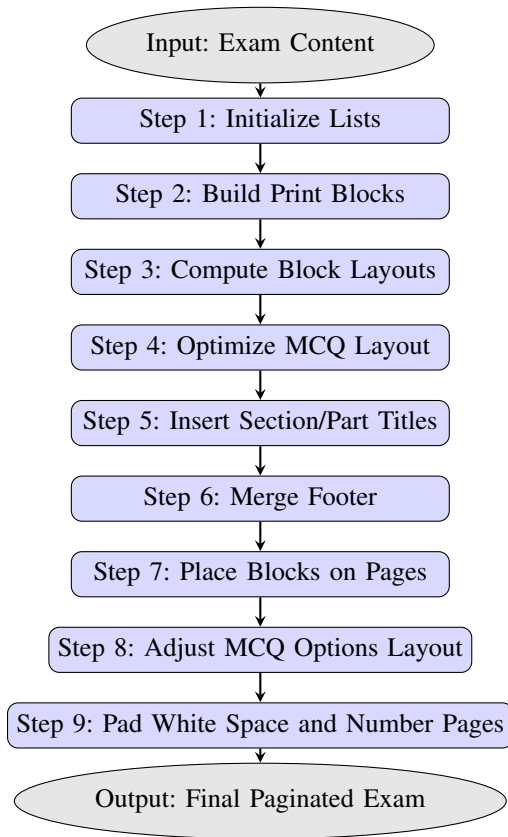


Fig. 1. Overall Flow of the Automatic Exam Pagination Algorithm

such as keeping questions intact and avoiding isolated headers or footers. The algorithm concludes by minimizing unused space and numbering the pages. Each step is elaborated in the following sub-algorithms.

Step 1: This step (Algorithm 1) initializes two primary data structures required for the layout process. The `pageList` will store the generated print pages, each of which will contain a sequence of blocks. The `blockList` will temporarily hold all the content blocks generated from the exam (questions, headers, footers, etc.) before they are assigned to pages.

Step 2: This step (Algorithm 2) translates exam elements into renderable print blocks. First, the exam header is wrapped in a block and added to the block list. The algorithm then iterates over the exam content. Each standalone question becomes a print block. For multiple-choice or true/false questions, options are laid out as individual compartments. In contrast, other types (e.g., multiple-response, matching) are grouped into a single compartment. For essay questions or context materials like images and passages, each subcomponent is encapsulated in its own block to ensure clear formatting and layout consistency.

Step 3: This step (Algorithm 3) ensures each print block and its compartments conform to page size constraints. Each block is assigned the full page width. For each compartment, width is initially estimated to accommodate its content. If this width exceeds the page width, it is capped at the page

width. The compartment's height is also calculated based on its content. All compartments in a block are stacked vertically (block display style), and the total height of the block is set accordingly. This layout step guarantees that each block fits neatly within a page's printable area.

Algorithm 1: Initialize Lists

Output: Empty lists for print pages and print blocks

- 1 Initialize `pageList` $\leftarrow$ empty list;
 - 2 Initialize `blockList` $\leftarrow$ empty list;
-

Algorithm 2: Build Print Blocks

Input: Exam content list

Output: Print blocks for header, questions, and groups

- 1 Create new block for header; add it to `blockList`;
 - 2 **foreach** *element* in *exam* **do**
 - 3 **if** *element* is a single MCQ **then**
 - 4 Create block; insert compartments for question prompt and options;
 - 5 Add block to `blockList`;
 - 6 **else if** *element* is a group question context **then**
 - 7 **foreach** *paragraph/image/table* in *context* **do**
 - 8 Create block for each; add to `blockList`;
 - 9 **else if** *element* is an essay question **then**
 - 10 Create block with compartments for question text;
 - 11 Add block to `blockList`;
-

Algorithm 3: Compute Block Layouts

Input: `blockList`

- 1 **foreach** *block* in *blockList* **do**
 - 2 Set block width = page width;
 - 3 **foreach** *compartment* in *block* **do**
 - 4 Set width = min(fit content, page width);
 - 5 Set height to fit content;
 - 6 Stack compartments vertically inside block;
 - 7 Set block height to fit all compartments;
-

Step 4: This step (Algorithm 4) optimizes the visual layout of answer options in multiple-choice questions. For single-answer questions with four options, it first calculates the width of each compartment. If any option is too wide, all options are displayed in block mode (each on a separate line). If all options are compact enough, they are arranged using a horizontal flex layout. In intermediate cases, the options are grouped into two lines, each containing two options, with compartments laid out horizontally within each line. For True/False questions, a fixed flex layout with two equal-width compartments is used. This layout improves readability and prevents option wrapping.

Step 5: This step (Algorithm 5) ensures that each section and part header appears immediately before the first relevant question block, instead of being mistakenly placed at the bottom of a page. By scanning through all blocks, it checks whether a block is the first question within its section or part, and inserts the corresponding header block above it. This

preserves the logical structure and readability of the exam layout.

Step 6: To prevent the footer from being isolated on a separate page, this step (Algorithm 6) merges the footer block with the last content block of the exam. If the footer exists and is not already part of the last block, its content is appended to the last block. This maintains visual continuity and avoids unnecessary white space or structural imbalance on the final page.

Step 7: This step (Algorithm 7) distributes all preformatted blocks into a sequence of pages while ensuring that no block is split across pages. It initializes a fresh page and iteratively checks whether a block can fit within the remaining vertical space. If it does not fit, a new page is created. This approach maintains layout consistency and readability while satisfying constraints that blocks must not span multiple pages.

Algorithm 4: Optimize MCQ Layout

```

1 foreach MCQ block do
2   if any option width > 48% page width then
3     | Set all options to block display;
4   else if all option widths ≤ 24% page width then
5     | Set to flex layout (4 per row);
6   else
7     | Group options as (2+2), display each group in
      | one row with 2 per row;
8 foreach True/False question do
9   | Set both options to flex layout (2 per row);

```

Algorithm 5: Insert Section/Part Titles

```

1 foreach first block of each part/section do
2   | Insert title content above block;

```

Algorithm 6: Merge Footer

```

1 Append footer content to last block in blockList;

```

Algorithm 7: Place Blocks on Pages

```

1 Create new page; set as currentPage;
2 foreach block in blockList do
3   if block height > remaining height on currentPage
      then
4     | Create new page; set as currentPage;
5   Add block to currentPage; stack it below
      previous blocks;

```

Step 8: This step (Algorithm 8) is designed to reduce bottom white space on each printed page by intelligently adjusting the display layout of multiple-choice questions (MCQs). In practice, many MCQs present four answer options on a single line to save space. However, when there is extra space at the bottom of a page, this algorithm selectively expands MCQ option layouts — from one line to two lines (e.g., A-B on one line, C-D on another), and if further room allows, into four separate lines. The algorithm iteratively performs this

adjustment until no further MCQs can be expanded without exceeding the available space. This step helps reduce visual gaps and improves page utilization (PU), especially for pages that would otherwise have significant bottom padding (BS).

Step 9: This final step (Algorithm 9) enhances the visual balance of each page by evenly distributing vertical white space (if any) between blocks. For pages with multiple blocks, the algorithm computes a uniform padding height (not exceeding a predefined maximum) and inserts it between consecutive blocks. It also assigns a unique page number to each page in sequential order. This ensures a clean and professional appearance for the printed test layout.

Algorithm 8: Adjust MCQ Options Layout

Input: A paginated exam with multiple pages

Output: Updated layout with reduced bottom white space per page

```

1 foreach page P in the exam do
2   while true do
3     hbottom ← remaining bottom white space on P
      // 1: Try expanding options from
      // 1-line to 2-line layout
4     foreach MCQ q on P with 4 options on 1 line
        do
5       hnew ← height of q if options shown as 2
          lines
6       if hnew - hq ≤ hbottom then
7         | Apply 2-line layout to q
8         | continue outer while loop
      // 2: Try expanding options from
      // 2-line to 4-line layout
9     foreach MCQ q on P with options shown in 2
        lines do
10      hnew ← height of q if options shown as 4
          lines
11      if hnew - hq ≤ hbottom then
12        | Apply 4-line layout to q
13        | continue outer while loop
      // 3: No expandable MCQ found
14    break

```

Algorithm 9: Pad White Space and Number Pages

```

1 foreach page in pageList do
2   Compute white space left = page height - total
      block heights;
3   Compute padding = white space / (number of
      blocks - 1);
4   Limit padding by maxPad;
5   Insert padding blocks between real blocks;
6   Assign page number;

```

IV. EVALUATION AND DISCUSSION

A. Experimental Setup

The proposed auto-pagination algorithm has been implemented as a core component of the iTest software system. The software is currently deployed in various educational institutions in Vietnam. Some of these institutions actively utilize the automatic pagination feature for preparing printed exam papers.

For experimental evaluation, we created a dataset of 150 exam papers with diverse structures and complexities. The experiments were conducted on a Intel Core i7 machine with 16GB RAM for client and a Intel(R) Xeon(R) CPU E5-2667 8-core with 16GB RAM for virtual private server, using the production version of iTest. Each test exam was formatted using the automatic pagination algorithm, and various performance and layout metrics were recorded.

B. Dataset

We collected a dataset of real-world exam papers. The dataset includes exam content across various subjects, difficulty levels, and structural formats.

In total, the dataset consists of:

- **150 exam papers** covering 15 different subjects. Each subject has 10 exam papers with the same length (i.e. the number of questions). These subjects span a diverse range of disciplines, including natural sciences (Mathematics, Physics, Chemistry, Biology), social sciences (History, Geography), and various languages (English, Chinese, Korean, Japanese, Russian, and Vietnamese). This diversity ensures that the proposed algorithm is evaluated under a wide variety of formatting challenges, such as mathematical expressions, diagrams, images, and multilingual text.
- Exams with **varying lengths**: from 30 to 90 questions per paper.
- **Diverse content types**, including:
 - Rich formatting (bold/italic text, bullet points)
 - Embedded images (diagrams, maps, photos) and tables
 - Complex math formulas (e.g., in Physics, Mathematics)

Each exam paper in the dataset was automatically paginated using the proposed algorithm. Objective layout metrics were computed as detailed in the next section.

C. Evaluation Metrics

To evaluate the effectiveness of the proposed algorithm, we used the following metrics:

- **Correctness**: The percentage of exam papers that satisfy all layout constraints (e.g., no block breaks across pages, no orphaned section headers or footers).
- **Efficiency**: The execution time required to paginate each exam, measured in seconds.
- **Layout Quality**: Measured by three sub-metrics:

- **Bottom Space (BS)**: The percentage of pages, except the last, that have large white space left at the bottom of pages. Smaller BS is better.
- **Page Utilization (PU)**: Ratio (%) between used space and total available page height. Larger PU is better.
- **Orphan Penalty (OP)**: Count of tables or images left at the top of a page.

D. Results and Discussion

Table I summarizes the experimental results. The results demonstrate that the proposed algorithm performs effectively across a variety of exam formats and contents. All test cases achieved 100% correctness, with no layout violations observed.

The average pagination time per exam was 7.14 seconds, with a standard deviation of 2.19 seconds across all test cases. This timing measurement includes the entire client-server interaction cycle, starting from the moment a user clicks to view an exam. It encompasses the network request sent to the server, the server's response containing the raw exam content, and the subsequent execution of the pagination algorithm on the client side. For subjects such as Mathematics or Physics, where exam content includes complex mathematical expressions, additional time is required to load the MathJax library and render LaTeX-based formulas. This comprehensive measurement reflects real-world usage and user-perceived performance of the system. This result indicates that the proposed algorithm maintains stable and efficient performance, with minimal variation even across exam papers of differing lengths and complexity. Such consistency makes it well-suited for integration into real-time educational platforms where rapid response and predictable behavior are essential.

The layout quality achieved by the proposed algorithm is notably high across all test cases. Specifically, the average bottom space (BS) was kept minimal, with the majority of subjects exhibiting 0% unused space at the bottom of pages. This indicates efficient vertical packing of content, reducing paper waste and enhancing visual coherence. The page utilization (PU) remained consistently high, ranging from 94.7% to 98.7%, with an average of 97.3%. This reflects the algorithm's ability to maximize the usable printable area without violating layout constraints. Besides, the orphan penalty (OP) score was zero in all cases, demonstrating that the algorithm effectively avoided isolating images or tables at the top of new pages.

These results collectively confirm that the proposed approach produces compact, visually balanced, and pedagogically appropriate layouts suitable for high-stakes examination.

It is worth noting that for certain exams—where questions frequently include images, tables, or lengthy text—the number of pages per exam may vary, even when the number of questions remains the same. This variation arises from the different permutations of question order across versions. If an image, table, or long question prompt cannot fit in the remaining bottom space of a page, the algorithm will initiate a page break and move the entire block to the next page.

TABLE I
EXPERIMENTAL RESULTS

Subject	#Exam	#Questions per exam	#Pages per exam	#Total of pages	Time (s) per exam	Correctness (%)	BS (%)	PU (%)	OP
Mathematics	10	50	7	70	7.94	100	0	97.4	0
Physics	10	60	7-8	74	8.52	100	1.2	96.7	0
Chemistry	10	60	6	60	11.2	100	0	98.1	0
Biology	10	40	8-9	61	6.23	100	2.4	95.6	0
History	10	65	6	60	5.37	100	0	98.7	0
Geography	10	70	6	60	5.48	100	0	98.4	0
Civic Education	10	90	10-11	108	7.4	100	0	98.1	0
Informatics	10	50	5	50	5.02	100	0	98.6	0
English	10	80	10	100	7.66	100	3	94.7	0
Chinese	10	30	3	30	5.25	100	0	98.4	0
Korean	10	50	8	80	5.89	100	0	97.7	0
Japanese	10	40	4	40	12.37	100	0	98.3	0
Russian	10	40	6	60	4.84	100	0	98.5	0
Vietnamese	10	30	5-6	52	5.67	100	1.05	96.4	0
Mixed	10	50	9-10	97	8.23	100	2.3	96.2	0

While this ensures layout correctness and avoids content fragmentation, it may result in additional pages being added and unused bottom space (BS) left on the preceding page. This trade-off reflects the algorithm’s strict adherence to structural and visual integrity when rendering complex exam content.

E. Public Deployment and Reviewer Access

To facilitate reproducibility and allow reviewers to experience the proposed algorithm in a real-world setting, we have deployed a dedicated instance of the iTest platform at <https://review.itest.com.vn>. A demo account with administrative access has been prepared¹:

- **Username:** admin
- **Password:** 123456

Reviewers can log in to the system, browse pre-imported exam datasets used in our experiments, and directly inspect the generated print layouts. Additionally, reviewers may create their own sample exams and observe how the auto-pagination feature arranges content for printing.

This deployment serves as a live demonstration of the proposed algorithm’s usability, correctness, and layout quality under real operational constraints.

V. CONCLUSION

This paper presents a novel algorithm for automatic pagination of printed exam papers, implemented and validated through the iTest platform. The algorithm intelligently arranges exam content across pages while respecting a wide range of layout constraints, including headers, footers, question group integrity, and space optimization.

Through deployment in real educational environments and extensive experimental evaluation, the proposed method has proven to be both efficient and robust. It significantly reduces manual workload, ensures correctness, and improves visual consistency in printed tests. Compared to existing systems and

manual practices, our approach offers substantial advantages in quality, speed, and usability.

Future work will focus on integrating user-controlled fine-tuning, adaptive template learning, and intelligent previewing mechanisms to further enhance the pagination experience. We also plan to extend the algorithm to support multilingual typesetting and more advanced page layout heuristics.

REFERENCES

- [1] Moodle Docs. (2025). *Quiz module*. Retrieved from https://docs.moodle.org/402/en/Quiz_activity
- [2] Instructure. (2025). *Canvas Instructor Guide*. Retrieved from <https://community.canvaslms.com/t5/Instructor-Guide/tkb-p/Instructor>
- [3] Anthology Inc. (2025). *Blackboard Learn Help*. Retrieved from <https://help.blackboard.com/Learn>
- [4] Turning Technologies. (2025). *ExamView Assessment Suite User Guide*. Retrieved from <https://turningtechnologies.com>
- [5] Respondus Inc. (2025). *Respondus 4.0 User Guide*. Retrieved from <https://web.respondus.com/he/respondus>
- [6] Wilkinson, P. (2023). *The exam document class for LaTeX*. Retrieved from <https://www.ctan.org/pkg/exam>
- [7] SpeedExam. (2025). Retrieved from <https://www.speedexam.net>
- [8] Creatrix Campus. (2025). Retrieved from <https://www.creatrixcampus.com>
- [9] McMIX. (2025). Retrieved from <https://aqtech.vn/chitiet-sp/AQEDuMcMixPro>
- [10] SmartTest. (2025). Retrieved from <https://smarttest.vn>
- [11] AZTest. (2025). Retrieved from <https://aztest.vn>
- [12] In.Test. (2025). Retrieved from <https://shrtlink.ai/8HyQ>
- [13] Knuth, D. E., & Plass, M. F. (1981). Breaking paragraphs into lines. *Software: Practice and Experience*, 11(11), 1119–1184.
- [14] Li, W., Du, R., & Tai, Y. (2019). LayoutGAN: Generating graphic layouts with wireframe discriminators. In *ICLR 2019*. Retrieved from <https://arxiv.org/abs/1901.06767>
- [15] Pfizmann et al. (2022). DocLayNet: DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation. In *Proc. of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3743–3751.
- [16] iTest. (2025). Retrieved from <https://itest.com.vn>

¹After logging in with the provided account, select "Exam Bank Management," choose the Subject at the top left corner, then select "Test Management."

Mathematical Kanban Under Uncertainty: A Systems Engineering Approach

1st Shee Weng LEE
Nanyang Technological University
Singapore
K2421251K@e.ntu.edu.sg

2nd TzeHoung LEE
Singapore University of Social Science
Singapore
thlee018@suss.edu.sg

Abstract—This paper presents a unified mathematical framework for modeling Kanban systems under uncertainty, with special emphasis on Systems Engineering processes and SysML v2 model development. While Kanban has been widely adopted as a visual workflow management methodology, a comprehensive mathematical foundation accounting for inherent uncertainties has remained largely unexplored. We develop an integrated mathematical approach combining stochastic queueing networks, Bayesian uncertainty quantification, and robust multi-objective optimization to characterize Kanban systems under both epistemic uncertainty (limited knowledge) and aleatory uncertainty (natural variability). The Work-In-Progress (WIP) limit emerges as the critical control parameter within this framework, enabling organizations to balance flow efficiency against robustness to uncertainty. A formal integration architecture between these mathematical models and SysML v2's metamodel is proposed, creating a unified theoretical foundation for uncertainty-aware systems modeling. Two simulation-based case studies in data-center capacity planning and pharmaceutical supply chain management indicate simulated improvements in development cycle time of 25-30 percent and enhanced model quality. This work establishes a theoretical bridge between process management under uncertainty and systems modeling disciplines.

Index Terms—Kanban, Uncertainty Quantification, Systems Engineering, SysML v2, Stochastic Queueing Theory, Bayesian Inference, Robust Optimization

I. INTRODUCTION

Systems engineering processes for complex systems development are increasingly challenged by the need to manage workflow under conditions of significant uncertainty, coordinate multiple engineering disciplines, and ensure consistent progress toward system objectives. Kanban, originating from Toyota's production system, has become a widely accepted visual management methodology to address these challenges by emphasizing work visualization, work-in-progress (WIP) limits, and flow optimization [1]. The adoption of Kanban in systems engineering, however, has largely been based on pragmatic considerations rather than on a rigorous theoretical foundation, particularly one that addresses the pervasive uncertainties in engineering processes.

No Funding Received for this Research.

At the same time, the systems engineering discipline has been enhancing its modeling capabilities with SysML, with the development of SysML v2 ongoing as of 2023 [4]. This presents an opportunity to link process management with systems modeling through rigorous mathematical frameworks that explicitly account for uncertainty. The aims of this paper are to develop a unified mathematical model of Kanban systems under uncertainty applicable to systems engineering workflows, to propose a theoretical integration between these models and SysML's metamodel, to demonstrate potential practical applications and benefits of this integrated approach, and to establish a research agenda for further development of these concepts.

By pursuing these aims, this paper addresses a substantial gap in the academic literature, where the intersection of Kanban, systems engineering, uncertainty quantification, and mathematical modeling has not yet been comprehensively explored. By establishing a unified theoretical foundation, it becomes possible to carry out quantitative analyses and robust optimizations of systems engineering processes in ways that have previously not been feasible.

II. LITERATURE REVIEW

A. Kanban Systems and Lean Approaches

Anderson [1] provides the foundational work on adapting Kanban for knowledge work and software development, describing the core principles of visualizing work, limiting work in progress, and managing flow. The theoretical foundations of flow-based systems, which underpin Kanban, are extensively described in Hopp and Spearman's work on factory physics [2]. Reinertsen [3] extends these concepts to product development, introducing the economic framework for analyzing queues and batch sizes in development processes.

B. Systems Engineering and SysML

The Systems Modeling Language (SysML) has become a standard for model-based systems engineering. The Object Management Group has been developing SysML v2, which aims to address limitations in the original specification [4]. While SysML v2 is still evolving, its planned improvements

in semantic precision and API-first architecture offer potential for better integration with process management methodologies. Friedenthal et al. [5] provide comprehensive coverage of model-based systems engineering using SysML, though based on earlier versions of the standard.

C. Uncertainty Quantification and Queueing Theory

Uncertainty quantification in engineering has traditionally focused on product design rather than process optimization. Smith [6] presents comprehensive methods for uncertainty quantification that could potentially be adapted to process contexts. Kennedy and O'Hagan [7] developed foundational work on Bayesian calibration of computer models, providing techniques that can be applied to process simulation. Queueing theory provides mathematical tools for analyzing waiting lines and service systems. Kleinrock [8] offers foundational work on queueing systems that can be adapted to model workflow in Kanban systems.

D. Research Gap

While each of these areas has a substantial body of literature, their integration remains largely unexplored. The mathematical formalization of Kanban systems under uncertainty, particularly in the context of systems engineering workflows, represents a significant gap. Furthermore, the potential integration of such mathematical models with emerging systems modeling languages like SysML v2 has not been comprehensively addressed in the current literature. Our paper aims to address this gap by proposing a unified theoretical framework that brings together these disparate fields, enabling quantitative analysis and optimization of Kanban-based systems engineering processes under conditions of uncertainty.

III. MATHEMATICAL FOUNDATIONS

A. Stochastic Queueing Network Model

We begin by extending traditional queueing theory to capture the stochastic nature of Kanban workflows. A Kanban system consists of n columns, each represented as a queue. We define $\lambda_i(t)$ as a stochastic process representing the arrival rate to column i at time t , and $\mu_i(t, \theta_i)$ as a stochastic process representing the service rate of column i at time t , parameterized by θ_i . The random variable $L_i(t)$ represents the number of items in column i at time t , while WIP_i denotes the maximum allowed value of $L_i(t)$, which is the WIP limit.

The utilization of column i becomes a random variable given by the ratio of arrival rate to service rate:

$$\rho_i(t) = \frac{\lambda_i(t)}{\mu_i(t, \theta_i)} \quad (1)$$

To capture the non-exponential nature of engineering tasks, we model each column as a G/G/c queue with bounded capacity. For such queues, approximation formulas exist for the expected waiting time W_i , which can be expressed as:

$$W_i \approx \left(\frac{c_i \rho_i^{2(c_i+1)}}{c_i! (1 - \rho_i)^2} \right) \left(\frac{CA_i^2 + CS_i^2}{2} \right) \cdot \frac{1}{c_i \mu_i} \cdot P_{WIP_i} \quad (2)$$

where CA_i and CS_i are the coefficients of variation of the arrival and service processes, and P_{WIP_i} is a correction factor for the finite capacity.

In a Kanban network, the output of one column becomes the input to the next. The blocking probability that an item completing service in column i cannot move to column $i + 1$ due to WIP limits directly impacts the effective service rate. This blocking probability affects the overall system performance by creating dependencies between adjacent workflow stages.

B. Bayesian Inference for Parameter Estimation

To address epistemic uncertainty, we employ Bayesian inference to estimate the unknown parameters of the queueing model. For each parameter θ_i in the service rate model, we specify a prior distribution $p(\theta_i)$ based on expert knowledge or historical data. Similarly, for the arrival process parameters, we specify priors where the parameters characterize the stochastic arrival process.

Given observed data D on arrival times, service times, and queue lengths, we can construct a likelihood function that quantifies how well the parameters explain the data. The posterior distribution of the parameters given the data is obtained through Bayes' rule:

$$p(\theta, \phi | D) \propto p(D | \theta, \phi) \cdot p(\theta) \cdot p(\phi) \quad (3)$$

To predict future performance metrics like cycle time, we integrate over the posterior distribution to obtain the posterior predictive distribution:

$$p(CT_{\text{future}} | D) = \int p(CT_{\text{future}} | \theta, \phi) \cdot p(\theta, \phi | D) d\theta d\phi \quad (4)$$

C. Robust Multi-Objective Optimization

With the stochastic model and Bayesian inference in place, we address the optimization of WIP limits under uncertainty. The WIP limit optimization problem is inherently multi-objective, requiring simultaneous consideration of multiple performance criteria. We formulate this as:

$$\min_{\mathbf{WIP}} F_1(\mathbf{WIP}), F_2(\mathbf{WIP}), \dots, F_m(\mathbf{WIP}) \quad (5)$$

$$\text{s.t. } \mathbf{WIP}_{\min} \leq \mathbf{WIP} \leq \mathbf{WIP}_{\max} \quad (6)$$

Typical objectives include expected cycle time, cycle time variance, 90th percentile of cycle time, probability of meeting a deadline, expected throughput, and blocking probability. To account for uncertainty in the parameters, we adopt a robust optimization perspective where we minimize the worst-case performance over an uncertainty set derived from the posterior distribution.

A key insight of our framework is that WIP limits function as uncertainty buffers in the stochastic process. By imposing WIP limits, the state space of the queueing network becomes bounded, ensuring that even under high variability or parameter uncertainty, the system remains stable and does not experience unbounded growth in queue lengths. This boundedness provides a fundamental trade-off between flow efficiency and robustness to uncertainty.

IV. SysML v2 INTEGRATION ARCHITECTURE

A. Metamodel Extension

We propose extending the SysML v2 metamodel to incorporate Kanban concepts and uncertainty elements as first-class modeling constructs. This extension includes several new elements that enable rigorous modeling and analysis. The KanbanBoard serves as the root container for Kanban workflow management, while KanbanColumn represents individual workflow stages with associated capacity constraints. WorkItem elements represent tasks being managed through the workflow, and FlowRule elements define the rules governing item movement between columns. UncertaintyParameter elements capture the stochastic characteristics of the process, while PerformanceDistribution elements track performance metrics and their statistical properties.

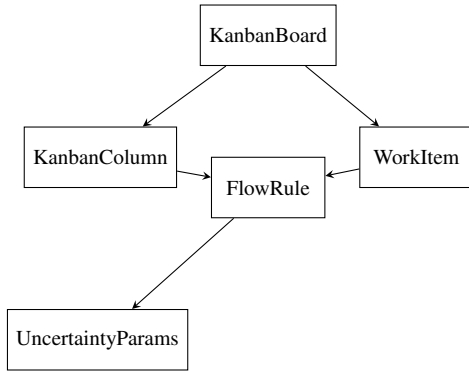


Fig. 1. SysML v2 Metamodel Extension for Kanban Under Uncertainty

B. Model Element Flow Mapping

A formal mapping connects SysML model elements to Kanban work items, ensuring bidirectional traceability between process and product development. This mapping allows forward propagation where changes in the system model create corresponding work items, and backward propagation where completion of work items updates the system model. This bidirectional relationship ensures that the engineering product (the SysML model) and the engineering process (the Kanban system) remain tightly coupled, even as they adapt to different uncertainty scenarios.

C. Uncertainty-Aware Constraint Specification

SysML v2 constraints can express conditions under which work items move between columns, incorporating uncertainty through probabilistic capacity checks and prerequisite satisfaction criteria that incorporate uncertainty in predecessor task completion. Quality thresholds can be specified with statistical tolerance bounds, enabling more sophisticated workflow management that accounts for the inherent variability in engineering tasks.

D. Implementation via SysML v2 API

The SysML v2 API-first architecture provides a programmatic interface for implementing the Kanban extensions. This includes REST API endpoints for querying and updating Kanban elements, webhook mechanisms for triggering updates when model elements or work items change, and query capabilities for analyzing the impact of uncertainty on system performance. The integration enables automatic generation of uncertainty-aware Kanban visualizations from SysML models, confidence interval overlays on WIP limits and flow metrics, and sensitivity analysis to identify which process parameters most affect performance variability.

V. CASE STUDIES

We conducted two simulation-based case studies to validate our framework and demonstrate its practical applicability. These studies illustrate how the unified mathematical framework can diagnose bottlenecks and optimize WIP limits under uncertainty in realistic engineering scenarios.

A. Datacenter Capacity Planning

A large-scale datacenter provider sought to streamline workflows for deploying server clusters and updating configurations while accounting for the inherent uncertainties in customer demand and configuration complexity. The simulation model featured a four-stage pipeline with non-stationary Poisson arrivals, log-normal service time distributions, and resource constraints. Epistemic uncertainty included limited historical data on service times for complex configurations, while aleatory uncertainty stemmed from high variability in customer request arrivals and configuration complexity.

The simulation results indicated a 25 percent reduction in expected provisioning lead time, decreasing from 10 days to 7.5 days relative to baseline. The 95th percentile lead time decreased from 15 days to 10 days, demonstrating improved predictability. The optimized WIP limits were surprisingly asymmetric with values of [5, 3, 7, 4] compared to the initial uniform limit of 5 per column. This asymmetry reflects the bottleneck in the Testing phase requiring higher WIP to maintain flow, which is counter-intuitive but optimal under uncertainty.

B. Pharmaceutical Supply Chain Management

A global pharmaceutical manufacturer applied the framework to engineering workflows for new production lines, with significant regulatory uncertainty affecting process performance. Epistemic uncertainty included limited knowledge about regulatory authority review times for novel products, while aleatory uncertainty stemmed from high variability in qualification test outcomes and regulatory decision times.

The simulation output showed a 30 percent decrease in expected total cycle time, reducing from 12 weeks to 8.4 weeks. Throughput increased from 1.5 to 2.2 completed production lines per month. The 90th percentile cycle time decreased from 18 weeks to 12 weeks, and the probability of meeting the target deadline increased from 70 percent to 88 percent. The optimal

WIP limits were found to be [4, 6, 3, 2], with the low limit on Regulatory Approval preventing upstream bottlenecks during periods of slower regulatory processing.

TABLE I
SIMULATION RESULTS SUMMARY

Metric	Datacenter	Pharma
Lead Time Improvement	25%	30%
Deadline Success Rate	65% → 92%	70% → 88%
Throughput Improvement	29%	47%

VI. DISCUSSION AND FUTURE WORK

A. Theoretical Implications

Our framework establishes a unified process-product-uncertainty theory that fuses the engineering product (SysML models), the process (Kanban), and the inherent uncertainties in a mathematically rigorous manner. This theory can be mathematically represented as a coupled stochastic differential equation system where the product state and process state evolve together under the influence of uncertainty and control inputs. This opens new research directions in computational process engineering under uncertainty, where engineering processes themselves are modeled, simulated, and optimized with the same rigor as technical product designs.

An intriguing theoretical direction is the interpretation of WIP limits through information theory. The entropy of the system state with WIP limits can be quantified, and the information gain from imposing WIP limits can be calculated as the difference between the entropy with infinite WIP limits and the entropy with the imposed constraints. This perspective provides additional theoretical insights into how WIP limits reduce system uncertainty.

B. Practical Implementation Challenges

A significant challenge in implementing mathematical Kanban models is effectively communicating uncertainty to stakeholders who may not be familiar with probabilistic concepts. We developed specific visualization templates in the SysML v2 environment to represent uncertainty in an accessible way, including confidence interval overlays and risk assessment dashboards.

The accuracy of the mathematical models depends heavily on the quality of the input data. We developed a Bayesian data augmentation approach to handle missing data in historical records, modeling the missing data as latent variables and integrating over their posterior distribution when estimating model parameters. For large-scale implementations, we recommend a hierarchical approach where each team optimizes its own workflow subject to constraints imposed by interfaces with other teams.

C. Future Research Directions

Future research should focus on developing analytical models for Kanban systems with time-varying arrival and service rates, extending the framework to handle deep uncertainty

where even the probability distributions are unknown. Reinforcement learning approaches could be applied to develop policies for adaptive WIP limits that respond to changing conditions. Causal inference methods could be used to identify which process changes would most effectively reduce uncertainty.

Integration of uncertainty-aware Kanban models with digital twins of the engineered system represents another promising direction. Methods for formally verifying properties of stochastic Kanban systems would provide additional confidence in the approach. Finally, creating open-source implementations of the SysML v2 plus Kanban framework with built-in uncertainty quantification capabilities would facilitate broader adoption and further research.

VII. CONCLUSION

This paper establishes a unified mathematical foundation for Kanban systems under uncertainty in systems engineering contexts, particularly when combined with SysML v2 for model-based development. By integrating stochastic queueing theory, Bayesian inference, and robust multi-objective optimization while explicitly addressing both epistemic and aleatory uncertainties, we enable rigorous analyses of cycle time distributions, throughput variability, WIP limit optimization, and multi-criteria trade-offs under uncertainty.

A key insight is the reframing of WIP limits as uncertainty buffers within a stochastic process framework, providing a theoretical foundation for what has traditionally been an empirical practice. This perspective enables systematic optimization of WIP limits based on the specific uncertainty characteristics of the development process. Our simulation-based case studies indicated that systematically optimizing WIP limits while accounting for uncertainty could significantly reduce both mean cycle times and cycle time variability while substantially increasing the probability of meeting deadlines.

The formal integration of these mathematical constructs with SysML v2 ensures that the engineering product and the engineering process remain tightly coupled, even as they adapt to different uncertainty scenarios. This work establishes a theoretical bridge between process management under uncertainty and systems modeling disciplines, opening new research directions in computational process engineering for complex systems development.

REFERENCES

- [1] D. J. Anderson, *Kanban: Successful evolutionary change for your technology business*. Blue Hole Press, 2010.
- [2] W. J. Hopp and M. L. Spearman, *Factory physics: Foundations of manufacturing management*, 3rd ed. Waveland Press, 2011.
- [3] D. G. Reinertsen, *The principles of product development flow: Second generation lean product development*. Celeritas Publishing, 2009.
- [4] Object Management Group, “Systems Modeling Language (SysML) v2 specification (Beta),” 2022. [Online]. Available: <https://www.omg.org/spec/SysML/2.0/>
- [5] S. Friedenthal, A. Moore, and R. Steiner, *A practical guide to SysML: The systems modeling language*. Morgan Kaufmann, 2014.
- [6] R. C. Smith, *Uncertainty quantification: Theory, implementation, and applications*. SIAM, 2014.

- [7] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B*, vol. 63, no. 3, pp. 425-464, 2001.
- [8] L. Kleinrock, *Queueing systems, volume I: Theory*. John Wiley & Sons, 1975.

Optimizing LLM with FP8

Vinh Pham Xuan
ETC Technology Systems JSC
Hanoi, Vietnam
phamxuanvinh023@gmail.com

Khoi Nguyen Le and Vinh Nguyen Van*
University of Engineering and Technology,
Hanoi VNU
khoi.n.le@vnu.edu.vn, vinhnv@vnu.edu.vn

Abstract—Recent advances in low-precision arithmetic have made FP8 [1] a compelling alternative to FP16 and BF16 for training and inference of large language models. Rather than using a uniform FP8 format, we propose a *layer-wise* specialization that assigns E4M3 to multi-layer perceptrons (MLPs) for higher mantissa precision and E5M2 to attention query-key paths for wider dynamic range. On Llama-3.2-3B and Llama-3.1-8B trained on 100K samples from OpenMathInstruct-2 [14], our method delivers (i) memory reduction on 3B from 82.4 GB $\rightarrow$ 74.2 GB ($\sim 10\%$), and (ii) training-time speedups on both scales (3B: ~ 3.6 h $\rightarrow$ ~ 2.1 h; 8B: 9.1 h $\rightarrow$ 6.6 h) on NVIDIA Blackwell-class GPUs. Crucially, we observe superior numerical stability, with *loss-variance* < 0.4 and approximately 50% lower than a Hybrid FP8 baseline that exhibits periodic spikes ≥ 0.8 [4], [10]. These results indicate that *component-aware* FP8 assignment improves stability and efficiency, with memory savings verified at 3B and time gains at both scales.

Index Terms—Large language models, FP8 precision, mixed precision training, transformer optimization, memory efficiency, computational acceleration

I. INTRODUCTION

The exponential growth in large language model (LLM) parameter counts and context lengths has created unprecedented computational and memory demands. Traditional mixed-precision approaches using FP16 and BF16 formats, while effective, still impose significant memory overhead that limits model scalability and training throughput. The introduction of 8-bit floating-point (FP8) formats [1] offers the potential to halve memory requirements while maintaining numerical stability, but current implementations face challenges in optimally leveraging the distinct characteristics of available FP8 variants.

NVIDIA’s standardization of two FP8 formats—E4M3 (4-bit exponent, 3-bit mantissa) and E5M2 (5-bit exponent, 2-bit mantissa)—presents a fundamental trade-off between precision and dynamic range. E4M3 provides higher mantissa precision within a limited range, making it suitable for stable, dense computations, while E5M2 offers broader exponent coverage at the cost of mantissa resolution, better suited for operations with wide dynamic ranges [2].

Existing FP8 training approaches predominantly employ uniform format assignment strategies. For instance, DeepSeek-V3 [3] applies E4M3 universally across all transformer components. However, our analysis reveals that different transformer components exhibit distinct computational patterns that benefit

from different FP8 formats. This work makes the following contributions:

- 1) We present a comprehensive analysis of computational patterns across transformer components, identifying optimal FP8 format assignments based on numerical characteristics and dynamic range requirements.
- 2) We propose a systematic layer-wise FP8 format assignment strategy that selectively applies E4M3 to MLPs and E5M2 to attention mechanisms based on their distinct computational profiles.
- 3) We provide extensive experimental validation across 2 models scales (3B, 8B) and comprehensive comparison with existing FP8 approaches, demonstrating consistent improvements in memory efficiency and training throughput.

II. RELATED WORK

A. Mixed-Precision Training Evolution

Mixed-precision training has evolved from early FP16 implementations [5] to more sophisticated approaches incorporating automatic loss scaling and gradient clipping [6]. The introduction of BF16 format addressed some numerical stability issues of FP16 by providing a wider exponent range at the cost of mantissa precision [7]. These approaches established the foundation for modern low-precision training strategies.

B. FP8 Format Specifications and Implementations

The IEEE 754-2019 standard introduced FP8 as a standardized low-precision format, with NVIDIA’s implementation defining two specific variants optimized for deep learning workloads [1]. The E4M3 format allocates 4 bits to the exponent and 3 bits to the mantissa, providing high precision for values within a moderate dynamic range. Conversely, E5M2 uses 5 bits for the exponent and 2 bits for the mantissa, enabling representation of values across a much wider range but with reduced precision.

Recent large-scale implementations have demonstrated FP8’s viability for transformer training. DeepSeek-V3 [3] successfully trained a trillion-parameter model using uniform E4M3 formatting, achieving competitive performance with significant memory savings. However, this approach does not leverage the complementary strengths of both FP8 formats within a single model.

* Corresponding author: Vinh Nguyen Van (vinhmv@vnu.edu.vn).

C. Quantization-Aware Training and Format Selection

Quantization-aware training approaches [8] have explored adaptive precision assignment based on layer sensitivity analysis. However, these methods primarily focus on post-training quantization or inference optimization rather than full-precision training with mixed FP8 formats. Recent work on block-wise quantization [9] demonstrates the benefits of fine-grained precision control, supporting our hypothesis that component-specific format assignment can yield significant improvements.

III. METHODOLOGY

A. Computational Pattern Analysis

To design an effective layer-wise FP8 assignment, we analyzed the numerical behavior of major transformer components. Three main patterns emerge:

1) *MLP Stability*: Feed-forward (MLP) layers exhibit concentrated activation and gradient distributions, largely confined within moderate ranges. Since their dense matrix multiplications evolve smoothly during training, they benefit from the higher mantissa precision of **E4M3**, which better preserves small value differences.

2) *Attention Variability*: In contrast, self-attention layers display wider dynamic ranges, especially in query-key interactions:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.$$

The softmax operation amplifies scale sensitivity, and attention weights often span several orders of magnitude. This motivates the use of **E5M2**, whose extended exponent range prevents overflow and underflow.

3) *Gradient Amplification*: Backpropagation through attention further increases gradient variance, particularly in query and key projections. The chain rule propagation multiplies scaling factors, producing distributions broader than those in MLP layers. Hence, stable training requires assigning formats with sufficient exponent headroom in these pathways.

B. Layer-Wise Format Assignment Strategy

Based on our analysis, we propose the following systematic assignment strategy:

MLP Components:

$$\forall \ell \in \mathcal{L}_{\text{MLP}} : W_\ell, A_\ell, G_\ell \mapsto \text{E4M3} \quad (1)$$

Attention Components:

$$\forall \ell \in \mathcal{L}_{\text{Attn}} : \begin{cases} Q_\ell, K_\ell \mapsto \text{E5M2} \\ V_\ell, O_\ell \mapsto \text{E4M3} \\ G_{Q,\ell}, G_{K,\ell} \mapsto \text{E5M2} \\ G_{V,\ell}, G_{O,\ell} \mapsto \text{E4M3} \end{cases} \quad (2)$$

where W , A , and G denote weights, activations, and gradients respectively, and Q , K , V , O represent the standard attention projections.

This assignment strategy optimizes the precision-range trade-off for each component type: - E4M3 for operations

requiring high precision within moderate ranges (MLP operations, value projections) - E5M2 for operations requiring wide dynamic range coverage (query-key interactions, attention gradients)

C. Implementation Architecture

Our implementation leverages NVIDIA's Transformer Engine to provide seamless integration with existing PyTorch workflows. Figure 1 illustrates the layer-wise FP8 format assignment within the transformer architecture, showing how E4M3 is applied to MLP components while E5M2 is selectively used for attention query-key operations.

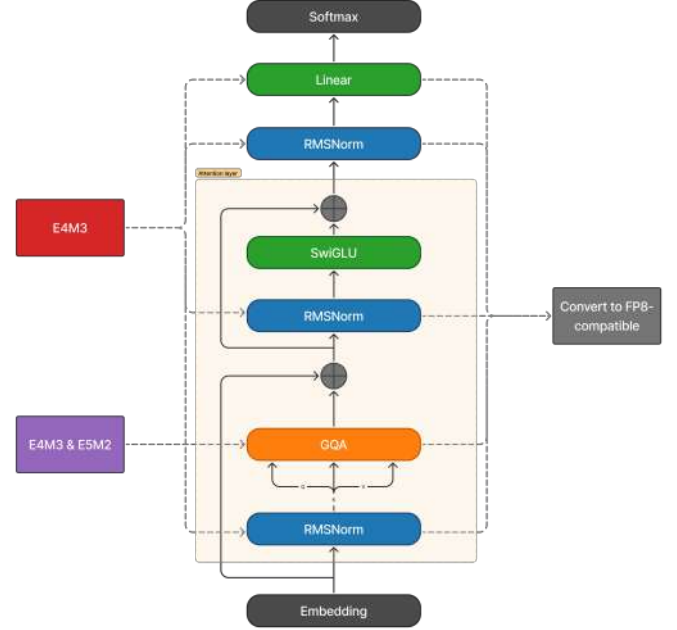


Fig. 1. Layer-wise FP8 format assignment architecture. The diagram shows a transformer block with selective application of E4M3 (red) for MLP layers and mixed E4M3/E5M2 (purple) for attention mechanisms. The GQA (Grouped Query Attention) module uses E5M2 for query-key operations requiring extended dynamic range, while E4M3 is applied to value projections and feed-forward networks for higher precision within stable ranges.

Algorithm 1 outlines our systematic module replacement strategy.

D. Training Loop Integration

Algorithm 2 presents our enhanced training procedure that incorporates dynamic format selection and optimized scaling strategies.

IV. EXPERIMENTAL SETUP

A. Model Architectures and Scales

Llama-3.2-3B [11] and **Llama-3.1-8B** [12].

B. Dataset and Training Configuration

Dataset: OpenMathInstruct-2 [14], 100K instruction-response pairs.

Schedule: 3 epochs, sequence length 512, AdamW, grad accumulation 4.

Algorithm 1 Layer-Wise FP8 Format Assignment**Require:** Original model $\mathcal{M}$, FP8 configuration $\mathcal{C}$ **Ensure:** FP8-optimized model $\mathcal{M}_{FP8}$

```

 $\mathcal{M}_{FP8} \leftarrow \text{copy}(\mathcal{M})$ 
for all  $\ell \in \text{transformer\_layers}(\mathcal{M}_{FP8})$  do
  // MLP format assignment (E4M3)
   $\ell.\text{mlp}.fc1 \leftarrow \text{FP8Linear}(\ell.\text{mlp}.fc1, \text{E4M3})$ 
   $\ell.\text{mlp}.fc2 \leftarrow \text{FP8Linear}(\ell.\text{mlp}.fc2, \text{E4M3})$ 
  // Attention format assignment (mixed)
   $\ell.\text{attn}.q\_proj \leftarrow \text{FP8Linear}(\ell.\text{attn}.q\_proj, \text{E5M2})$ 
   $\ell.\text{attn}.k\_proj \leftarrow \text{FP8Linear}(\ell.\text{attn}.k\_proj, \text{E5M2})$ 
   $\ell.\text{attn}.v\_proj \leftarrow \text{FP8Linear}(\ell.\text{attn}.v\_proj, \text{E4M3})$ 
   $\ell.\text{attn}.o\_proj \leftarrow \text{FP8Linear}(\ell.\text{attn}.o\_proj, \text{E4M3})$ 
end for
return  $\mathcal{M}_{FP8}$ 

```

Algorithm 2 FP8 Training Step with Layer-Wise Format Assignment**Require:** Mini-batch $\mathcal{B}$, parameters Θ_t , FP8 config $\mathcal{C}$ **Ensure:** Updated parameters Θ_{t+1} , loss $\mathcal{L}$ **Forward Pass**

```

for all  $\ell \in \mathcal{L}_{MLP}$  do
   $A_\ell^{E4M3} \leftarrow \text{quantize}(A_\ell, \text{E4M3}, \mathcal{C})$ 
   $W_\ell^{E4M3} \leftarrow \text{quantize}(W_\ell, \text{E4M3}, \mathcal{C})$ 
   $Z_\ell \leftarrow \text{GEMM}(A_\ell^{E4M3}, W_\ell^{E4M3})$ 
end for
for all  $\ell \in \mathcal{L}_{Attn}$  do
   $Q_\ell^{E5M2} \leftarrow \text{quantize}(Q_\ell, \text{E5M2}, \mathcal{C})$ 
   $K_\ell^{E5M2} \leftarrow \text{quantize}(K_\ell, \text{E5M2}, \mathcal{C})$ 
   $V_\ell^{E4M3} \leftarrow \text{quantize}(V_\ell, \text{E4M3}, \mathcal{C})$ 
   $\text{Attn}_\ell \leftarrow \text{ScaledDotProduct}(Q_\ell^{E5M2}, K_\ell^{E5M2}, V_\ell^{E4M3})$ 
end for
 $\mathcal{L} \leftarrow \text{ComputeLoss}(\text{model\_output}, \text{targets})$ 

```

Backward Pass

Apply format-specific gradient quantization according to Eq. 1-2

Parameter Update

Update parameters with FP32 accumulation
return $\Theta_{t+1}, \mathcal{L}$

C. Baseline Comparisons

- 1) **BF16**: standard mixed precision baseline [7]
- 2) **Hybrid FP8**: NVIDIA Hybrid (delayed scaling) using E4M3/E5M2 across layers [4], [10]
- 3) **Ours**: layer-wise FP8 (E5M2 for Q/K; E4M3 for MLP/V/O)

D. Hardware and Software Environment

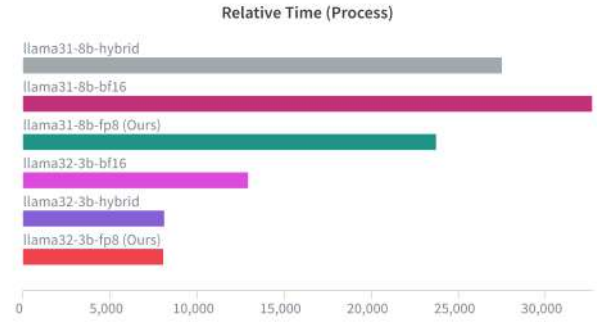
NVIDIA Blackwell-class GPUs (96 GB), CUDA 12.9, PyTorch 2.7+, Transformer Engine 2.5.0, Accelerate 0.34+.

E. Evaluation Metrics

Training efficiency (time, tokens/s, VRAM), model quality (perplexity, reasoning), numerical stability (loss variance), and resource utilization.

TABLE I
TRAINING CONFIGURATION ACROSS MODEL SCALES

Configuration	Llama-3.2-3B	Llama-3.1-8B
Batch Size (BF16)	24	6
Batch Size (FP8)	24	6
Sequence Length	512	512
Learning Rate	1e-5	1e-5
Epochs	3	3
Training Samples	100K	100K
Optimizer	AdamW	AdamW
Gradient Accumulation	4	4

V. RESULTS AND ANALYSIS**A. Memory Efficiency and Training Time**Fig. 2. Relative training time comparison. In our runs FP8 configurations achieve **1.3–1.7**× speedups over BF16 depending on scale.TABLE II
MEMORY UTILIZATION AND TRAINING TIME

Model	Precision	VRAM (GB)	Time
Llama-3.2-3B	BF16	82.4	~3.6 h
	Hybrid	74.2	~2.2 h
	FP8 (Ours)	74.2	~2.1 h
Llama-3.1-8B	BF16	80.0	~9.1 h
	Hybrid	88.0	~7.6 h
	FP8 (Ours)	88.3	~6.6 h

Findings. FP8 reduces memory by **10%** on **3B** (82.4→74.2 GB) and improves training time from ~3.6h (BF16) to ~2.1h (ours), closely matching the Hybrid FP8 baseline (~2.2h). For **8B**, our method achieves **6.6h** vs **7.6h** (Hybrid) and **9.1h** (BF16). In our setup, **8B** FP8 configurations show *higher allocated VRAM* (~88 GB) than BF16 (~80 GB), likely due to workspace and scaling-metadata overheads; thus memory savings are **confirmed at 3B** and *not universal across scales* in our current implementation.

B. Convergence and Model Quality

Both methods converge from ~0.47 to < 0.36; our layer-wise FP8 maintains lower variance and smoother dynamics.

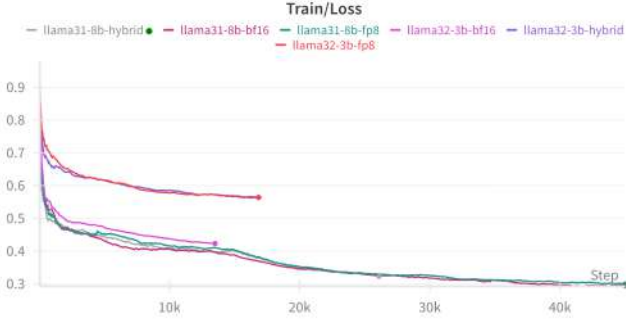


Fig. 3. Average training loss for BF16 vs FP8 Hybrid/Ours. Final loss values are comparable; our layer-wise FP8 shows smoother dynamics.

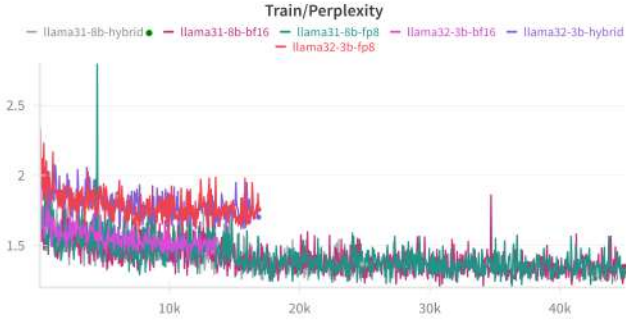


Fig. 4. Training perplexity across scales and precisions. Our layer-wise FP8 maintains perplexity in the 1.30–1.32 range with stable behavior.

C. Numerical Stability Analysis

Selective use of E5M2 for high-dynamic-range paths (Q/K) and E4M3 elsewhere yields **lower variance** (about **50%** less than Hybrid), improving stability.

VI. DISCUSSION

A. Performance Analysis

Memory–Speed Trade-off. At **3B**, FP8 reduces memory by 10% and shortens training time (2.1 h vs 3.6 h). At **8B**, we

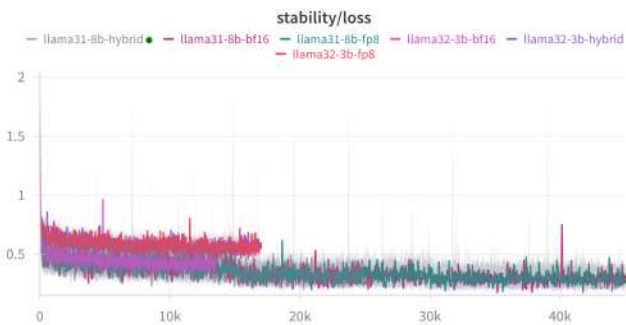


Fig. 5. Loss variance over steps. Our layer-wise FP8 (gray) stays < 0.4 ; Hybrid FP8 shows periodic spikes reaching ≥ 0.8 .

observe *time speedups* but *no memory reduction* in our implementation (likely due to additional workspaces/metadata).

Method Comparison. Layer-wise FP8 achieves **superior stability** (variance < 0.4) compared to Hybrid FP8 (spikes ≥ 0.8) [4], [10].

Practical Implications. Memory savings at **3B** enable tighter resource budgets; time gains appear at both scales, larger at **8B**.

B. Limitations and Future Work

Evaluation is limited to two Llama scales; generalization to larger models (e.g., 70B), other architectures (MoE, sparse), dynamic per-layer or per-token formats, and tighter hardware co-design are promising directions.

VII. CONCLUSION

We presented a **layer-wise FP8** assignment strategy—E5M2 for attention query–key and E4M3 for MLP/value/output projections—motivated by component-specific numerical behavior. On **Llama-3.2-3B** and **Llama-3.1-8B** with **OpenMathInstruct-2** [14], we observed: (1) **3B memory savings** (82.4→74.2 GB, -10%), and (2) **training-time speedups** at both scales (3B: ~ 2.1 h vs ~ 3.6 h; 8B: 6.6 h vs 9.1 h). Our method also delivers **50% lower loss variance** than a Hybrid FP8 baseline [4], [10]. While memory reductions were *not universal* at 8B in our setup, the combination of stability and time gains highlights component-aware FP8 as a practical path to efficient LLM training.

ACKNOWLEDGMENTS

We thank the open-source community for tools/datasets, NVIDIA for Transformer Engine and hardware documentation, and the research community for feedback.

REFERENCES

- [1] P. Micikevicius, D. Stolic, N. Burgess, M. Cornea, P. Dubey, R. Grisenthwaite, S. Ha, A. Heinecke, P. Judd, J. Kamalu, N. Mellempudi, S. Oberman, M. Shoeybi, M. Siu, and H. Wu, “FP8 Formats for Deep Learning,” *arXiv preprint arXiv:2209.05433*, 2022.
- [2] NVIDIA Corporation, “FP8 formats for deep learning,” NVIDIA Technical Report, 2022.
- [3] DeepSeek-AI, “DeepSeek-V3: A Strong, Economical, and Efficient Mixture-of-Experts Language Model,” *arXiv preprint arXiv:2412.19437*, 2024.
- [4] NVIDIA Corporation, “MXFP8: Advanced Mixed-Precision Training for Transformer Models,” NVIDIA Technical Report, 2024.
- [5] S. Narang, G. Diamos, S. Sengupta, and E. Elsen, “Mixed precision training,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [6] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, “Mixed precision training,” in *International Conference on Learning Representations*, 2018.
- [7] D. Kalamkar, D. Mudigere, N. Mellempudi, D. Das, K. Banerjee, S. Avancha, et al., “A Study of BFloat16 for Deep Learning Training,” *arXiv preprint arXiv:1905.12322*, 2019.
- [8] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference,” in *CVPR*, 2018, pp. 2704–2713.
- [9] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, “GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale,” in *NeurIPS*, vol. 35, 2022, pp. 30318–30332.

- [10] NVIDIA, “Transformer Engine: Accelerating Transformer Models with FP8 Precision,” NVIDIA Deep Learning Documentation, 2025.
- [11] Meta AI, “Llama 3.2: Collection of foundation models for global languages and vision,” Meta AI Blog, September 25, 2024.
- [12] Meta AI, “Llama 3.1: Our most capable models to date,” Meta AI Blog, July 23, 2024.
- [13] Qwen Team, “Qwen2.5: A Party of Foundation Models,” Qwen Technical Report, 2024.
- [14] S. Toshniwal, W. Du, I. Moshkov, B. Kisacran, A. Ayrapetyan, and I. Gitman, “OpenMathInstruct-2: Accelerating AI for Math with Massive Open-Source Instruction Data,” *arXiv preprint arXiv:2410.01560*, 2024.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017, pp. 5998–6008.
- [16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language models are few-shot learners,” in *NeurIPS*, vol. 33, 2020, pp. 1877–1901.
- [17] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [18] M. Wortsman, B. Rozière, and A. Szlam, “Stable and Low-Precision Training for Large-Scale Vision Models,” *arXiv preprint arXiv:2310.04407*, 2023.
- [19] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers,” *arXiv preprint arXiv:2210.17323*, 2022.
- [20] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, “SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models,” in *ICML*, 2023.

Transforming Operationalised iStar Models into Temporal Semantics and Event-B for Consistency Assurance

Hoang-Viet La Trinh, Duc-Hanh Dang*

Faculty of Information Technology, VNU University of Engineering and Technology, Hanoi, Viet Nam

Email: {24025181, hanhdd}@vnu.edu.vn

Abstract—In goal-oriented requirements engineering, models such as iStar effectively capture stakeholder intentions but lack temporal semantics and formal executability, which hinders verification and traceability in critical systems. To address this issue, this paper proposes a systematic transformation approach from iStar models to formal specifications, incorporating linear temporal semantics to ensure completeness and precision of requirements. The proposed approach bridges the gap between declarative requirement models and verifiable system specifications, while preserving goal semantics throughout the development process. Our contribution lays the foundation for the formalization of goal-oriented requirements in a consistent and extensible manner, supporting automated verification in complex software systems.

Index Terms—iStar, goal-oriented requirements engineering, linear temporal logic, Event-B, operationalisation

I. INTRODUCTION

In modern software engineering, many types of systems are no longer concerned solely with what to do, but also with when to do it and in what order. Such systems often require adherence to *temporal obligations*, such as responding within a limited time frame, maintaining valid execution order, or ensuring timeliness in decision-making processes. In this context, Goal-Oriented Requirements Engineering (GORE), particularly the iStar framework [1], has been widely used to model stakeholder intentions through constructs such as *goals*, *tasks*, and *dependencies*. When combined with operationalisation, iStar models help bridge the gap between abstract goals and concrete system behaviours.

However, ensuring that an iStar model and its accompanying operational specifications are consistent and correct remains a significant challenge—especially when applied to systems that must satisfy temporal obligations. Potential inconsistencies, such as misalignments between preconditions and actions, or violations of timing constraints, are often difficult to detect by visual inspection or through standard modelling tools. The root cause lies in the lack of *temporal semantics* and formal behavioural verification capabilities in standard iStar.

*Corresponding author

Previous research has attempted to bring iStar closer to formal specification by translating models into languages such as Z [2], Alloy [3], or Promela [4], which support structural consistency checking or static logic properties. Other approaches have formalized iStar using ontologies, such as iStar2.0-OWL [5], or description logics like BiStar [6], providing a semantically grounded model representation. While these efforts add value, they do not address operationalisation or temporal constraints—both of which are essential in reactive and real-time systems.

In parallel, the increasing complexity of adaptive and cyber-physical systems (CPS) demands goal models that specify not only *what* should be achieved, but also *when* and *how*. Models such as FLAGS [7], GRL-Compliant extensions [8], and the integration of iStar with SysML [9] have attempted to address this challenge, but have yet to provide a fully formal and verifiable operational framework.

To address these limitations, we propose a novel approach for verifying the correctness and consistency of operationalised iStar models through a two-step process: (1) extracting temporal behavioural obligations from iStar constructs (including goals, tasks, dependencies, and refinements), and encoding them using Linear Temporal Logic (LTL) [10], [11], which allows the specification of constraints such as “if A occurs, then B must eventually follow”; and (2) transforming these LTL formulas into Event-B models, consisting of state variables, invariants, and events, using structured transformation patterns and automated verification tools such as Rodin and ProB [12], [13].

The main contributions of this work are as follows:

- A technique for enhancing the expressiveness of iStar models by incorporating SES specifications and the corresponding LTL formulas.
- A method for ensuring the consistency of operationalized iStar models by transforming them into Event-B and LTL specifications.
- A case study and experiments illustrating our proposed method.

The rest of the paper is organized as follows: Section II

motivates our work with examples. Section III explains the basic idea of our approach. Section IV introduces the mapping from iStar constructs to Linear Temporal Logic formulas. Section V details the transformation of LTL obligations into Event-B models. Section VI reports on tool-supported verification using Rodin and ProB. Section VII surveys related work. Finally, Section ?? concludes the paper with a discussion of future work.

II. MOTIVATING EXAMPLE AND BACKGROUND

This section motivates our study with an example and provides the necessary background.

A. Motivating Example

In this study, we use the Air Traffic Control (ATC) domain, adapted from the EUROCONTROL technical specification [14], as our motivating example. Figure 1 presents the goal model of the ATC system, built using the Goal-Oriented Requirements Engineering (GORE) approach and modeled in the iStar language [1]. The model consists of three main actors: the *ATC_System*, which provides radar and runway data to support decision-making; the *ATC_Officer*, who evaluates landing requests and issues operational decisions; and the *Pilot*, who sends landing requests and acts upon the *ATC_Officer*'s instructions. Each rounded rectangle in the model represents a goal that must be achieved to ensure safe aircraft landing.

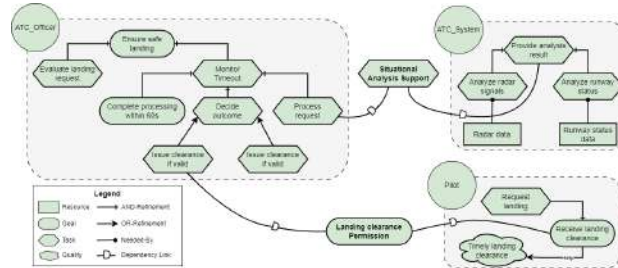


Fig. 1. iStar model of the ATC system.

Within this model, operationalisation plays a crucial role in refining abstract goals into concrete, executable tasks that human or system actors can perform [15]. For instance, the goal *Ensure safe landing* of the *ATC_Officer* is decomposed into two mandatory tasks: *Evaluate landing request* and *Monitor timeout*, forming an AND-decomposition. The latter is further refined into *Complete processing within 60 seconds*, *Process request*, and *Decide outcome*, also in AND form. The decision-making step is then OR-decomposed into *Issue clearance if valid* and *Abort landing if timeout*, ensuring that only one branch is taken at runtime depending on the context. Similarly, the *ATC_System*'s task *Provide analysis results* is refined into *Analyze radar signals* and *Analyze runway status*, both of which are required for a reliable assessment. On the Pilot side, the softgoal *Timely*

landing clearance is operationalised into *Request landing* and *Receive landing clearance*, both necessary to fulfill the timeliness requirement.

To describe these tasks in a way that is both human-readable and amenable to formalisation, we adopt the *Structured English Specification* (SES) [16]. SES captures behavioral requirements using a constrained natural language syntax, enhancing traceability and providing a bridge to formal logic. The SES statements for the ATC goal model are as follows:

P1. WHEN a landing request is received THEN the ATC Officer shall evaluate the request. (*Corresponds to task T1*).

P2. WHEN a request is received THEN the ATC System shall provide radar and runway status. (*Corresponds to task T7*).

P3. IF the radar and runway data are valid THEN the request shall be processed. (*Corresponds to task T4; resources R1, R2*).

P4. IF the request is not completed within 60 seconds THEN the landing must be aborted. (*Corresponds to task G2*).

P5. IF the request is completed on time and valid THEN issue landing clearance. (*Corresponds to decision branch T5*).

P6. WHEN clearance is issued or landing is aborted THEN the decision shall be recorded immediately. (*Corresponds to post-decision action – ATC_Officer*).

P7. WHEN a pilot sends a landing request THEN the system shall eventually respond with a decision. (*Corresponds to dependency D2*).

P8. WHEN a request is received THEN the system shall separately analyze radar signals and runway status. (*Corresponds to subtasks T8, T9*).

P9. WHEN clearance is issued THEN the pilot shall receive landing confirmation. (*Corresponds to dependency T5 → T11*).

This integrated operationalisation and SES specification make the iStar model both precise in its behavioral obligations and ready for systematic transformation into formal representations such as Linear Temporal Logic (LTL) and Event-B, enabling rigorous verification of consistency and correctness.

B. Background

iStar [1] is an agent-oriented modeling language used in early requirements engineering to capture stakeholder intentions and system rationale. Within iStar models, actors pursue goals through tasks, utilize resources, and interact via relationships such as refinements, contributions, qualifications, and dependencies. Figure 2 presents the iStar metamodel from two perspectives: Strategic Dependency (SD), showing inter-actor requirements, and Strategic Rationale (SR), detailing internal goals and

reasoning. Typical link types in iStar include AND/OR refinements, needed-by dependencies, and contributions to softgoals.

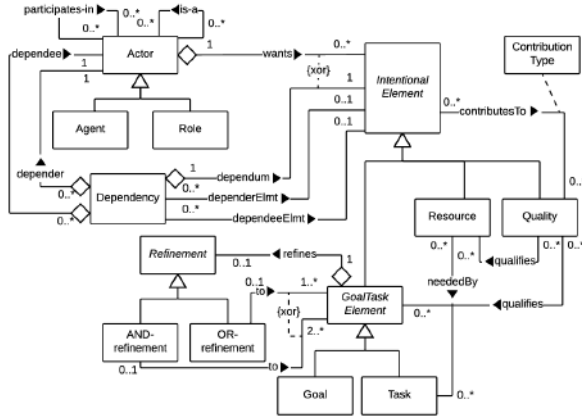


Fig. 2. Metamodel of iStar model.

Linear Temporal Logic (LTL) [10], [11] is a logic formalism used to express time-dependent behavioural constraints, such as safety (‘something bad never happens’) and liveness (‘something good eventually happens’). In this work, LTL formulas are derived from SES to capture temporal obligations implied by operationalised iStar models.

Event-B [12] is a formal modelling language that supports refinement-based development. A model comprises a static context (sets, constants, axioms) and a dynamic machine (variables, invariants, and events). Temporal obligations specified in LTL are translated into Event-B structures such as events and invariants for formal verification.

Rodin and *ProB* are toolsets used to build and validate Event-B models. Rodin supports theorem proving and invariant checking, while ProB offers model checking for temporal properties. Together, they ensure that the final specification is logically sound and behaviourally consistent with the original iStar model.

Our transformation pipeline addresses two questions: (1) How can iStar elements be mapped to temporal obligations using LTL? (2) How can these obligations be operationalised into verifiable Event-B models.

III. OVERVIEW OF OUR APPROACH

The proposed approach follows a four-step workflow, illustrated in Figure 3. Among these, two core steps constitute the main technical contribution, while the remaining steps provide necessary context and supporting activities.

Step 1 – Problem domain modelling: Define the operationalised iStar model for the target system and prepare the Structured English Specification (SES).

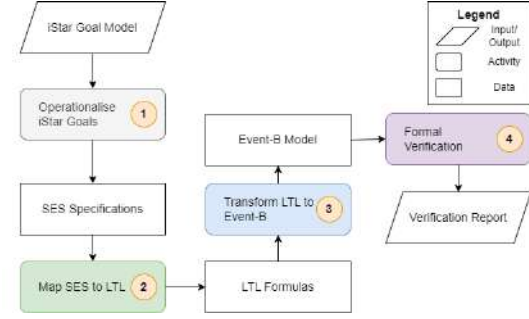


Fig. 3. Overview of our approach.

Step 2 – Extracting and formalising temporal obligations (Core Step 1): From the SES statements, behavioural constraints are mapped into Linear Temporal Logic (LTL) formulas using a predefined set of mapping rules. This step enriches the iStar model with explicit temporal semantics, capturing causal relationships and liveness/safety obligations.

Step 3 – Transforming LTL into Event-B (Core Step 2): Each LTL formula is systematically translated into Event-B components, including state variables, invariants, and events with corresponding guards and actions. This produces a complete Event-B model ready for formal verification.

Step 4 – Formal verification: Verify the Event-B model using Rodin and ProB.

In this workflow, Steps 2 and 3 are the primary focus, forming the bridge between high-level goal models and verifiable Event-B specifications.

IV. TEMPORAL SEMANTICS VIA LTL

This section introduces the temporal semantics layer by mapping operationalised iStar models into LTL formulas, enabling precise specification of time-dependent behaviors.

A. Mapping iStar Elements to LTL

The iStar framework provides a structural and intentional representation of system elements such as goals, tasks, and dependencies between actors. We focus on mapping iStar models into Linear Temporal Logic (LTL) formulas, enabling temporal reasoning and serving as a bridge toward formal modeling in Event-B.

The mapping is guided by the following principles:

- **Tasks** represent concrete executable actions. In LTL, they are modeled as propositions expected to occur in the future: $\Diamond(task)$.
- **Goals** are modeled as temporal constraints over task fulfillment. For operational goals, the structure $\Box(goal \rightarrow \Diamond(task))$ is used, expressing that the goal leads to eventual task execution.

- **AND-decomposition** is expressed through logical conjunction of sub-tasks: $\Box(goal \rightarrow (task_1 \wedge task_2 \wedge \dots))$, indicating that all branches must be fulfilled.
- **OR-decomposition** is expressed through disjunction: $\Box(goal \rightarrow (task_1 \vee task_2 \vee \dots))$, requiring at least one sub-task to succeed.
- **Dependencies** between actors are represented via implication: $\Box(request \rightarrow \Diamond(response))$, meaning that a request must eventually trigger a response.

This mapping introduces explicit temporal semantics to iStar models, capturing causality, ordering, and liveness. The resulting formulas enable traceable behavioral modeling and support formal verification in Event-B, which requires well-defined logical constraints.

B. LTL Specifications for the ATC Model

Based on the mapping rules above, we define the following LTL formulas that correspond to the nine Structured English specifications (P1–P9) derived from the ATC model. These formulas capture the system’s expected behavior in temporal terms.

TABLE I
MAPPING STRUCTURED ENGLISH TO LTL

P1	$\Box(\text{landing\_request\_received} \rightarrow \Diamond \text{evaluate\_request})$	$\rightarrow$
P2	$\Box(\text{request\_received} \rightarrow \Diamond(\text{radar\_provided} \wedge \text{runway\_provided}))$	
P3	$\Box((\text{radar\_valid} \wedge \text{runway\_valid}) \rightarrow \Diamond \text{process\_request})$	$\rightarrow$
P4	$\Box(\text{processing\_started} \rightarrow \Diamond(\text{abort\_landing} \vee \text{process\_completed\_within\_60s}))$	
P5	$\Box((\text{on\_time} \wedge \text{valid}) \rightarrow \Diamond \text{issue\_clearance})$	
P6	$\Box((\text{issue\_clearance} \vee \text{abort\_landing}) \rightarrow \Diamond \text{record\_decision})$	$\rightarrow$
P7	$\Box(\text{pilot\_request} \rightarrow \Diamond \text{decision\_returned})$	
P8	$\Box(\text{request\_received} \rightarrow \Diamond(\text{analyze\_radar} \wedge \text{analyze\_runway}))$	
P9	$\Box(\text{issue\_clearance} \rightarrow \Diamond \text{receive\_confirmation})$	$\rightarrow$

These LTL specifications model the progression of actions and decisions throughout the landing clearance workflow:

- **P1–P3** capture the initial stages: request reception, data provisioning, and request validation.
- **P4–P6** define time-sensitive constraints and decision outcomes, ensuring timely and correct responses.
- **P7** guarantees liveness by requiring that every pilot request eventually leads to a system response.
- **P8** refines system-side behavior by explicitly decomposing data analysis tasks into radar and runway evaluation.
- **P9** completes the communication loop by modeling the pilot’s confirmation receipt, thus ensuring end-to-end traceability.

This structure forms a solid foundation for translating these behavioral expectations into Event-B models in the section V.

V. TRANSLATING TO EVENT-B

After the behavioral properties of the iStar model are specified using LTL formulas, the next step is to transform these specifications into an Event-B model, in order to support formal verification of consistency and correctness. Event-B allows system states to be modeled via logical variables, constraints via invariants, and behavior via guarded events.

A. Mapping LTL to Event-B

Following the transformation pattern proposed by Machado et al. [17], we map each LTL obligation of the form $\Box(\varphi \rightarrow \Diamond\psi)$ into corresponding Event-B constructs:

- φ (trigger condition) is mapped to the WHEN clause (guard) of an Event-B event.
- ψ (expected behavior) is mapped to the THEN clause (action) of the event.
- The causal constraint is also encoded as an INVARIANT to preserve the intended relation.

For example, the obligation:

$\Box(\text{landing\_request\_received} \rightarrow \Diamond \text{evaluate\_request})$

is mapped to:

- **Variables:** `landing_request_received, evaluated`
- **Invariant:** `landing_request_received = TRUE  $\Rightarrow$  evaluated = TRUE`
- **Event:** `Evaluate_Request`
`WHEN landing_request_received = TRUE`
`THEN evaluated := TRUE`

This pattern ensures that each operationalised requirement is directly represented by an executable event, with its expected behavior formally constrained and verifiable in Event-B.

B. Structure of the ATC Event-B Model

We implemented a complete Event-B machine named `ATC_Machine`, covering all behavioral specifications P1–P9 explained in Section II. Due to space limitations, the full specification of the Event-B model is provided only in the extended version¹. The model contains:

¹<https://drive.google.com/file/d/1Hw-OQEPUYKkIQHJd3e93y6k14nUEfv9e/view?usp=sharing>

- 20 Boolean variables representing the state of actors and system tasks
- 12 invariants derived from LTL formulas, capturing safety and behavioral constraints
- 17 events corresponding to system activities such as request handling, data validation, landing clearance, and pilot confirmation

Each event is modeled in a guarded action style, ensuring that if a certain condition holds, the expected behavior will eventually occur. Some events also involve timing constraints (e.g., processing within 60 seconds), which are captured qualitatively using a Boolean variable `process_completed_within_60s` rather than modeling actual time.

VI. EXPERIMENTAL EVALUATION

This section presents the tool environment used to support the proposed transformation approach and reports the experimental results obtained from verifying the generated Event-B models.

A. Tool Support

The proposed approach was implemented and validated within the Event-B ecosystem, using two main tools: *Rodin* and *ProB*. Rodin supports the development of Event-B models by automatically generating and checking *proof obligations* (POs) to ensure structural correctness and invariant preservation. ProB complements this by performing dynamic model checking, detecting invariant violations, deadlocks, and verifying temporal properties.

The Event-B model generated from the extracted temporal obligations consists of *Boolean state variables*, *invariants* corresponding to the LTL formulas, and a set of *events* that simulate the system's behavior.

B. Experimental Results

When verified in Rodin, a total of 38 *invariant proof obligations* (INV) were generated. Some invariants failed to be automatically discharged, such as:

- `Start_Processing/inv32/INV` – corresponds to the `Start_Processing` event in the Event-B model, which is derived from **P3**
- `Receive_Request/inv27/INV` – corresponds to the `Receive_Request` event, mapped from **P1** and **P2**
- `Mark_On_Time_Valid/inv23/INV` – corresponds to the `Mark_On_Time_Valid` event, associated with **P5**

These failures indicate potential *inconsistencies* between the intended state-preservation conditions and the modeled event logic.

Subsequently, ProB was applied for dynamic model checking with the options *Find Deadlocks* and *Find Invariant Violations* enabled. The results showed:

- No deadlocks or event-related errors detected.

- An invariant violation involving the formula:

$$landing\_request\_received \Rightarrow evaluated$$

Specifically, the state `landing_request_received = TRUE` occurred while `evaluated = FALSE`.

This finding aligns with the primary goal of the proposed approach—to identify inconsistencies in operationalised requirements—allowing the development team to revisit the original iStar model or the Structured English Specification (SES) to make appropriate corrections.

The screenshot shows a table with columns 'Name', 'Value', and 'Previous value'. It lists various system variables and their current and previous states. The 'landing_request_received' variable is highlighted in green, indicating it is TRUE, while 'evaluated' is highlighted in red, indicating it is FALSE. This combination of values represents an invariant violation.

Name	Value	Previous value
ATC Machine		
about	FALSE	FALSE
clearance_requested	FALSE	FALSE
confirmation_received	FALSE	FALSE
decision_recorded	FALSE	FALSE
decision_returned	FALSE	FALSE
evaluated	FALSE	FALSE
landing_request_received	TRUE	TRUE
on_time	FALSE	FALSE
pilot_request	TRUE	FALSE
process_completed	FALSE	FALSE
processing_started	FALSE	FALSE
radar_analyzed	FALSE	FALSE
radar_provided	FALSE	FALSE
radar_valid	FALSE	FALSE
request_processed	FALSE	FALSE
request_received	FALSE	FALSE
runway_analyzed	FALSE	FALSE
runway_provided	FALSE	FALSE
runway_valid	FALSE	FALSE
valid	FALSE	FALSE
formulas		
variables		
constants		
invariants		
event guards		
Invariants satisfied		No event errors detected

Fig. 4. ProB model checking results showing invariant violation.

VII. RELATED WORK

This work contributes to bridging goal-oriented models and formal methods by transforming iStar into verifiable Event-B specifications via LTL. Related studies are grouped into: (1) using iStar for formal modeling, (2) formal semantics of iStar, and (3) operationalizing LTL in Event-B.

iStar as a formal foundation. Several works have explored translating iStar into formal languages. Zambon et al. [4] mapped iStar SR models to Alloy for consistency checking and constraint validation. Sebastiani and Zini [3] used Promela and SPIN for verifying temporal properties. Cabot and Teniente [18] integrated iStar with OCL constraints in the USE environment to assess structural consistency. These approaches link iStar with formal tools but mostly focus on static or propositional logic, lacking temporal expressiveness and traceability to executable models.

Formal semantics of iStar. Although widely adopted, iStar initially lacked a formal semantic foundation. Yu's seminal work introduced strategic rationale modeling, later refined in the iStar 2.0 metamodel [1]. Horkoff et al. [15] proposed a lifecycle roadmap for goal models. Dalpiaz et al. [19] formalized refinement and contribution links using rule-based logic, supporting goal propagation and conflict detection. These efforts improve semantic

clarity and reasoning capability but do not directly enable verified or executable models.

LTL operationalisation in Event-B. Machado et al. [17] introduced systematic patterns to encode LTL obligations—such as eventuality, bounded response, and urgency—into Event-B using guards, clocks, and refinement. This facilitates temporal verification in Event-B. However, the LTL formulas are crafted independently, without grounding in a requirements-level model like iStar, limiting traceability.

Complementary directions. Souza et al. [20] proposed *awareness requirements* and *live goals* for runtime adaptation. While relevant to dynamic systems, their work emphasizes flexibility over formal verifiability.

In summary, existing research offers valuable contributions in goal modeling, temporal reasoning, and formal verification. However, a unified and tool-supported pipeline from high-level iStar models, through temporal obligation modeling, to verified Event-B specifications remains underexplored. This paper aims to close that gap by integrating the intuitive modeling of iStar, the semantic rigor of LTL, and the verification strength of Event-B and ProB.

REFERENCES

- [1] F. Dalpiaz, X. Franch, J. Horkoff, iStar 2.0 language guide, arXiv preprint arXiv:1605.07767 (2016).
- [2] L. Kadakolmath, U. D. Ramu, iStar goal model to Z formal model translation and model checking of CBTC moving block interlocking system, *Formal Aspects of Computing* 36 (1) (2024) 1–45.
- [3] R. Ali, F. Dalpiaz, P. Giorgini, A goal-based framework for contextual requirements modeling and analysis, *Requirements Engineering* 15 (4) (2010) 439–458. doi:10.1007/s00766-010-0110-z.
- [4] E. Zambon, J. C. do Prado Leite, J. Horkoff, From i* models to alloy: Towards formal verification of goal-oriented requirements models, *Journal of Systems and Software* 104 (2015) 30–43.
- [5] C. Almendra, C. Silva, V. Souza, R. Guizzardi, C. Bernabe, iStar2.0-owl: an operational ontology for iStar, in: *Proc. of the 12th International i* Workshop (iStar 2019)*, co-located with ER 2019, Vol. 2490 of *CEUR Workshop Proceedings*, 2019, pp. 1–7.
- [6] C. Djeddi, N. E. Zarour, P. Charrel, Formal verification of the extension of iStar to support big data projects, *Computer Science* 22 (3) (2021) 4035. doi:10.7494/csci.2021.22.3.4035.
- [7] R. Morgan, S. Pulawski, M. Selway, W. Mayer, G. Grossmann, M. Stumptner, A. Ghose, R. Kyprianou, Modeling rates of change and aggregations in runtime goal models, in: *Conceptual Modeling, 2022*, pp. 397–412.
- [8] M. Daun, J. Brings, L. Krajinski, V. Stenkova, T. Bandyszak, A GRL-compliant iStar extension for collaborative cyber-physical systems, *Requirements Engineering* 26 (3) (2021) 325–370.
- [9] A. A. Anda, D. Amyot, Self-adaptation driven by SysML and goal models – a literature review, *e-Informatica Software Engineering Journal* 16 (1) (2022) 220101.
- [10] F. Ferrarotti, P. Riviere, K. Schewe, N. K. Singh, Y. A. Ameer, A complete fragment of LTL(EB), arXiv preprint arXiv:2401.16838 (2024).
- [11] M. Broy, B. Rumpe, Development use cases for semantics-driven modeling languages, *Communications of the ACM* 66 (5) (2023) 62–71.
- [12] J. Abrial, *Modeling in Event-B: System and Software Engineering*, Cambridge University Press, 2010.
- [13] M. Leuschel, M. Butler, ProB: A model checker for B, *International Journal on Software Tools for Technology Transfer (STTT)* 10 (2) (2008) 185–203.
- [14] A. Guillaume, Eurocontrol specification for airport collaborative decision making (a-cdm), Tech. rep., EUROCONTROL (2025).
- [15] J. Horkoff, T. Li, F. Li, M. Salnitri, E. F. B. Cardoso, P. Giorgini, J. Mylopoulos, J. C. S. do Prado Leite, J. Castro, L. Liu, E. S. K. Yu, Taking goal models downstream: A systematic roadmap, in: *2014 IEEE Int. Conf. on Research Challenges in Information Science (RCIS)*, IEEE, 2014, pp. 1–12.
- [16] D. Giannakopoulou, T. Pressburger, A. Mavridou, J. Schumann, Generation of formal requirements from structured natural language, in: *From Requirements to Software*, 2020, pp. 19–35.
- [17] R. J. Machado, A. Mota, J. C. Campos, From goal-oriented requirements to event-b specifications, in: *Proc. of the Int. Conf. on Software Engineering and Formal Methods (SEFM)*, 2015, pp. 97–113.
- [18] J. Cabot, E. Teniente, Integrating OCL and i* with USE, in: *Proc. of the Int. Conf. on Advanced Information Systems Engineering (CAiSE)*, 2008, pp. 204–218.
- [19] F. Dalpiaz, P. Giorgini, J. Mylopoulos, Tropos 2.0: Specification, analysis, and planning of multi-agent systems, *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 8 (1) (2013) 1–33.
- [20] V. Souza, A. Lapouchnian, J. Mylopoulos, Awareness requirements for adaptive systems, in: *Proc. of the 6th Int. Symp. on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, 2011, pp. 60–69. doi:10.1145/1988008.1988018.

SAT-Based Approaches for Two-Dimensional Bin Packing: A Comprehensive Comparison with MIP and CP Methods

Tuyen Van Kieu

VNU University of Engineering and Technology
Hanoi, Vietnam
tuyenkv@vnu.edu.vn

Chi Linh Hoang

VNU University of Engineering and Technology
Hanoi, Vietnam
22028132@vnu.edu.vn

Khanh Van To

VNU University of Engineering and Technology
Hanoi, Vietnam
khanhtv@vnu.edu.vn

Abstract—The Two-Dimensional Bin Packing Problem (2D-BPP) is an NP-hard problem where traditional methods like Mixed Integer Programming (MIP) and Constraint Programming (CP) can be slow. This paper introduces a comprehensive SAT-based framework for 2D-BPP, featuring dual non-stacking and stacking strategies with rotation and symmetry-breaking constraints. We evaluated three solving paradigms: binary search with a standard SAT solver, incremental SAT, and a direct MaxSAT approach. Extensive experiments show that our methods are highly competitive, often outperforming leading commercial solvers—including Google OR-Tools, IBM CPLEX, and Gurobi—on their respective CP and MIP models, establishing SAT as a powerful alternative for this class of geometric optimization problems.

Index Terms—Two-Dimensional Bin Packing, SAT encoding, MaxSAT

I. INTRODUCTION

The Two-Dimensional Bin Packing Problem (2D-BPP) is a cornerstone of combinatorial optimization, challenging researchers for decades due to its NP-hard nature [1] and its wide-ranging industrial relevance in fields like logistics, manufacturing, and resource allocation [2], [3]. The objective is to pack a given set of rectangular items into the minimum number of identical bins, a task that becomes computationally intractable for large-scale instances.

Historically, exact methods for 2D-BPP have evolved significantly. Early approaches were dominated by sophisticated branch-and-bound algorithms [4], [5]. Subsequently, Mixed Integer Programming (MIP) became a standard tool, though formulations based on grid coordinates or relative item positioning often suffer from an explosion in the number of variables and weak linear relaxations, limiting their scalability [2], [6]. A major breakthrough came with set covering models solved via branch-and-price frameworks, rooted in the seminal work of Gilmore and Gomory [7], where the column generation subproblem (finding a valid packing pattern) is itself a difficult task. A highly successful hybrid

technique uses Constraint Programming (CP) to solve this pricing subproblem, leveraging CP's strength in handling geometric constraints [6]. Standalone CP models, using global constraints like `diffn`, are also effective but can be sensitive to modeling choices and search heuristics [8]. Alongside these exact methods, numerous heuristic and metaheuristic approaches have been proposed to find high-quality solutions quickly, though without guarantees of optimality [9]. Recent work continues this trajectory, with learning-based methods such as deep reinforcement learning demonstrating significant improvements over classical heuristics [10].

In parallel, the field of Boolean Satisfiability (SAT) has witnessed revolutionary advances [11]. Modern SAT solvers, powered by techniques like conflict-driven clause learning (CDCL), efficient restart policies, and incremental solving capabilities, have become formidable tools for complex discrete optimization [12]. The extension to Maximum Satisfiability (MaxSAT) further allows for direct optimization of objective functions [13]. However, applying SAT to complex geometric problems like 2D-BPP is not straightforward. While SAT has shown promise in related areas like strip packing [14], adapting these techniques to bin packing presents significant technical hurdles. The challenge lies in efficiently encoding the dual objective of placing items without overlap while simultaneously minimizing the number of bins used. This represents a more complex constraint interaction than what is encountered in single-bin problems. A naive encoding can easily lead to an intractable number of variables and clauses, negating the power of modern solvers.

This paper addresses this gap by developing and evaluating a complete SAT-based framework for 2D-BPP. Our primary contributions are:

- 1) We propose two innovative encoding frameworks. The first is a *non-stacking* model that provides a direct representation of item placement. The second is a *stacking* model that introduces a key technical innovation: it

strategically decomposes the bin packing problem into a sequence of virtual strip packing problems, a non-trivial adaptation of prior art that is purpose-built for BPP.

- 2) We develop and integrate crucial extensions for handling 90-degree item rotations and adapt powerful symmetry-breaking (SB) constraints [14]. We detail the technical adaptations required to make these constraints effective within our BPP-specific models, a critical step for managing the combinatorial explosion in the search space.
- 3) We design and evaluate three distinct solving paradigms built upon our encodings: a classic binary search, an incremental SAT approach that leverages learned clauses, and a direct optimization method using a state-of-the-art MaxSAT solver.
- 4) To validate the performance of our framework, we conduct a large-scale empirical study comparing our methods against leading commercial solvers (Google OR-Tools, IBM CPLEX, and Gurobi).

The remainder of this paper is organized as follows. Section II details the problem formulation. Section III presents our SAT encoding frameworks. Section IV describes the solving paradigms. Section V provides a comprehensive experimental evaluation, and Section VI concludes the paper.

II. PROBLEM FORMULATION

A. Mathematical Definition

Let $I = \{1, 2, \dots, n\}$ be a set of rectangular items, where each item $i \in I$ is characterized by its original width $w_i > 0$ and height $h_i > 0$. Let $B = \{1, 2, \dots, k\}$ be a set of identical rectangular bins, each with width $W > 0$ and height $H > 0$. All bin and item dimensions (W, H, w_i, h_i) , as well as coordinates, are assumed to be positive integers. We consider two variants: one with fixed item orientation and one where items can be rotated by 90 degrees.

For each item $i \in I$, we define decision variables:

- (x_i, y_i) : The integer coordinates of the bottom-left corner.
- b_i : The bin assignment.
- $rot_i \in \{0, 1\}$: $rot_i = 1$ if item i is rotated and 0 otherwise (including the non-rotational variant).

The effective dimensions (w'_i, h'_i) of item i are determined by its rotation status:

$$(w'_i, h'_i) = \begin{cases} (w_i, h_i) & \text{if } rot_i = 0 \\ (h_i, w_i) & \text{if } rot_i = 1 \end{cases} \quad (1)$$

The optimization problem is to find the minimum number of bins k^* such that:

$$\text{minimize } k \quad (2)$$

$$\text{subject to } b_i \in \{1, \dots, k\}, \forall i \in I \quad (3)$$

$$x_i + w'_i \leq W, \quad y_i + h'_i \leq H, \forall i \in I \quad (4)$$

$$\forall i, j \in I, i < j, \text{ if } b_i = b_j \text{ then}$$

$$(x_i + w'_i \leq x_j) \vee (x_j + w'_j \leq x_i) \vee$$

$$(y_i + h'_i \leq y_j) \vee (y_j + h'_j \leq y_i) \quad (5)$$

B. Bounds for Search Space Pruning

To accelerate the search for an optimal solution, we establish a lower bound (LB) and an upper bound (UB) on the number of bins required. These bounds are crucial for effectively pruning the search space of the solving paradigms discussed in Section IV.

1) *Lower Bound*: A straightforward LB on the number of bins can be derived from the total area of all items [15]. It is the sum of the areas of all items divided by the area of a single bin, rounded up to the nearest integer. The formula is:

$$LB = \left\lceil \frac{\sum_{i \in I} (w_i \cdot h_i)}{W \cdot H} \right\rceil \quad (6)$$

where w_i and h_i are the width and height of item i , and W and H are the dimensions of the bin.

2) *Upper Bound*: For an initial UB , we implement a simple and fast heuristic: the First-Fit (FF) algorithm [16]. The items are considered one by one and each is placed into the first bin where it can fit. If an item cannot be placed in any existing bin, a new bin is created for it. The total number of bins used by this heuristic provides a valid UB for the optimal solution.

III. SAT ENCODING FRAMEWORK

A. Stacking Strategy Encoding

Our first approach conceptually stacks bins vertically to form a strip of dimensions $W \times (k \cdot H)$, reducing the problem to an instance of the Two-Dimensional Strip Packing Problem (see Figure 1). This strategy follows the order encoding approach from strip packing literature and transforms the bin packing objective from minimizing the number of bins to minimizing the total height of the strip.

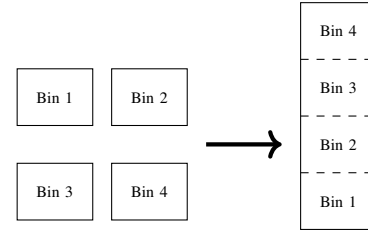


Fig. 1. Illustration of the stacking strategy. Multiple logical bins are stacked into a single, larger bin to represent the problem structure.

We employ a position-based encoding strategy where item positions are discretized to a finite grid of potential placement coordinates. The coordinates are assumed to be non-negative integers, and items are placed on integer grid positions. The key insight of the stacking strategy is that minimizing the number of bins k in the original 2D-BPP is equivalent to minimizing the height $H_{total} = k \cdot H$ in the transformed strip packing formulation. This allows us to leverage efficient strip packing encodings while solving the bin packing objective.

Variable Design for Stacking:

Following the order encoding methodology, we use:

- *Position variables*: $px_{i,e}$ (true if item i has x -coordinate $\leq e$), $py_{i,f}$ (true if item i has y -coordinate $\leq f$)

- *Relative position variables:* $lr_{i,j}$ (true if item i is left of item j), $ud_{i,j}$ (true if item i is below item j)

Order Encoding Axioms:

$$\neg px_{i,e} \vee px_{i,e+1}, \forall i \in I, \forall e \in \{0, \dots, W-1\} \quad (7)$$

$$\neg py_{i,f} \vee py_{i,f+1}, \forall i \in I, \forall f \in \{0, \dots, k \cdot H-1\} \quad (8)$$

Non-Overlap Constraint: For each pair of items (i, j) :

$$lr_{i,j} \vee lr_{j,i} \vee ud_{i,j} \vee ud_{j,i}, \forall i \neq j \quad (9)$$

Position-Relation Consistency: These constraints link relative positions to absolute coordinates. For the rotational variant, all instances of w_i and h_i are replaced by their effective dimensions w'_i and h'_i , which are implemented using clauses conditioned on the rot_i variables.

$$\neg lr_{i,j} \vee \neg px_{j,e}, \forall e \in \{0, \dots, w'_i-1\} \quad (10)$$

$$\neg lr_{i,j} \vee px_{i,e} \vee \neg px_{j,e+w'_i}, \forall e \in \{0, \dots, W-w'_i-1\} \quad (11)$$

$$\neg ud_{i,j} \vee \neg py_{j,f}, \forall f \in \{0, \dots, h'_i-1\} \quad (12)$$

$$\neg ud_{i,j} \vee py_{i,f} \vee \neg py_{j,f+h'_i}, \forall f \in \{0, \dots, k \cdot H-h'_i-1\} \quad (13)$$

Domain Constraint: Items must fit within the strip:

$$px_{i,W-w'_i}, \forall i \in I \quad (14)$$

$$py_{i,k \cdot H-h'_i}, \forall i \in I \quad (15)$$

Bin Separation Constraint: This ensures that items are not split across bin boundaries.

$$\neg py_{i,j \cdot H-1} \vee py_{i,j \cdot H-h'_i}, \forall j \in \{1, \dots, k-1\} \quad (16)$$

$$\neg py_{i,j \cdot H-h'_i} \vee py_{i,j \cdot H-1}, \forall j \in \{1, \dots, k-1\} \quad (17)$$

B. Non-Stacking Strategy Encoding

The non-stacking approach maintains explicit bin assignment variables and enforces constraints within each bin independently. This strategy treats each bin as a separate orthogonal packing problem and is the primary encoding used in our implementation.

Variable Design for Non-Stacking:

- *Bin assignment variables:* $x_{i,j}$ (true if item i is assigned to bin j)
- *Position variables:* $px_{i,e}$ (true if item i has x -coordinate $\leq e$), $py_{i,f}$ (true if item i has y -coordinate $\leq f$)
- *Relative position variables:* $lr_{i,j}$ (true if item i is left of item j), $ud_{i,j}$ (true if item i is below item j)
- *Bin usage variables:* b_j (true if bin j is used)

Assignment Constraints: Each item must be assigned to exactly one bin.

$$\bigvee_{j=1}^k x_{i,j}, \forall i \in I \quad (18)$$

$$\neg x_{i,j_1} \vee \neg x_{i,j_2}, \forall i, j_1 \neq j_2 \quad (19)$$

Order Encoding Axioms: The position variables must be consistent.

$$\neg px_{i,e} \vee px_{i,e+1}, \forall i \in I, \forall e \in \{0, \dots, W-1\} \quad (20)$$

$$\neg py_{i,f} \vee py_{i,f+1}, \forall i \in I, \forall f \in \{0, \dots, H-1\} \quad (21)$$

Non-Overlap Constraint: For any two items i_1, i_2 in the same bin j , their geometric relationship is enforced.

$$\neg x_{i_1,j} \vee \neg x_{i_2,j} \vee lr_{i_1,i_2} \vee lr_{i_2,i_1} \vee ud_{i_1,i_2} \vee ud_{i_2,i_1}, \quad (22)$$

$$\forall i_1 < i_2, \forall j$$

Position-Relation Consistency: These constraints link the relative and absolute position variables.

$$\neg x_{i_1,j} \vee \neg x_{i_2,j} \vee \neg lr_{i_1,i_2} \vee \neg px_{i_2,e}, \forall e \in \{0, \dots, w'_{i_1}-1\} \quad (23)$$

$$\neg x_{i_1,j} \vee \neg x_{i_2,j} \vee \neg ud_{i_1,i_2} \vee \neg py_{i_2,f}, \forall f \in \{0, \dots, h'_{i_1}-1\} \quad (24)$$

$$\neg x_{i_1,j} \vee \neg x_{i_2,j} \vee \neg lr_{i_1,i_2} \vee px_{i_1,e} \vee \neg px_{i_2,e+w'_{i_1}}, \quad (25)$$

$$\forall e \in \{0, \dots, W-w'_{i_1}-1\}$$

$$\neg x_{i_1,j} \vee \neg x_{i_2,j} \vee \neg ud_{i_1,i_2} \vee py_{i_1,f} \vee \neg py_{i_2,f+h'_{i_1}}, \quad (26)$$

$$\forall f \in \{0, \dots, H-h'_{i_1}-1\}$$

Domain and Bin Usage Constraints:

$$\neg x_{i,j} \vee px_{i,W-w'_i}, \forall i \in I, j \in \{1, \dots, k\} \quad (27)$$

$$\neg x_{i,j} \vee py_{i,H-h'_i}, \forall i \in I, j \in \{1, \dots, k\} \quad (28)$$

$$\neg x_{i,j} \vee b_j, \forall i \in I, j \in \{1, \dots, k\} \quad (29)$$

$$\neg b_{j+1} \vee b_j, \forall j \in \{1, \dots, k-1\} \quad (30)$$

C. Symmetry Breaking Constraints

To prune the search space, we adapt symmetry-breaking constraints from the strip packing literature [14]. These are applied to both models. For the non-stacking model, they are conditioned on pairs of items being in the same bin.

Large Item Constraints: If two items i, j cannot be placed adjacent along an axis, the corresponding relative position is forbidden. For the rotational variant, this check is performed for all four orientation combinations of the pair.

$$\text{if } w'_i + w'_j > W \implies \neg lr_{i,j} \wedge \neg lr_{j,i} \quad (31)$$

$$\text{if } h'_i + h'_j > H \implies \neg ud_{i,j} \wedge \neg ud_{j,i} \quad (32)$$

Note that for the stacking model, the vertical check uses the total strip height $k \cdot H$.

Same-Sized Item Ordering: For identical items i, j where $i < j$, we enforce a lexicographical ordering to eliminate redundant placements.

$$\text{if } (w_i, h_i) = (w_j, h_j) \text{ and } i < j \implies \neg lr_{j,i} \quad (33)$$

Rotation-Specific Symmetry Breaking: For square items, rotation is disabled as it produces an identical configuration.

$$\text{if } w_i = h_i \implies \neg rot_i \quad (34)$$

IV. SOLVING PARADIGMS

A. Non-Incremental SAT with Binary Search

Our baseline approach, detailed in Algorithm 1, employs a binary search over the number of bins k with the given LB and UB . We repeatedly solve decision problems of the form: "Can all items be packed into k bins?"

Algorithm 1 Binary Search for Minimum Bins

```

1: Initialize lower bound  $LB \leftarrow \lceil (\sum_{i \in I} w_i \cdot h_i) / (W \cdot H) \rceil$ 
2: Initialize upper bound  $UB$  from a fast heuristic.
3: while  $LB < UB$  do
4:    $k \leftarrow \lfloor (LB + UB) / 2 \rfloor$ 
5:    $\phi \leftarrow \text{GenerateSATEncoding}(I, W, H, k)$ 
6:   if  $\text{SolveSAT}(\phi) = \text{SATISFIABLE}$  then
7:      $UB \leftarrow k$   $\triangleright$  A solution with  $k$  bins was found
8:   else
9:      $LB \leftarrow k + 1$   $\triangleright k$  bins are not enough
10:  end if
11: end while
12: return  $LB$   $\triangleright$  The optimal number of bins

```

B. Incremental SAT Solving

To accelerate the search, we enhance the binary search framework with an incremental SAT solver. Instead of rebuilding the SAT formula from scratch for each k , we reuse learned clauses. When the solver proves k bins are insufficient (UNSAT), the learned conflict clauses that represent fundamental geometric impossibilities are retained. For the next iteration with $k + 1$ bins, we only add clauses for the newly available space, allowing the solver to build upon prior knowledge and solve subsequent problems faster, as formalized in Algorithm 2.

Algorithm 2 Incremental SAT Solving Framework

```

1: Initialize solver  $S$  with base constraints  $\phi_{base}$ .
2: Initialize  $LB, UB$  as in Algorithm 1.
3: while  $LB < UB$  do
4:    $k \leftarrow \lfloor (LB + UB) / 2 \rfloor$ 
5:   Let  $A = \{-b_{k+1}\}$  be the assumption for this iteration.
6:    $\triangleright$  Test if  $k$  bins are sufficient
7:   if  $S.\text{solve}(A) = \text{SATISFIABLE}$  then
8:      $UB \leftarrow k$   $\triangleright$  Solver state and clauses are kept
9:   else
10:     $LB \leftarrow k + 1$ 
11:  end if
12: end while
13: return  $LB$ 

```

C. Incremental MaxSAT Optimization

As an alternative to decision-based search, we formulate the 2D-BPP as a direct optimization problem using a MaxSAT solver. The geometric and assignment rules become hard clauses that must be satisfied. The bin usage variables (b_j) are treated as soft clauses, and the solver's objective is to find a valid packing that minimizes the number of satisfied soft clauses (i.e., used bins).

Hard Constraints: All geometric and assignment constraints from the encoding section.

Soft Constraints: Bin usage minimization:

$$\text{Soft: } \neg b_j, \forall j \in \{LB, \dots, UB\}, \text{ weight} = 1 \quad (35)$$

TABLE I
COMPARATIVE ANALYSIS OF SAT/MAXSAT STRATEGIES FOR
NON-ROTATIONAL 2D-BPP.

Configuration	#Opt	Opt Rate (%)	$\sum \tau$ (s)	#Vars ($\times 10^6$)	#Cls ($\times 10^9$)	Gap (%)
<i>Non-Incremental SAT with Binary Search</i>						
BPP	286	91.67	26212.38	8.88	3.20	0.56
BPP_C1	292	93.59	25672.54	8.88	2.46	0.42
BPP_S	212	67.95	21968.41	48.02	3.54	0.63
BPP_S_C1	235	75.32	18902.99	48.08	3.48	0.66
<i>Incremental SAT Solving</i>						
BPP_INC_C1	246	78.85	19372.02	8.98	2.93	0.51
BPP_INC_S_C1	249	79.81	10902.70	53.36	3.95	0.89
<i>Direct MaxSAT Optimization</i>						
BPP_MS_C1	280	89.74	40685.03	9.04	2.98	1.02
BPP_MS_S_C1	276	88.46	40219.68	55.46	4.04	1.66

S: Stacking; C1: Symmetry Breaking.

#Opt: Number of optimal solutions found; Opt Rate: Percentage of optimal solutions; $\sum \tau$: Total runtime (s); Gap: Average optimality gap.

Objective: Minimize the number of used bins by minimizing the total weight of violated soft constraints.

V. EXPERIMENTAL EVALUATION**A. Setup and Benchmarks**

Our experimental evaluation employs a comprehensive collection of 510 benchmark instances aggregated from established datasets, including BENG [17] and the CLASS benchmark [5], [18]. This suite covers diverse geometric configurations and packing densities. For metrics requiring known optimal solutions, such as optimality rate, our analysis is based on the set of instances for which an optimal solution was found by at least one of the evaluated configurations. This comprises 312 instances for the non-rotational variant and 290 for the rotational variant. All experiments were conducted on Google Cloud Platform e2-highmem-8 instances (8 vCPUs, 64 GB RAM) with a uniform time limit of 900 seconds per instance, covering both model generation and solving. For our SAT-based methods, we used Glucose 4.2 [12] for SAT and TT-Open-WBO-Inc [13] for MaxSAT. We benchmarked these against a suite of leading solvers: Google OR-Tools 9.1 [19] (CP-SAT and MIP), IBM CPLEX 22.1.1 [20] (CP and MIP), and Gurobi Optimizer 10.0 [21] (MIP), representing the state of the art in commercial optimization solvers. For all solvers, we used their default parameter settings. No problem-specific tuning was performed, ensuring a fair comparison of their out-of-the-box performance.

B. Performance of SAT/MaxSAT Strategies

We first evaluate the performance of different SAT/MaxSAT strategies for the 2D-BPP. Tables I, II present comprehensive results for non-rotational and rotational variants, respectively.

For non-rotational instances, the standard binary search approach with the non-stacking encoding (BPP_C1) achieves the highest optimality rate (93.59%) and exhibits a very low average gap of 0.42%. This significantly outperforms all other methods, including state-of-the-art CP and MIP solvers. Contrary to theoretical expectations that stacking encodings might

TABLE II
IMPACT OF 90-DEGREE ROTATION ON SAT/MAXSAT SOLVER
PERFORMANCE.

Configuration	#Opt	Opt Rate (%)	$\sum \tau$ (s)	#Vars ($\times 10^6$)	#Cls ($\times 10^9$)	Gap (%)
<i>Non-Incremental SAT with Binary Search</i>						
BPP_R	275	94.83	31106.90	10.96	12.84	0.39
BPP_S_R	243	83.79	23542.88	47.54	6.95	0.90
BPP_R_SB	267	92.07	29452.13	10.55	9.77	0.50
BPP_S_R_SB	237	81.72	24213.38	47.48	6.90	1.07
<i>Incremental SAT Solving</i>						
BPP_INC_R_SB	261	90.00	30149.30	11.07	13.15	0.10
BPP_INC_S_R_SB	241	83.10	14279.58	54.16	7.91	1.26
<i>Direct MaxSAT Optimization</i>						
BPP_MS_R_SB	257	88.62	20293.83	10.49	10.83	1.87
BPP_MS_S_R_SB	255	87.93	22405.73	52.69	7.67	2.30

S: Stacking; R: Rotation; SB: Symmetry Breaking.

#Opt: Number of optimal solutions found; Opt Rate: Percentage of optimal solutions; $\sum \tau$: Total runtime (s); Gap: Average optimality gap.

be more compact, the non-stacking approach proves more robust across this diverse benchmark suite. The stacking model (BPP_S_C1), while efficient in runtime for solved instances, achieves a lower optimality rate (75.32%), suggesting potential scalability challenges. The MaxSAT approach (BPP_MS_C1) is competitive, finding 280 optimal solutions (89.74%), but at the cost of higher computational time and a slightly larger gap (1.02%), indicating that the overhead of weighted optimization can impact performance on complex instances.

When item rotation is permitted, our specialized SAT model (BPP_R) demonstrates exceptional performance, achieving a 94.83% optimality rate and the best gap of 0.39%. This performance significantly surpasses all general-purpose solvers. The effectiveness of the integrated rotation and SB constraints is evident, as they allow the solver to navigate a much larger combinatorial space efficiently. Again, the stacking model (BPP_S_R_SB) is less effective, solving fewer instances optimally (81.72%) compared to its non-stacking counterpart. MaxSAT (BPP_MS_R_SB) remains a strong contender, securing 88.62% of optimal solutions, reinforcing its viability as a direct optimization approach, albeit with a trade-off in solution quality for unsolved instances (1.87% gap).

A key and counter-intuitive finding is the superior performance of the non-incremental binary search (BPP_C1) over its incremental counterpart (BPP_INC_C1), which solved fewer instances to optimality (292 vs. 246). Our analysis points to two factors explaining this phenomenon. First, the data in Table I indicates that the incremental approach generates a significantly larger set of clauses on average (e.g., 2.93 billion for BPP_INC_C1 vs. 2.46 billion for BPP_C1). This suggests that the overhead of managing a growing database of learned clauses becomes a substantial computational bottleneck, outweighing the benefits of avoiding a restart. Second, the conflict clauses learned while proving unsatisfiability for k bins may be largely irrelevant to the subsequent search for a solution in $k - 1$ bins. These "stale" clauses can pollute the search, misguiding the solver's heuristics. In contrast, the non-incremental approach benefits from a "clean slate" in each step

TABLE III
PERFORMANCE COMPARISON OF SAT/MAXSAT CONFIGURATIONS AND
GENERAL OPTIMIZATION SOLVERS.

Configuration	#Opt	Opt Rate (%)	$\sum \tau$ (s)	Gap (%)
<i>SAT/MaxSAT (Selected Best)</i>				
BPP_C1	292	93.59	25672.54	0.42
BPP_R	275	94.83	31106.90	0.39
<i>OR-Tools Solvers</i>				
OR-TOOLS_CP_C1	291	93.27	17340.02	0.70
OR-TOOLS_CP_R_SB	257	88.62	19129.24	1.28
OR-TOOLS_MIP_C1	202	64.74	8614.88	9.32
OR-TOOLS_MIP_R_SB	151	52.07	7509.12	11.76
<i>CPLEX Solvers</i>				
CPLEX_CP_C1	192	61.54	86147.59	9.54
CPLEX_CP_R_SB	153	52.76	68990.01	11.34
CPLEX_MIP_C1	211	67.63	19053.85	8.95
CPLEX_MIP_R_SB	174	60.00	14839.29	10.06
<i>Gurobi Solvers</i>				
GUROBI_MIP_C1	230	73.72	12666.57	7.47
GUROBI_MIP_R_SB	181	62.41	13745.75	9.00

CI, SB: Symmetry Breaking; R: Rotation.

#Opt: Number of optimal solutions found; Opt Rate: Percentage of optimal solutions; $\sum \tau$: Total runtime (s); Gap: Average optimality gap.

of the binary search, allowing its heuristics to focus only on the current problem. A deeper investigation into the lifecycle and relevance of learned clauses is a compelling direction for future work.

C. Comparison with General Optimization Solvers

The key finding of this research is the definitive outperformance of specialized SAT-based methods over state-of-the-art general optimization solvers. As shown in Table III, our best SAT configurations (BPP_C1 and BPP_R) solve more instances to optimality and provide higher-quality solutions (lower gap) than the best configurations of Google OR-Tools, IBM CPLEX, and Gurobi. The performance gap is particularly pronounced in the rotational setting. Our BPP_R model solves 18 more instances to optimality than the best general solver, OR-Tools CP_R_SB (275 vs. 257), and reduces the solution gap from 1.28% to a mere 0.39%. This underscores the power of a tightly integrated, domain-specific encoding.

An important secondary observation, confirming a known trend in packing literature [2], is the superiority of CP-based solvers over MIP-based ones for this problem. For instance, OR-Tools' CP-SAT solver achieves a 93.27% optimality rate, while its MIP counterpart only reaches 64.74%. This is largely because global constraints used in CP (like `diffn` [22]) are more powerful for expressing non-overlapping conditions than the large sets of linear inequalities required in MIP, leading to more effective pruning.

D. Discussion

The superior performance of our SAT-based framework can be attributed to several key factors. First, the declarative nature of SAT allows for a direct and powerful representation of the complex geometric and combinatorial constraints of 2D-BPP. Modern CDCL solvers [23] are exceptionally skilled at

navigating the resulting search space. Their core mechanisms, consisting of unit propagation, conflict analysis, and clause learning, appear more effective at pruning infeasible regions for this problem class than the linear relaxations of MIP or the domain-filtering techniques of CP.

Second, our results highlight the critical importance of problem-specific modeling. The non-stacking encoding, despite being potentially larger in terms of variables, consistently provided a more tractable problem structure for the solver. This suggests that the simpler, more direct constraints of the non-stacking model allow the solver’s heuristics to make more effective branching decisions. Our experimental data supports this observation: while the stacking approach requires significantly more variables (47-55 million vs. 9-11 million), both approaches generate comparable clause counts, indicating that the variable overhead rather than constraint complexity may be the limiting factor for the stacking model’s performance.

Finally, the success of our rotation-enabled models underscores the power of integrating domain-specific knowledge, such as adapted symmetry-breaking constraints directly into the encoding. These constraints are crucial for managing the exponential growth of the search space when rotation is permitted, establishing specialized SAT models as a formidable alternative to general-purpose commercial solvers.

VI. CONCLUSION

This paper introduced a comprehensive SAT-based framework that redefines the state of the art for solving the 2D Bin Packing Problem. Our core contribution is the development of specialized encodings, namely the direct *non-stacking* and innovative *stacking* approaches, that transform complex geometric and assignment constraints into structures that modern SAT solvers can exploit efficiently. We demonstrated that adapting techniques from related problems, particularly symmetry-breaking constraints from strip packing, is crucial for achieving superior performance.

The effectiveness of our approach was validated through extensive empirical evaluation on 510 benchmark instances. Our SAT-based framework consistently outperformed state-of-the-art CP and MIP solvers from Google, IBM, and Gurobi. This provides definitive evidence that engineering domain-specific representations for powerful paradigms like SAT is more effective than relying on general-purpose solvers for challenging combinatorial problems. While our focus was on exact methods to guarantee optimality, a comparative study against modern metaheuristics remains a valuable direction for future work to assess the trade-offs between solution quality and runtime.

This research opens promising avenues for future work, including extensions to three-dimensional packing, dynamic scenarios with real-time item arrivals, and portfolio-based solvers that leverage complementary strengths of multiple formulations. The source code and datasets are publicly available at <https://github.com/bin-packing/BPP>.

REFERENCES

- [1] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.
- [2] A. Lodi, S. Martello, and M. Monaci, “Two-dimensional packing problems: A survey,” *European Journal of Operational Research*, vol. 141, no. 2, pp. 241–252, 2002.
- [3] M. Iori, V. L. de Lima, S. Martello, F. K. Miyazawa, and M. Monaci, “Exact solution techniques for two-dimensional cutting and packing,” *European Journal of Operational Research*, vol. 289, no. 2, pp. 399–415, 2021.
- [4] N. Christofides and C. Whitlock, “An algorithm for two-dimensional cutting problems,” *Operations Research*, vol. 25, no. 1, pp. 30–44, 1977.
- [5] S. Martello and D. Vigo, “Exact solution of the two-dimensional finite bin packing problem,” *Management Science*, vol. 44, no. 3, pp. 388–399, 1998.
- [6] D. Pisinger and M. Sigurd, “Using decomposition techniques and constraint programming for solving the two-dimensional bin-packing problem,” *INFORMS Journal on Computing*, vol. 19, no. 1, pp. 36–51, 2007.
- [7] P. C. Gilmore and R. E. Gomory, “A linear programming approach to the cutting-stock problem,” *Operations Research*, vol. 9, no. 6, pp. 849–859, 1961.
- [8] F. Clautiaux, J. Carlier, and A. Moukrim, “A new exact method for the two-dimensional bin-packing problem with fixed orientation,” *Operations Research Letters*, vol. 36, no. 2, pp. 173–179, 2008.
- [9] E. Hopper and B. C. Turton, “An empirical investigation of meta-heuristic and heuristic algorithms for a 2d packing problem,” *European Journal of Operational Research*, vol. 128, no. 1, pp. 34–57, 2001.
- [10] B. Wang, Z. Lin, W. Kong, and H. Dong, “Bin packing optimization via deep reinforcement learning,” *IEEE Robotics and Automation Letters*, 2025.
- [11] A. Biere, M. Heule, H. van Maaren, and T. Walsh, *Handbook of Satisfiability*. IOS Press, 2009, vol. 185.
- [12] G. Audemard and L. Simon, “Glucose: a solver that predicts learnt clauses quality,” *Proceedings of the International Conference on Theory and Applications of Satisfiability Testing (SAT)*, pp. 7–8, 2009.
- [13] A. Nadel, “TT-Open-WBO-Inc: an efficient anytime MaxSAT solver,” *Journal of Satisfiability, Boolean Modeling and Computation*, vol. 15, no. 1, pp. 1–7, 2024.
- [14] T. Soh, K. Inoue, N. Tamura, M. Banbara, and H. Nabeshima, “A sat-based method for solving the two-dimensional strip packing problem,” in *Fundamenta Informaticae*, vol. 102, no. 3-4. IOS Press, 2010, pp. 467–487.
- [15] M. Dell’Amico, S. Martello, and D. Vigo, “A lower bound for the non-oriented two-dimensional bin-packing problem,” *Discrete Applied Mathematics*, vol. 118, no. 1-2, pp. 13–24, 2002.
- [16] E. G. Coffman Jr, M. R. Garey, and D. S. Johnson, “Approximation algorithms for bin packing: a survey,” *Approximation algorithms for NP-hard problems*, pp. 46–93, 1980.
- [17] B. E. Bengtsson, “Packing rectangular pieces—a heuristic approach,” *The Computer Journal*, vol. 25, no. 3, pp. 353–357, 1982.
- [18] J. O. Berkey and P. Y. Wang, “Two-dimensional finite bin packing algorithms,” *Journal of the Operational Research Society*, vol. 38, no. 5, pp. 423–429, 1987.
- [19] Google Inc., “OR-Tools: Operations Research Tools,” <https://developers.google.com/optimization>, Google, 2023, software library.
- [20] IBM Corporation, *IBM ILOG CPLEX Optimization Studio: CPLEX User’s Manual*, IBM Corporation, 2023, version 22.1.
- [21] Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, Gurobi Optimization, LLC, 2023, version 10.0.
- [22] N. Beldiceanu and E. Contejean, “Introducing global constraints in chip,” *Mathematical and Computer Modelling*, vol. 20, no. 12, pp. 97–123, 1994.
- [23] M. W. Moskewicz, C. F. Madigan, Y. Zhao, L. Zhang, and S. Malik, “Chaff: Engineering an efficient sat solver,” in *Proceedings of the 38th Design Automation Conference (DAC)*, 2001, pp. 530–535.

An Efficient Method for Mining Closed Frequent High Average-Utility Occupancy Patterns

Thang La
Faculty of Information
Technology, University of
Science, Ho Chi Minh City
Vietnam National University,
Ho Chi Minh City, Vietnam
Faculty of Information
Technology, Dalat University,
Vietnam
thanglq@dlu.edu.vn

Khanh Tran
Faculty of Information
Technology, Dalat University,
Vietnam
khanhtnn@dlu.edu.vn
Anh Tran
FPT University, Danang,
Vietnam
anhtn35@fe.edu.vn

Hai Duong*
Department of Mathematics
and Computer Science, Dalat
University, Vietnam
haidv@dlu.edu.vn

Bac Le
Faculty of Information
Technology, University of
Science, Ho Chi Minh City
Vietnam National University,
Ho Chi Minh City, Vietnam
lhbac@fit.hcmus.edu.vn

Abstract—Discovering all Frequent High Average-Utility Occupancy Itemsets (FHAUOIs) in quantitative transaction databases has emerged as a promising direction in pattern mining. By integrating both average-utility occupancy and frequency, these itemsets are better aligned with real-world applications, as they not only represent the preferences of the majority of users but also contribute significantly to the total utility of input transactions in which they appear. However, the set of exploited FHAUOIs can be extremely large, especially when dealing with dense or large-scale datasets, or when low minimum threshold values are used. This can make it difficult for users to interpret and utilize the results effectively. To overcome this challenge, the paper designs a novel algorithm called C-FHAUOIM for mining Closed FHAUOIs (CFHAUOIs). The algorithm leverages a newly designed upper bound on average-utility occupancy to efficiently reduce the database by eliminating low-average-utility occupancy items. In addition, it introduces a novel weak lower bound, along with associated pruning strategies, to effectively discard non-closed FHAUOIs early in the mining process. These theoretical advances are integrated into an efficient algorithm for discovering CFHAUOIs. Experimental evaluations conducted on multiple synthetic and real quantitative databases reveal that the proposed method achieves superior performance compared to the baseline, particularly in execution speed, memory efficiency, and scalability.

Keywords—Closed itemset, high average-utility occupancy itemset, pruning strategy, upper bound, weak upper bound.

I. INTRODUCTION

The problems of mining high utility itemsets (HUIM) [1] and high average-utility itemsets (HAUIM) [2] over quantitative transaction databases (QTDBs) have recently emerged as prominent research areas, driven by their wide-ranging applications across domains such as identifying profitable product combinations, examining shopping behavior on digital commerce systems, exploring biomedical information, and designing integrated marketing approaches [3]. In these problems, each item in a transaction is associated with both a quantity and a quality value, which together determine the utility of that item.

HAUIM is generally considered more advantageous than HUIM because it employs a balanced evaluation metric based on average utility [2], which accounts for both the total utility and the length of the itemset. However, HAUIM over QTDBs faces a fundamental challenge: the anti-monotonicity property is not satisfied by the average-utility measure, a key characteristic commonly used to prune the search space in many pattern mining tasks. To overcome this limitation, researchers have introduced various upper bounds (UBs) that fully or partially preserve anti-monotonic behavior, enabling more effective pruning strategies. Over time, several tight UBs and weak upper bounds (WUBs) have been developed, such as *mrau* [4], *vmaub* [5] and *tvwaub* [6].

Related work. While HAUIM and HAUIM have contributed significantly to utility-based pattern mining, a critical limitation remains: the discovered patterns do not inherently reflect how much utility they contribute within each supporting transaction. In other words, these approaches lack a measure of utility representativeness at the transaction level.

To address this gap, the concept of utility occupancy (UO) was proposed [7]. UO quantifies the proportion of a pattern's utility relative to the total utility of the transaction in which it appears. This metric offers a finer-grained perspective by highlighting a pattern's significance and representativeness within individual transactions, rather than just its global utility or frequency. Based on the definition of UO, researchers introduced frequent high utility occupancy itemset mining (FHUOIM) [8], [9], [10], which has proven effective in various real-world applications [8].

Building on utility occupancy, Kumar et al. [11] introduced frequent high average-utility (AU) occupancy itemset mining (FHAUOIM) by combining AU occupancy with support, offering a more balanced measure than FHUOIM. However, FHAUOIM can produce an overwhelming number of patterns, especially in large, dense QTDBs or with low thresholds, which makes it challenging for users to interpret the results and derive meaningful insights, thus limiting the practical utility of the mining process. To address this, researchers mine concise representations such as closed and generator patterns [12][13], which reduce redundancy while preserving essential information; closed patterns, in particular, capture the most representative high-utility patterns in each equivalence class, groups of patterns having the same transaction coverage. In the context of FHUOIM, prior studies have proposed methods to mine both closed FHUOIs [14] and their generators [15]. While closed FHAUOIs are even more desirable than closed FHUOIs, no work has addressed the Closed FHAUOIM (CFHAUOIM) problem. This paper fills that gap by establishing its theoretical foundations and proposing an efficient mining algorithm.

Contributions. The main contributions of this study can be summarized in three aspects. First, two new bounds are introduced, *lubauo* and *wlbauo*, together with effective pruning mechanisms to substantially narrow the exploration space. Second, an efficient algorithm named C-FHAUOIM is developed to identify closed FHAUOIs based on the proposed theoretical framework. Finally, comprehensive experiments conducted on both real-world and synthetic datasets confirm that the proposed approach consistently surpasses the baseline method regarding execution time, memory efficiency, and scalability.

The structure of this paper is outlined as follows. Section II introduces the essential preliminaries and basic definitions. Section III describes the newly developed bounds, pruning mechanisms, and the C-FHAUOIM algorithm. Section IV reports the experimental setup and discusses the obtained results. Lastly, Section V concludes the study and suggests potential directions for future investigation.

II. PRELIMINARY CONCEPTS AND PROBLEM STATEMENT

In this section, we introduce the essential notations and definitions used for discovering FHAUIs. Let $\mathcal{A} \triangleq \{a_j, j \in J = \{1, 2, \dots, M\}\}$ denote a finite collection of distinct items, and let $A \subseteq \mathcal{A}$ represent an itemset. If A includes k elements, it is termed a k -itemset, that is $k = |A|$. For convenience, we assume that all itemsets are sorted in alphabetical order. Each item a_j has a profit, denoted as $p(a_j)$, representing its relative importance or external utility. The set of all item profits is denoted as $\mathcal{P} = \{p(a_j), j \in J \text{ and } p(a_j) > 0\}$. The notation (a, q) is a q -item, in which $a \in \mathcal{A}$ and q is its internal utility. In addition, a quantitative transaction database (QDB) $\mathcal{D}$ is composed of multiple transactions, $\mathcal{D} \triangleq \{t_i, i \in I \triangleq \{1, 2, \dots, N\}\}$. Each transaction $t_i \triangleq \{(a_j, q_{ij}), j \in J_i\}$ contains a set of q -items and is uniquely identified by its transaction identifier ($TID = i$), where $J_i \subseteq J$. Let $u(a, q) = q \times p(a)$ be the utility of (a, q) , representing the profit generated by selling item a with quantity q in a transaction.

To eliminate redundant utility calculations for q -items (a_j, q_{ij}) across transactions t_i in $\mathcal{D}$, these values are precomputed once and kept in an integrated QDB corresponding to $\mathcal{D}$ and $\mathcal{P}$. This database is denoted as $\mathcal{D}' \triangleq \{t_i \triangleq \{(a_j, q'_{ij}), j \in J_i, i \in I\}$, where $q'_{ij} = q_{ij} \times p(a_j)$ if $(a_j, q_{ij}) \in t_i, \forall i \in I, j \in J_i$.

Definition 1 (Itemset support [14]).

- For any non-empty itemset $A \subseteq \mathcal{A}$, let $\rho(A) \triangleq \{t \in \mathcal{D}' \mid t \supseteq A\}$ denote the collection of transactions containing A , where $t \supseteq A$ indicates that every item a_j in A must be included within a corresponding q -item (a_j, q_{ij}) of t (or t is called a supporting transaction of A). By convention, $\rho(\emptyset) \triangleq \mathcal{D}'$.
- Let $\text{supp}(A) = |\rho(A)|$ denote the support of A , given by the transaction count containing A .
- Let $\mathcal{IS} \triangleq \{A \subseteq \mathcal{A} \mid (A \neq \emptyset) \wedge (\rho(A) \neq \emptyset)\}$ be the collection of all itemsets appearing at least once in the transactions.

Running example. For example, consider a QDB $\mathcal{D}$ consisting of 6 transactions $t_1 = \{(a, 2), (b, 5), (c, 56), (d, 25), (e, 14), (f, 4)\}$, $t_2 = \{(b, 2), (c, 2), (e, 80), (f, 8)\}$, $t_3 = \{(b, 60), (e, 30), (f, 2)\}$, $t_4 = \{(a, 5), (d, 5)\}$, and $t_5 = \{(a, 2), (b, 6), (c, 3), (d, 2), (e, 55), (f, 10)\}$, and the profit vector $\mathcal{P} = (2; 4; 2; 4; 3; 2)$ of 6 items a, b, c, d, e and f . The integrated QDB $\mathcal{D}'$ according to $\mathcal{D}$ and $\mathcal{P}$ in Table 1 will be used as the running example. It is observed that there are 2 transactions containing bc , $\rho(bc) = \{t_1, t_2, t_5\}$, $\text{supp}(bc) = 2$.

Definition 2 (Utility and average utility [2]). Let us consider any transaction $t \in \mathcal{D}'$.

- Let $u(a, t) = q$ be the utility of item a in t , and $u(t) \triangleq$

$\sum_{(a,q) \in t} u(a, t)$ be the utility of t .

- Let $u(A, t) \triangleq \sum_{a \in A} u(a, t)$ and $au(A, t) \triangleq \frac{u(A, t)}{|A|}$ be the utility and average-utility of itemset A in t , respectively, where $|A|$ denotes the length of A .

Table 1. The integrated QDB $\mathcal{D}'$.

<i>TID</i>	<i>Transaction (item, utility)</i>
t_1	$(a, 4), (b, 20), (c, 112), (d, 100), (e, 42), (f, 8)$
t_2	$(b, 8), (c, 4), (e, 240), (f, 16)$
t_3	$(b, 240), (e, 90), (f, 4)$
t_4	$(a, 10), (d, 20)$
t_5	$(a, 4), (b, 24), (c, 6), (d, 8), (e, 165), (f, 20)$

Definition 3 (Utility occupancy [8] and average-utility occupancy [11]). Consider any itemset A .

- The utility occupancy of A in a supporting transaction $t \in \mathcal{D}'$, $A \subseteq t$, is termed as the relative utility of A in t compared to the transaction utility of t , given by $uo(A, t) \triangleq \frac{u(A, t)}{u(t)}$.
- The utility occupancy of itemset A over all supporting transactions $t \in \rho(A)$ in the entire database $\mathcal{D}'$ is defined as $uo(A) \triangleq \frac{1}{\text{supp}(A)} \sum_{t \in \rho(A)} uo(A, t)$.
- The average-utility occupancy of A (in $\mathcal{D}'$) is determined as $auo(A) \triangleq \frac{uo(A)}{|A|}$. If we define $auo(A, t) \triangleq \frac{uo(A, t)}{|A|} = \frac{au(A, t)}{u(t)}$, then $auo(A)$ can be represented in another form as $auo(A) = \frac{1}{\text{supp}(A)} \sum_{t \in \rho(A)} auo(A, t)$.

Obviously, for any $A \in \mathcal{IS}$, since $auo(A, t) \in (0; 1], \forall t \in \rho(A)$, we always have $auo(A) \in (0; 1]$.

Example 1. For example, consider itemset $A = bc$, the utility and average-utility of A in t_1 are respectively $u(A, t_1) = 20 + 112 = 132$ and $au(A, t_1) = \frac{132}{2} = 66$. Since the transaction utility of t_1 is $u(t_1) = 286$, the utility occupancy of A in t_1 is $uo(A, t_1) = \frac{u(A, t_1)}{u(t_1)} = \frac{132}{286} = 0.46$. Thus, the average utility occupancy of A (with length 2) in t_1 , $auo(A, t_1) = \frac{0.46}{2} = 0.23$. Similarly, by considering all transactions containing A , $\rho(A) = \{t_1, t_2, t_5\}$, the AU occupancy of itemset A is determined as $auo(A) = (0.23 + (\frac{12}{268} + \frac{30}{227})/2)/3 = 0.106$.

Definition 4 (Frequent high AU occupancy itemset). Let mo and ms be two thresholds representing the minimum AU occupancy and the minimum support, respectively ($mo > 0$ and $ms > 0$). An itemset A is identified as a high AU occupancy itemset (HAUOI) if its AU occupancy in $\mathcal{D}'$ satisfies $auo(A) \geq mo$; otherwise, when $auo(A) < mo$, A is categorized as a low AU occupancy itemset (LAUOI). Furthermore, an HAUOI A is regarded as a frequent HAUOI (FHAUOI) if its support meets the condition $\text{supp}(A) \geq ms$. The objective of FHAUOI mining (FHAUOIM) is to determine the complete set of itemsets satisfying both criteria, formally defined as $\mathcal{FHAUOI} \triangleq \{A \in \mathcal{IS} \mid auo(A) \geq mo \wedge \text{supp}(A) \geq ms\}$.

Example 2. Consider itemset $A = bc$ in Example 1, $ms = 1$ and $mo = 0.1$. Since $auo(A) = 0.106 > mo$ and $supp(A) = 3 > ms$, the A itemset is an FHAUOI.

Definition 5 (Set CFHAUOI). Let A be an FHAUOI, $A \in \mathcal{FHAUOI}$. A is defined as a closed FHAUOI (CFHAUOI) if there does not exist any FHAUOI B such that $A \subset B$ and $supp(B) = supp(A)$. Then, let $\mathcal{CFHAUOI}$ denote the set of all CFHAUOIs, defined as

$$\mathcal{CFHAUOI} \stackrel{\text{def}}{=} \{A \in \mathcal{FHAUOI} \mid \nexists B \in \mathcal{FHAUOI}: B \supset A \wedge supp(B) = supp(A)\}.$$

Problem Definition. Given a quantitative database $\mathcal{D}'$ and two user-specified positive thresholds, minimum AU occupancy mo and minimum support ms , the objective is to identify all patterns belonging to the set of $\mathcal{CFHAUOI}$.

Example 3. For $ms = 2$ and $mo = 0.1$, the result set $\mathcal{CFHAUOI}$ discovered from the QDB $\mathcal{D}'$ in Table 1 is. $\mathcal{CFHAUOI} = \{abcdef, ad, bcef, bef\}$.

Definition 6 (Forward extension [5]). Assume that the elements of $\mathcal{A}$ are arranged following a predefined total order relation $<$. For any two itemsets A, B , and item y , we define $A < y$ iff $x < y$ for $\forall x \in A$; similarly, $A < B$ iff $x < y$, $\forall x \in A, \forall y \in B$. If A and B are disjoint, their union is written as $A + B$. Moreover, when $A \cap B = \emptyset$ and $A < B$, the notation $A + B$ is replaced with $A \oplus B$, and the resulting itemset $S = A \oplus B$ is referred to as a *forward extension* (FE) of A by B .

III. MINING CFHAUOIS

A. Pruning of low average utility occupancy itemsets

The reduction of the search space can be achieved effectively by exploiting the anti-monotonicity ($\mathcal{AM}$) property of the support function together with UBs derived from the auo measure. Specifically, when an itemset A is identified as infrequent, all of its supersets are guaranteed to be infrequent as well. Consequently, the branch $Br(A)$, which contains only such infrequent extensions of A , can be safely eliminated from further exploration. A function ub is a UB on auo if it satisfies $auo(C) \leq ub(C)$, $\forall C \in \mathcal{IS}$.

It is well-known that $supp$ is $\mathcal{AM}$. However, as auo does not satisfy the $\mathcal{AM}$ property, it is necessary to design $\mathcal{AM}$ UBs on auo . Inspired by the $\mathcal{AM}$ UB $auoub$ on auo [11], which represents the maximum item utility occupancy, we generalize it and develop a tighter new $\mathcal{AM}$ UB, named $lubauo$. Its value is the average of the top- k largest item utility occupancies. The $\mathcal{AM}$ property of $supp$ and $lubauo$ can be utilized to remove infrequent or LAUO branches from the prefix tree.

Definition 7 (Upper bound $lubauo$ on auo). For any frequent itemset $A \in \mathcal{FI}$ and transaction $t \in \rho(A)$, after sorting all item utilities of t in descending order, we obtain the series $T = (a_1, a_2, \dots, a_n)$, where $n = |t|$.

a. The k -largest average utility of t is defined as $lau(t, k) \stackrel{\text{def}}{=} \frac{1}{k} \sum_{i=1..k, a_i \in T} a_i$ for $1 \leq k \leq n$, and $lubauo(A, t) \stackrel{\text{def}}{=} \frac{lau(t, |A|)}{u(t)}$.

b. An upper bound (UB) on auo , called $lubauo$ (largest

average utility occupancy upper bound), is defined as

$$lubauo(A) \stackrel{\text{def}}{=} \frac{1}{ms} \sum_{i=1..ms, v_i \in V} v_i,$$

where $V = \{v_i, i = 1..|\rho(A)|\}$ that is obtained by sorting the series $\{lubauo(A, t) \mid t \in \rho(A)\}$ in descending order.

Example 4. For $ms = 2$, consider itemset $A = bc$ with $supp(A) = 3$, $length(A) = 2$ and $\rho(A) = \{t_1, t_2, t_5\}$. The computation of the new UB $lubauo(A)$ is as follows. First, for transaction t_1 , we have $lubauo(A, t_1) = (112 + 100)/2/286 = 0.37$ because 112 and 100 are the two largest utilities of t_1 . Similarly, for the two remaining transactions t_2 and t_5 , $lubauo(A, t_2) = (240 + 16)/2/268 = 0.48$ and $lubauo(A, t_5) = (165 + 24)/2/227 = 0.42$. Since $lubauo(A, t_2) > lubauo(A, t_5) > lubauo(A, t_1)$, we have $lubauo(A) = (0.48 + 0.42)/2 = 0.45$.

Proposition 1 (The $\mathcal{AM}$ properties of $supp$ and UB $lubauo$).

- $lubauo$ is a UB on auo , i.e. $auo(A) \leq lubauo(A)$, $\forall A$
- $lubauo$ is $\mathcal{AM}$, meaning $lubauo(B) \leq lubauo(A)$ for any super-itemset B of A in $\mathcal{FI}$, $B \supseteq A$.

Example 5. For $A = bc$ and $mo = 0.5$, the $Br(A)$ branch can be eliminated utilizing the new UB $lubauo$ since $lubauo(A) = 0.45 < mo$. For example, $auo(A) = 0.106 < lubauo(A) < mo$. Hence, the tighter UB $lubauo$ will be used in the pruning strategies, as stated in Theorem 1 below.

Let $PC(A) \stackrel{\text{def}}{=} (supp(A) < ms) \vee (lubauo(A) < mo)$ represent the *pruning condition* used to check whether A is an infrequent or LAUO itemset. Since the functions $supp$ and $lubauo$ are $\mathcal{AM}$, the necessary condition for any FE of A to be an FHAUOI is $not(PC(A))$. The following theorem is obtained based on the definition of PC .

Theorem 1 (Pruning strategy 1 based on $supp$ and $lubauo$ – PS1). For any item $y \in \mathcal{A}$, if $PC(y)$ is true, then this irrelevant item can be safely discarded from $\mathcal{D}'$. Furthermore, for any itemset A , if $PC(A)$ holds, the whole branch $Br(A)$ can also be pruned from the search tree.

B. Pruning of non-closed FHAUOIs

For brevity, for any itemset B , if $B \notin \mathcal{FHAUOI}$ (the set of all FHAUOIs) or B is not a CFHAUOI, then B is referred to as a *non-closed FHAUOI* (NCFHAUOI) itemset. Similarly, if the $Br(B)$ branch, which includes B and all its FEs, has no CFHAUOIs, it is referred to as an *NCFHAUO branch*.

Note that the above UB targets fast elimination of LAUOIs, but FHAUOI mining can still yield many results, especially in large QDBs or with low mo and ms . Extracting CFHAUOIs from all FHAUOIs is challenging due to costly inclusion checks. This motivates designing weak lower bounds (WLB) on auo to enable early pruning of non-closed FHAUOIs and unpromising branches.

Definition 8 (WLB and LB on auo). For any transaction and itemset A , a function $wlb(A, t)$ is called a WLB on $auo(A, t)$ if $wlb(A, t) \leq auo(C, t)$ for any proper FE C of A . Similarly, wlb is referred to as a WLB on auo if $wlb(A) \leq auo(C)$ for any proper FE C of A .

Definition 9 (WLB $wlbauo$ on auo). Consider any itemset A and transaction $t \in \delta(A)$. A WLB on $auo(A, t)$, denoted $wlbauo(A, t)$, is calculated as

$$wlbauo(A, t) = \frac{sau(A, t)}{u(t)},$$

where $sau(A, t)$ is the smallest average utility of proper forward extensions of A in t , which is defined as

$$sau(A, t) = \begin{cases} f(r), & \text{if } mu(A, t) \leq au(A, t) \\ f(1), & \text{if } mu(A, t) > au(A, t) \end{cases}$$

where $mu(A, t) = \min\{u(a, t) \mid (a, q) \in rem(A, t)\}$, $r = |rem(A, t)|$, $l = length(A)$ and $f(k) \stackrel{\text{def}}{=} \frac{l \cdot au(A, t) + k \cdot mu(A, t)}{l + k}$ is an intermediate value between $au(A, t)$ and $mu(A, t)$, $\forall k = 1 \dots r$.

To construct a novel WLB on the auo function, we utilize the values $wlbauo(A, t)$ corresponding to each transaction t containing itemset A . Specifically, for an itemset A such that $|\delta(A)| \geq ms$, the series $\{wlbauo(A, t) \mid t \in \delta(A)\}$ is arranged in ascending order to form a new ordered sequence $T \stackrel{\text{def}}{=} \{t_i, i = 1 \dots n\}$. Here, we have $supp(A) \geq n \stackrel{\text{def}}{=} |\delta(A)| \geq ms$ as $\delta(A) \subseteq \rho(A)$.

Definition 10 (WLB on auo). Given any itemset A such that $|\delta(A)| \geq ms$, a novel WLB on auo is formally represented and defined as $wlbauo(A) = \frac{1}{ms} \sum_{i=1 \dots ms} t_i$.

Proposition 2 (Properties of WLB on auo). Consider any an itemset A , transaction $t \in \delta(A)$, and proper FE $B = A \oplus C$ of A with $C \neq \emptyset$. Then,

a) $wlbauo(A, t) \leq auo(B, t)$.

b) $wlbauo$ is a WLB on au , i.e. $wlbauo(A) \leq auo(B)$.

Example 6. For $ms = 2$, consider itemset $A = bc$, $l = length(A) = 2$, and $\delta(A) = \{t_1, t_2, t_5\}$. For transaction t_5 , then $mu(A, t_5) = 8$, $r = 3$ and $au(A, t_5) = (24 + 6)/2 = 15$. Since $mu(A, t_1) < au(A, t_5)$, so $sau(A, t_5) = f(r) = f(3) = \frac{2 \cdot 15 + 3 \cdot 8}{2+3} = 10.8$. Thus, $wlbauo(A, t_5) = \frac{10.8}{227} = 0.048$. For t_1 , since $mu(A, t_1) = 8 < au(A, t_1) = 66$, so $sau(A, t_1) = f(3) = \frac{2 \cdot 66 + 3 \cdot 8}{2+3} = 31.2$, and $wlbauo(A, t_1) = \frac{31.2}{286} = 0.109$. Similarly, for t_2 , since $mu(A, t_2) = 16 > au(A, t_2) = 6$, so $sau(A, t_2) = f(1) = \frac{2 \cdot 6 + 16}{2+1} = 9.33$, and $wlbauo(A, t_2) = \frac{9.33}{268} = 0.035$. After sorting the series $\{wlbauo(A, t) \mid t \in \delta(A)\} = \{0.109, 0.035, 0.048\}$ in ascending order, the new series $T = \{0.035, 0.048, 0.109\}$ is obtained. Then, $wlbauo(A) = (0.035 + 0.048)/2 = 0.0415$. For the proper FE B of A , $B = A \oplus e = bce$, then $auo(B) = 0.608 > wlbauo(A)$.

Based on the proposed WLB, $wlbauo$, this paper introduces the first novel pruning strategy aimed at early elimination of NCFHAUO branches in the prefix search tree. This strategy leverages itemsets already generated in the CFHAUOI set and is presented as follows.

Theorem 2 (Pruning Strategy 2 – PS2). For the current itemset C being examined when constructing the CFHAUOI set, if there is a CFHAUOI $P \in CFHAUOI$ such that $P \supset C$,

$lastItem(P) = lastItem(C)$, and $supp(C) = supp(P)$, then if $wlbauo(P) \geq mo$, the entire NCFHAUO proper branch $pBr(C)$ is pruned early. Additionally, if $auo(P) \geq mo$ or $C \notin CFHAUOI$, the NCFHAUO node C is also eliminated.

Corollary 1 (Pruning Strategy 3 – PS3). For any two nodes $C_1 = D \oplus x$ and $C_2 = D \oplus y$ having the same parent in the prefix search tree such that $y > x$, when extending C_1 with item y to have the larger itemset $P = C_1 \oplus y$, if $supp(P) = supp(C_2)$ and $wlbauo(P) \geq mo$, then the whole NCFHAUO $pBr(C_2)$ proper branch is safely eliminated early. Additionally, the NCFHAUO C_2 node is also discarded, if $auo(P) \geq mo$ or $C_2 \notin CFHAUOI$.

Note that, due to space limitations, the proofs of the theorems and propositions are omitted from this paper.

C. Proposed C-FHAUOIM Algorithm

Building upon the theoretical foundations established earlier, this part introduces an innovative algorithm, referred to as C-FHAUOIM, which is designed to efficiently identify the collection of CFHAUOIs. The proposed approach specifically aims to overcome the inherent difficulties associated with extracting such patterns. For its implementation, the algorithm employs the NVUV-list data representation introduced in [5], which facilitates effective computation and pruning during the mining process.

C-FHAUOIM ($\mathcal{D}'$, mo , ms)

Input: the QDB $\mathcal{D}'$, mo and ms .

Output: the set CFHAUOI.

1. **Scan** $\mathcal{D}'$ to calculate the support and WUB $lubauo$ of all items, and determine the set $I \stackrel{\text{def}}{=} \{x \in \mathcal{A} \mid lubauo(x) \geq mo \wedge supp(x) \geq ms\}$; //Pruning strategy 1
2. **Remove** all irrelevant items in $\mathcal{A} \setminus I$ from $\mathcal{D}'$; // PS 1
3. **Sort** I in ascending order of item support;
4. **Scan** $\mathcal{D}'$ again to construct the structure NVUV-list for all relevant items of I ;
5. $CFHAUOI = \emptyset$;
6. **for each** $x \in I$ **do** {
7. **Mine-CI**(x , mo , ms , $CFHAUOI$);
8. }
9. **return** $CFHAUOI$;

Fig. 1. Algorithm C-FHAUOIM

The pseudo-code of the proposed algorithm is depicted in Fig. 1, with the objective of mining the CFHAUOI. It takes $\mathcal{D}'$, mo and ms as input. C-FHAUOIM initially scans $\mathcal{D}'$ to calculate the $lubauo$ and support values for items in $\mathcal{A}$ and determine the I set (line 1). Then, C-FHAUOIM removes irrelevant items from $\mathcal{A} \setminus I$ in $\mathcal{D}'$ to minimize the candidate set for extension (line 2). After sorting the I set in ascending order of item support, the algorithm scans the QDB $\mathcal{D}'$ again to construct the structure NVUV-list for items in I (lines 3-4). Then, with each item x in I , the Mine-CI procedure is invoked at line 7 to discover CFHAUOIs in the $pBr(x)$ branch. Finally, C-FHAUOIM offers the CFHAUOI set as its output (line 9).

The Mine-CI procedure depicted in Fig. 2 first verifies if $supp(C)$ is less than ms . If yes, it stops immediately (line 1). Then, it checks if the node C or the $pBr(C)$ proper branch is pruned by Pruning Strategy 3 in Theorem 5 or not (line 2-6). If yes, the fields *Prune_NC_Node* and/or *Prune_NC_pBr* of

itemset C are marked to prune in subsequent steps. Next, if C is an FHAUOI ($auo(C) \geq mo$) and the value of its field $Prune\_NC\_Node$ is false (lines 7-8), then the procedure Update-Closed-FHAUOI is called at lines 9-10 to update the $CFHAUOI$ set. After that, Mine-CI evaluates whether any of the conditions outlined at line 12 are met. If affirmative, the $pBr(C)$ is pruned prematurely (line 13). In the case of negative, a new set of items, $newI$, is determined for expanding C (line 17), utilizing the criteria specified at line 16. Subsequently, each item from $newI$ is added to extend the itemset C , generating a new itemset $P = Cy$, and establishing the NVUV-list structure of P at line 20. The NCFHAUO-Branch-Pruning procedure is invoked to identify NCFHAUO branches and nodes for subsequent pruning steps (line 21). Ultimately, Mine-CI calls itself recursively to explore larger CFHAUOIs

```

Mine-CI ( $C, mo, ms, CFHAUOI$ )
Input: the itemset  $C$ ,  $mo$  and  $ms$  thresholds, and set  $CFHAUOI$ .
Output:  $CFHAUOI$  is updated.
1. if ( $supp(C) < ms$ ) return;
2. for each  $P \in CFHAUOI$  and  $P \supset C$  do {
3.   NCFHAUO-Branch-Pruning ( $C, P$ ); // Theorem 2
4.   If ( $C.Prune\_NC\_pBr = true$ ) then
5.     break;
6. }
7. if ( $auo(C) \geq mo$ ) then {
8.   if ( $C.Prune\_NC\_Node = false$ ) then
9.     Update-Closed-FHAUOI ( $C, CFHAUOI$ );
10. }
11. if ( $C.Prune\_NC\_pBr = true$ ) // Pruning strategies 2-3
12.   return;
13.  $x = lastItem(C)$ ;  $D = prefix(C)$ ;  $newI = \emptyset$ ;
14. for each  $y \in I$  such that  $y > x$  do {
15.   if ( $lubauo(Cy) \geq mo$  and  $supp(Cy) \geq ms$ ) then // PS 1
16.      $newI = newI \cup \{y\}$ ;
17.  $Exts\_of\_C = \emptyset$ ;
18. for each  $y \in newI$  do {
19.   Add  $P = Cy$  to  $Exts\_of\_C$  and construct NVUV-list of  $P$ ;
20.   NCFHAUO-Branch-Pruning ( $Dy, P$ ); // PS3
21. }
22. for each  $P \in Exts\_of\_C$  do
23.   Mine-CI( $P, mo, ms, CFHAUOI$ );

```

Fig. 2. Procedure Search-FHAUOS.

IV. EXPERIMENTAL EVALUATION

The experiments were conducted on 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.4 GHz (8 CPUs) with 12.288 GB RAM, running Windows 10 Home 64 bit. All comparison algorithms were implemented in Java SE 8 and executed using the NetBeans IDE. The evaluation was carried out on two QSDBs, including one real-world dataset BMS [16] and one synthetic dataset T2014N600D70K, where T2014N600D70K is dense and BMS is sparse. The characteristics of the QSDBs are shown in Table 2. Utilities were randomly generated using the IBM Quest data generator (obtained from [16]). To the best of our knowledge, no algorithm has been proposed for discovering CFHAUOIs to date. Therefore, we compare the performance of the C-FHAUOIM algorithm with a Baseline algorithm that does not utilize the novel theoretical results proposed this paper.

TABLE 2. CHARACTERISTICS OF DATABASES

Database	#Tran	#Item	Avg. length	Data type
BMS	59,602	497	2.5	Real-life
T2014N600D70K	70,000	600	20	Synthetic

A. Influence of the parameter mo

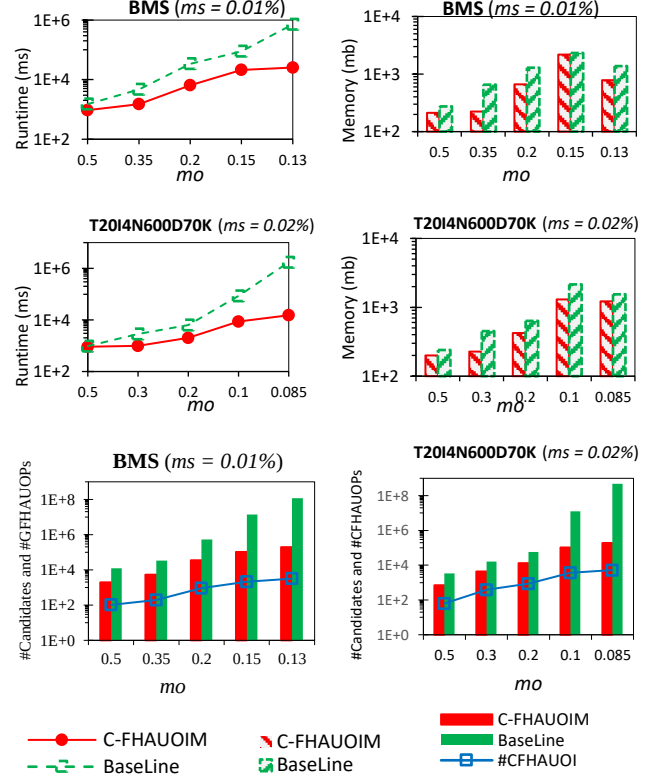


Fig. 3. The influence of varying the mo threshold.

Firstly, we examine the impact of mo on the performance of the algorithms using the QSDBs listed in Table 3, with results shown in Fig. 3. As mo decreases, execution time, memory usage, and the number of candidates grow substantially, since more patterns satisfy the threshold. The figure also shows that C-FHAUOIM consistently outperforms the Baseline in all three metrics. For example, on the QTDBs BMS and T2014N600D70K, C-FHAUOIM is approximately 15 times and 66 times faster, respectively. This advantage stems from its novel strategies and the proposed bounds, which frequently satisfy pruning conditions and effectively shrink the search space. Moreover, C-FHAUOIM avoids frequent, costly inclusion-relationship checks, unlike the Baseline, further improving efficiency.

B. Influence of the parameter ms

Next, we assess the effect of varying the ms parameter while keeping mo fixed, as shown in Fig. 4. Lower ms values lead to a sharp rise in execution time, memory usage, and #Candidates, since more patterns qualify for evaluation. Across all datasets, C-FHAUOIM clearly surpasses the Baseline in performance. For instance, on BMS and T2014N600D70K, it achieves speedups of about 245 times and 39 times, respectively, thanks to its effective pruning strategies and

proposed UB/WLB bounds that drastically reduce the search space and minimize costly inclusion checks.

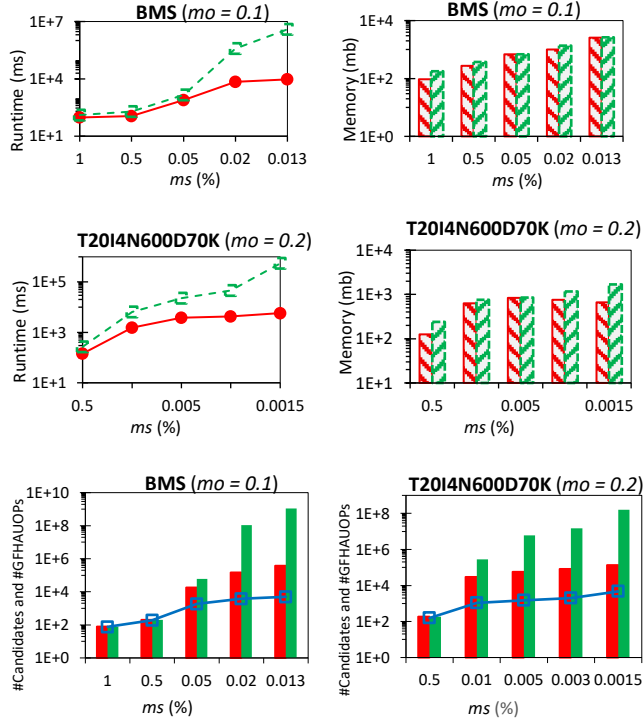


Fig. 4. The influence of varying the ms threshold.

In summary, both mo and ms have a substantial impact on runtime, memory usage, and the number of candidates. In every setting, C-FHAUOIM consistently outperforms the Baseline by achieving faster runtimes and lower memory consumption. These improvements are largely due to its effective UB/WLB-based pruning strategies, which greatly reduce the search space and avoid costly inclusion checks.

V. CONCLUSION

This paper presents new theoretical contributions to the problem of mining CFHAUOIs. To address this problem efficiently, we propose a novel upper bound and a weak lower bound for the auo function, along with associated pruning strategies that significantly reduce the search space during mining. Building on these theoretical insights, we introduce C-FHAUOIM, an algorithm specifically designed to effectively identify CFHAUOIs. Experimental evaluations demonstrate that C-FHAUOIM consistently outperforms the Baseline algorithm in terms of runtime, memory usage, and candidate count. Looking ahead, we plan to integrate parallel processing techniques to further boost the efficiency of CFHAUOI mining, particularly for large-scale datasets.

ACKNOWLEDGMENT. This research is funded by Dalat University, Vietnam under decision number 1153/QĐ-ĐHĐL, 2024.

REFERENCES

[1] H. Yao and H. J. Hamilton, “Mining itemset utilities from transaction databases,” *Data Knowl Eng*, vol. 59, no. 3 SPEC. ISS., pp. 603–626, 2006.

[2] T.-P. Hong, C. H. Lee, and S. L. Wang, “Mining high average-utility itemsets,” in *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 2009, pp. 2526–2530.

[3] B. E. Shie, P. S. Yu, and V. S. Tseng, “Mining interesting user behavior patterns in mobile commerce environments,” *Applied Intelligence*, vol. 38, no. 3, pp. 418–435, 2013.

[4] H. Kim, U. Yun, Y. Baek, J. Kim, B. Vo, and E. Yoon, “Efficient list based mining of high average utility patterns with maximum average pruning strategies,” *Inf Sci (N Y)*, vol. 543, pp. 85–105, 2021.

[5] T. Truong, H. Duong, B. Le, P. Fournier-Viger, and U. Yun, “Efficient high average-utility itemset mining using novel vertical weak upper-bounds,” *Knowl Based Syst*, vol. 183, no. 1, p. 104847, 2019.

[6] T. Tran, H. Duong, T. Truong, and A. Tran, “High Average-Utility Itemset Mining with A Novel Vertical Weak Upper Bound,” in *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Dec. 2023, pp. 342–347.

[7] B. Shen, Z. Wen, Y. Zhao, D. Zhou, and W. Zheng, “OCEAN: Fast discovery of high utility occupancy itemsets,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, pp. 354–365.

[8] W. Gan, J. C. W. Lin, P. Fournier-Viger, H. C. Chao, and P. S. Yu, “HUOPM: High-Utility Occupancy Pattern Mining,” *IEEE Trans Cybern*, vol. 50, no. 3, pp. 1195–1208, Mar. 2020.

[9] H. Kim *et al.*, “Mining high occupancy patterns to analyze incremental data in intelligent systems,” *ISA Trans*, vol. 131, pp. 460–475, 2022.

[10] L. T. T. Nguyen, T. Mai, G. H. Pham, U. Yun, and B. Vo, “An efficient method for mining high occupancy itemsets based on equivalence class and early pruning,” *Knowl Based Syst*, vol. 267, p. 110441, 2023.

[11] M. J. K. Kumar and D. Rana, “HAUOPM: High Average Utility Occupancy Pattern Mining,” *Arab J Sci Eng*, vol. 49, pp. 3397–3416, 2024.

[12] H. Duong, T. Hoang, T. Tran, T. Truong, B. Le, and P. Fournier-Viger, “Efficient Algorithms for Mining Closed and Maximal High Utility Itemsets,” *Knowl Based Syst*, vol. 257, p. 109921, 2022.

[13] H. Duong, T. Truong, B. Le, and P. Fournier-Viger, “CG-FHAUI: an efficient algorithm for simultaneously mining succinct pattern sets of frequent high average utility itemsets,” *Knowl Inf Syst*, vol. 66, pp. 5239–5280, 2024.

[14] H. Duong, H. Pham, T. Truong, and P. Fournier-Viger, “Efficient algorithms to mine concise representations of frequent high utility occupancy patterns,” *Applied Intelligence*, vol. 54, pp. 4012–4042, Mar. 2024.

[15] D. Hai and T. Tin, “An Efficient Algorithm for Mining Frequent High Utility Occupancy Generators,” in *CCIOT ’24: Proceedings of the 2024 9th International Conference on Cloud Computing and Internet of Things*, Association for Computing Machinery (ACM), Jan. 2025, pp. 101–109.

[16] P. Fournier-Viger *et al.*, “SPMF: a Java Open-Source Pattern Mining Library Version 2,” *Machine Learning Research*, vol. 15, no. 1, pp. 3389–3393, 2014.

An interactive time suggestion for multi-objective evolutionary algorithms

Thu To Thi
MIITE
Academy of Military Science
and Technology
Hanoi, Vietnam
Email: thuto88.hvktqs@gmail.com

Long Nguyen
Faculty of Logistics and Engineering
National Defense Academy
Hanoi, Vietnam
Email: longit76@gmail.com

Dinh Nguyen Duc
MIITE
Academy of Military Science
and Technology
Hanoi, Vietnam
Email: nddinh76@gmail.com

Abstract—Multi-objective optimization has more and more practical applications in all areas of life. Evolutionary algorithms for multi-objective optimization are an effective choice and are widely used in applying mathematics to computer science. However, because the result of the algorithm is a solution set, the choice of the most optimal solution depends a lot on the choice of the decision maker, not only after the algorithm finishes the search process but also during the search process, the interaction results of the decision maker will create an evolutionary trend for the next generations, close to the target area that the decision maker wants without being a hard constraint.

However, the multi-objective evolutionary optimization algorithm has good quality when maintaining a good balance between convergence and diversity as well as exploration and exploitation capabilities, so the use of decision maker's reference information needs to be very careful, avoiding the search process that is unbalanced between exploration and exploitation capabilities, as well as not ensuring convergence and diversity quality at the same time. In this paper, we propose to suggest interaction timing for the decision maker based on the analysis of the quality and characteristics, evolutionary trends of the algorithm, in order to ensure the best interaction quality while preserving the balance between exploration and exploitation, and maintaining population convergence and diversity during the evolutionary process. The experimental results on the interactive multi-objective optimization algorithm give relatively competitive results, with high application value.

Keywords—Interactive evolutionary multi-objective optimization, DMEA-II, NSGA-II, MOEA/D

I. INTRODUCTION

In the real world, optimization problems usually have multiple objectives, and they need to be solved simultaneously, when they conflict with each other, we have the class of multi-objective optimization problems and multi-objective optimization problems (at least 2 objectives). In general, the class of k -objective problems is represented by the following formula:

$$\vec{f}(\vec{x}) = [f_1(\vec{x}), f_2(\vec{x}), \dots, f_k(\vec{x})] \quad (1)$$

Here, $\vec{x}$ denotes a decision vector composed of v variables, each defined in the real-valued space $\mathbb{R}^v$. In the framework of evolutionary computation (EC), $\vec{x}$ corresponds to a candidate solution (or individual) within the evolving population. Each objective function $f_j(\vec{x})$ evaluates the quality or fitness of the

individual $\vec{x}$ with respect to the j -th objective of the multi-objective optimization problem.

There are many methods to solve optimization problems, classical methods applied to solve multi-objective optimization problems can be mentioned as: weighted sum, epsilon constraint, mathematical programming, simulation... however, these methods encounter many difficulties with complex problems in practice, especially problems where the objective function is not differentiable, even black box. With the effective application of evolution, on the basis of random operation, multi-objective optimization evolutionary algorithm is an effective and popular choice due to the nature of working on the population, random mechanism in the condition that the optimization problem needs to find a set of feasible solutions.

Through the principle of dominance relationship, Pareto theory, algorithms all aim at a set of solutions that approach and spread evenly on the Pareto layer of the problem. The solutions in the feasible solution set are solutions that are not dominant to each other. There are many ways to classify methods for solving multi-objective optimization problems, which is the use of reference information. Non-reference method: the decision maker will not use reference information but will accept or reject the solution recommended by the optimal search process. Prior method: reference information is included before the search process as a basis for choosing the final solution. Posterior method: reference information is included after the optimal search process to choose the optimal solution. And finally, the interactive method, reference information is included one or more times during the optimal search process to give priorities and choices during the search process...

In the interactive model, the decision maker tries to provide his desired information in various forms of information: reference points, reference lines, value functions... to provide for the evolution process, to have a priority close to the decision maker's preferences. In the field of interactive multi-objective optimization, there are many research problems such as: how is the decision maker's reference information represented? how is it used to control the evolution process? how to achieve effective interaction? ... On the other hand, whether or not interaction is used, the algorithm quality is good

when it maintains a good balance between exploration and exploitation, while ensuring both convergence and diversity within the solution population. In this paper, we would like to propose the analysis and evaluation of population quality, evolutionary characteristics, and interaction trends to develop a strategy on the number of times and reasonable time of interaction to achieve effective interaction while ensuring the quality of the algorithm both in terms of the search process and the quality of the solution set.

The paper is structured into 4 sections, section I is a general introduction, section II presents recent related works, section III presents the methodology of time suggestion for multi-objective evolutionary algorithms, section IV is the experiment and results. Section 5 is the Conclusion.

II. RELATED WORKS

A. Interactive evolutionary multi-objective optimization

Interactive Multi-Objective Evolutionary Optimization (IMOE) has emerged as a significant research direction within the field of evolutionary computation. Its primary focus lies in solving multi-objective optimization problems by incorporating human decision-making into the optimization process. Unlike traditional multi-objective optimization, which aims to approximate the Pareto front comprehensively, IMOE seeks to identify solutions that align closely with the preferences of decision-makers, thereby reducing computational complexity and enhancing the relevance of the outcomes.

The research in IMOE can be categorized into several key areas.

The first, preference elicitation methods are central to the field, enabling decision-makers to express their preferences effectively. These methods include explicit preference articulation, such as ranking or weighting objectives, and implicit articulation based on interaction during the search process. Advances in machine learning techniques have further enriched these methods, allowing for more adaptive and accurate preference modeling. In [1] the authors analyzed three interactive multi-objective optimization (IMO) algorithms — I-NSGA-II, I-GDE3, and I-GDE3+APM — which progressively incorporate reference points during the search process. These methods were tested on benchmark multi-objective single-objective problems (MOSOPs) using standard performance indicators. The results showed that the interactive variants generally outperformed their corresponding *a priori* versions and can achieve competitive results without predefined reference points, allowing decision makers to make more informed and reliable choices. The authors in [2] proposed an interactive final solution selection (IFSS) method for multi-objective optimization. The approach enables the decision maker to reach a satisfactory final solution through a limited number of interaction rounds. Notably, it allows effective decision support even when the decision maker's preferences are not explicitly specified at the beginning of the process.

The second, the development of interactive evolutionary algorithms (IEAs) is a critical aspect of IMOE research.

IEAs integrate user feedback iteratively, updating the search process to focus on regions of the solution space that align with the decision-maker's preferences. This involves designing algorithms capable of balancing exploration and exploitation, as well as ensuring robustness against noisy or inconsistent feedback. In [3] the authors proposed to exploit the wellknown Lower Confidence Bound acquisition function within Bayesian optimization, to interactively infer the DM's preferred solution in a data-efficient manner, even in scenarios where the DM experiences uncertainty in their decision-making process. In [4], The authors introduced an interactive framework that engages decision makers in identifying their most preferred robust solutions through multi-objective optimization. The framework incorporates a learning phase that enables decision makers to better understand the problem structure, the relationships among objectives and scenarios, the practicality of their expressed preferences, and the influence of uncertainty on decision outcomes.

The third, there is not much research on interactive adaptive adjustment. In this area, there is a paper [5] that introduces an adaptive interactive approach based on decision neural network. The neural network is used to capture and represent the decision maker's preferences as input for finding the desired solution. Then, by applying a genetic algorithm, the most desired solution will be provided. The authors in [6] proposed an interactive particle swarm meta-algorithm for multi-objective optimization (MOO) that integrates Pareto dominance with interactive decision-making. Pareto dominance is employed to evaluate particles during the search, while non-dominated solutions are maintained in an external archive updated through an adaptive grid mechanism. The decision maker (DM) interacts with the approximate Pareto front, and his or her preferences are used to infer a utility function guiding the search toward preferred regions. By combining adaptive grid management, a self-adaptive mutation operator, and a preference-based decision strategy, the method maintains a good balance between convergence and diversity, making it both effective and innovative.

Finally, the IMOE research also emphasizes the real-world applications of these methods. Domains such as engineering design, resource allocation, and environmental management have benefited significantly from IMOE, as these problems often involve competing objectives and require human expertise to achieve practical solutions [7], [8], [9].

In summary, IMOE combines human intelligence with computational optimization, creating a synergistic approach to solving complex multi-objective problems. By advancing preference elicitation methods, interactive algorithms, and visualization tools, and by demonstrating its applicability in real-world scenarios, IMOE continues to push the boundaries of evolutionary optimization research.

B. The usage of DM's reference information

In interactive multi-objective evolutionary optimization algorithms, the decision maker expresses his desired information to find the optimal solution suitable for his problem in many

forms, which can be reference points, reference lines, reference regions, value functions, constraint values... provided to the search process (population evolution process) to have priority in choosing solutions that are close to the ideal solutions of the decision maker. The decision maker's reference information can be provided one or more times during the search process, depending on the satisfaction or desired adjustment of the decision maker. In interactive multi-objective evolutionary optimization algorithms, depending on the principle of each algorithm, the decision maker's information will be used to adjust the control parameters, to create preferences generated from that reference information. Sometimes it is the adjustment of techniques in the selection, crossover or mutation operators; stores the external sets containing elite individuals, combining the individuals selected from the preferences from that reference information, to maintain and generate the next generations, reaching the regions containing the decision maker's ideal solutions.

The use of the decision maker's reference information to prioritize the search for the desired ideal solutions would not be worth mentioning, if the use of reference information could create an imbalance between the ability to explore and exploit, as well as the quality of convergence and diversity of the solution set. In fact, these are very important factors if a multi-objective optimization evolutionary algorithm is to be evaluated as good and effective. The interaction can cause some other problems such as: if using the preference with strong forcing, it can make the evolutionary algorithm fall into a local optimum, losing global validity; Or, with the ever-flexible mechanism to maintain the balance of the multi-objective evolutionary optimization algorithm, the reference information is introduced at times when there is little priority, and the interaction has no obvious effect.

From the analysis of those issues, we propose some techniques to improve the quality of the multi-objective optimization evolutionary algorithm, through analyzing the quality of the population, analyzing the interaction trends of the decision maker to suggest effective interaction times; reducing the number of ineffective interactions to achieve the best simultaneous between the desire of the decision maker and ensuring the quality and efficiency of the algorithm.

III. METHODOLOGY

Analyzing the evolutionary characteristics of evolutionary algorithms, there is a common feature of all multi-objective optimization evolutionary algorithms: in the early stage of the evolution process, from the initial random initialization mechanism, the diversity quality of the basic population is quite good, but the convergence quality is very poor. At this stage, the algorithms all try to create a balance between convergence quality and diversity by increasing depth-first search, or the so-called exploration ability, prioritizing the exploitation ability. On the contrary, at the final stage, after the evolution process, the quality of the population is now very good in both convergence and diversity, the development for adjacent generations is very small. In the intermediate stage, depending

on the mechanism of each algorithm, the balance between convergence quality and diversity can vary, before achieving a good balance in the generations in the final stage. Thus, it can be seen that there are two stages where user interaction with reference information will not be effective. Specifically, if the interaction is in the early stage, there is a high probability that there will be no priority if the reference area created is biased towards increasing diversity quality, on the contrary, if it is biased towards increasing convergence quality, the algorithm can easily fall into a local convergence point, because at this time, the algorithms are trying to increase convergence to create a balance between convergence quality and diversity by increasing exploration ability. From that analysis, we propose a technique to suggest effective interaction timing as follows: In the search process, we will identify non-interactive timing and effective interaction timing.

The first, we identify two non-interactive stages as the initial stage and the final stage because these stages have limited interaction efficiency.

Next, in the intermediate stage, we will evaluate the balance between convergence quality and diversity based on the calculation of the generation measure, if the rate approaches the equilibrium, it is determined to be the appropriate generation for interaction, otherwise, when there is a large enough imbalance, the timing is determined to be ineffective for interaction. The calculate is described as follows: The generation distance measurements are defined in [10]

- **The generational distance (GD)** is defined as the average distance from a set of solutions, denoted P , found by evolution to the global POF. The first-norm equation is defined as follows:

$$GD = \frac{\sum_{i=1}^{N_P} d_i}{N_P} \quad (2)$$

In this metric, d_i represents the Euclidean distance (in objective space) between the i solution in the obtained set P and its nearest counterpart on the true POF, while N_P denotes the total number of solutions in P . This measurement is only considered for the convergence aspect of performance. Therefore, it may happen that the solution set is very close to the POF (good convergence) but not evenly distributed over the whole POF (poor diversity).

- **The inverse generational distance (IGD)** considers both the convergence and diversity of the solution set. The first-norm equation for IGD is as follows:

$$IGD = \frac{\sum_{j=1}^{N_{POF}} d_j}{N_{POF}} \quad (3)$$

where d_j denotes the Euclidean distance (in the objective space) between the j reference point on the true Pareto optimal front (POF) and its nearest neighbor in the obtained solution set P , while N_{POF} is the size of the POF. In order to get a good value for IGD (ideally zero), the solution set P should provide a uniform and comprehensive coverage of the POF. However, since the

IGD metric only measures distances from the POF to the closest elements in P , certain solutions in P may not contribute to this computation.

In this research, the variation in population quality is examined through the changes in the generational distance (GD) and inverted generational distance (IGD) metrics during the early phase of evolution, denoted by n_1). Within these initial n_1 generations, a segmented evaluation approach is applied: after every n_r generations, the GD and IGD values of the current population are compared with those obtained n_r generations earlier). By analyzing the relationship between the fluctuations of GD and IGD, this study uses the observed correlation as a reference indicator to adjust algorithmic parameters, aiming to minimize this correlation value toward zero. The detailed procedure is outlined as follows:

- 1) Initialization parameters: Number of surveyed generations n_1 ; number of generations to use as segmentation n_r ; the maximum measure value of each metric $maxGD = 0$, $maxIGD = 0$.
- 2) The main generation loop of the algorithm:
 - a) 1: Calculate and store GD_i and IGD_i values for current generation i , update values for $maxGD$, $maxIGD$ on GD_i , IGD_i respectively.
 - b) 2: Within the inner loop of the main evolutionary process (using the number of generations or stopping condition), execute:
 - Determine if the segmentation condition is true ($i \bmod n_r = 0$), do the following:
 - i) Calculate the change in the GD metric
$$\Delta_{GD_i} = \frac{GD_{i-n_r} - GD_i}{maxGD} \quad (4)$$
 - ii) Calculate the change in the IGD metric
$$\Delta_{IGD_i} = \frac{IGD_{i-n_r} - IGD_i}{maxIGD} \quad (5)$$
 - iii) Calculate the correlation between GD and IGD metrics
$$\Delta_i = \Delta_{GD_i} - \Delta_{IGD_i} \quad (6)$$
 - Adjust the algorithm based on Δ reference information.
 - c) 3: Ends sub-loop after n_1 generations.

During the evolution of the algorithm, the ratio between the two metrics is always calculated, according to the results of the ratio, green, red and yellow moments will be given to recommend when the user should interact or not to ensure the effectiveness of the interaction and maintain The algorithm's capability to balance exploration and exploitation plays a crucial role in maintaining both convergence quality and population diversity, which are key criteria for assessing the performance of multi-objective evolutionary optimization algorithm. In which, the boundaries between regions are determined based on the absolute value of the ratio between the two above measurements. Specifically, when the absolute value of the ratio $|\Delta_i|$ is less than ϵ_1 , that moment is a green

time; when the value is in the range $[\epsilon_1, \epsilon_2]$ then that moment is a golden time; finally when the value is greater than ϵ_2 it is a red time. The meanings of the moments are as follows: When the green moment is, it is the best moment for users to provide interactive information; the golden moment is the time when interaction is acceptable and finally the red moment is the time when users are not recommended to interact to ensure that user information does not hinder the balance between the exploration and exploitation capabilities of the population. The colors of the times are chosen as indicator colors to decide the delicate interaction time in the solution-finding process of the algorithm. The determination of the values of ϵ_1 and ϵ_2 depends on the value of the Δ ratio in the last generation without interaction; it is usually 0 or approaching 0, which is called the ideal ratio point. Then, ϵ_1 will be equal to this ideal ratio point plus a value to determine the boundary for the most suitable interaction region, similarly for the ratio limit for possible interaction. Note that the value of the ideal ratio point depends on the algorithms, on different problems it is necessary to run it once without interaction to determine this value. The illustration of state of moments is described

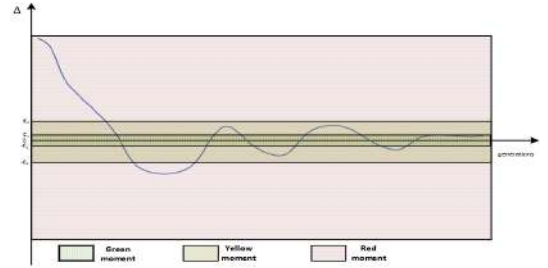


Fig. 1: The illustration of state of moments

on Fig 1 Thus, when the decision maker wants to interact, from the state of checking the population quality balance, the interactive loop framework will suggest whether to interact or not? The technique is represented in the form of an interaction function as follows:

- **Step 1:** DM asked to interact with evolutionary process?
- **Step 2:** Check the ratio between GD and IGD of current population then set the interaction status for the interaction. The moment status is indicated by green, yellow, red moment.
- **Step 3:** Say 'Yes' when the time is green, 'Maybe' when the time is yellow, and 'No' if the time is red, and perform the interaction function at the DM's discretion.

IV. EXPERIMENTS

A. Building up on reference points interactive methods

To test the proposed technique, experiments were conducted on two algorithms MOEA/D [11], DMEA-II [12] with an interactive method using published reference points. The experimental problem uses the on ZDTs benchmark sets [13] with two objectives.

Through the analysis of the evolution characteristics of multi-objective optimization algorithms, both DMEA-II and

MOEA/D have the characteristics of the early stage evolution process to prioritize convergence and the late stage to try to maintain diversity. Therefore, the interaction in the early stage will not be very effective, it will be effective when the balance reaches a good value.

In addition to that specific characteristic, in the interactive method with DMEA-II [14], using the reference point set, the reference points are used to determine the new ray system (In DMEA-II, the ray system is used to maintain the diversity of the population), in which some rays pass through the reference point, and the rays in the far area of interaction are removed, ensuring that the number of rays is preserved. Therefore, the creation of new ray density will greatly affect the process of maintaining the balance of exploration and exploitation capabilities.

And with MOEA/D, the interactive method introduced in [15], the reference point set is used in combination with the reference points of MOEA/D to create new reference points. In which, the reference points from the original MOEA/D algorithm are used to create boundaries, defining subproblems in the decomposition method. Therefore, the creation of new reference points has a strong impact on the evolution process of the MOEA/D algorithm.

B. Results

With the method of determining the green, yellow, and red moments in the proposed technique, applied to the DMEA-II and MOEA/D algorithms with reference information represented through sets of reference points. The results determine the moment regions as shown in the figure below Fig. 2 - Fig.6

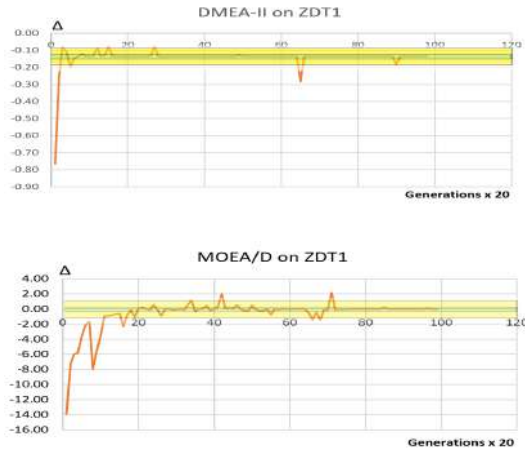


Fig. 2: Results of identifying interactive suggestion regions on ZDT1 problem

Through the experimental results of the interactive algorithm with DMEA-II and MOEA/D on ZDT problems, it can be seen that. With the DMEA-II algorithm, the ideal ratio point has a negative value, so the ideal feasible suggestion regions are below the zero ratio. With MOEA/D, the ideal

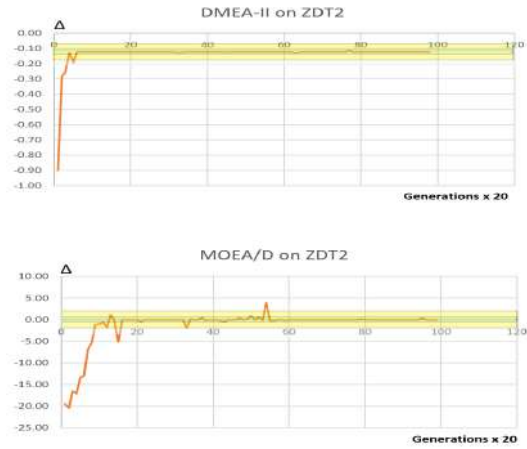


Fig. 3: Results of identifying interactive suggestion regions on ZDT2 problem

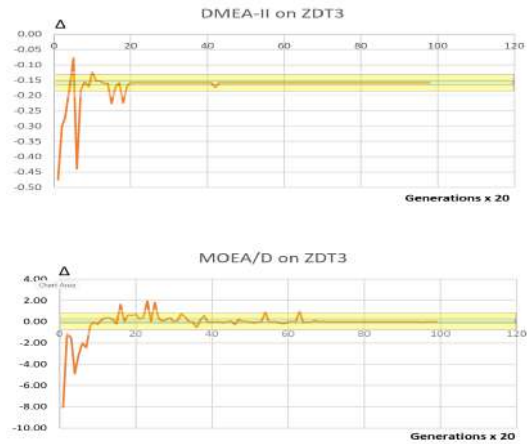


Fig. 4: Results of identifying interactive suggestion regions on ZDT3 problem

ratio with basic ZDT problems reaches a high balance of 0 at the end of the search process, so the suggestion regions are evenly distributed along the zero ratio axis. However, when there is interaction, the Δ ratio will change, the interactive suggestion regions will change; In some cases, with complex problems, the ideal ratio will change, which will determine the determination of the boundary values ϵ_1 , ϵ_2 . But, one thing is clear, by always taking into account the ratio between convergence and diversity measures, suggesting the time of selection for DM interaction will ensure the interaction effect (has a strong effect) while still maintaining the population's exploration and exploitation capabilities, thereby preserving the balance between convergence accuracy and diversity.

ACKNOWLEDGEMENT

The study was funded by National Project No. KC-4.0-53/19-30

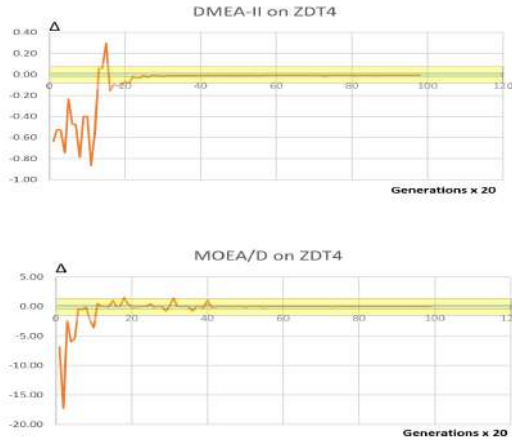


Fig. 5: Results of identifying interactive suggestion regions on ZDT4 problem

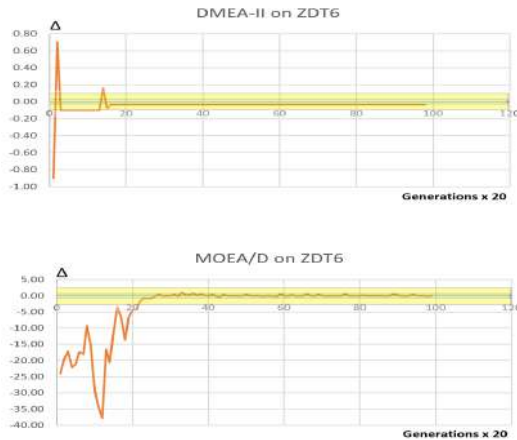


Fig. 6: Results of identifying interactive suggestion regions on ZDT6 problem

V. CONCLUSION

This paper has proposed a novel technique for determining effective interaction timing in interactive multi-objective evolutionary optimization. Recognizing the critical balance between convergence and diversity, we introduced a method that analyzes the generational behavior of population quality using GD and IGD metrics to guide when a decision maker (DM) should interact. By classifying interaction opportunities into green, yellow, and red moments based on the calculated Δ ratio, the framework ensures that user preferences are incorporated at the most beneficial stages of the evolutionary process—enhancing interaction effectiveness while avoiding disruptions to exploration and exploitation dynamics.

Experimental results using the DMEA-II and MOEA/D algorithms on standard ZDT benchmark problems demonstrated that the proposed time suggestion strategy effectively identifies high-impact interaction periods. This method helps to

preserve the algorithm's convergence-diversity balance, minimizes unnecessary or detrimental interactions, and supports better alignment with the DM's evolving preferences.

Overall, this approach contributes a practical and adaptive mechanism to enhance the performance and applicability of interactive evolutionary algorithms. Future work will aim to generalize the technique across more complex problem domains and integrate advanced learning mechanisms for automated interaction timing refinement.

REFERENCES

- [1] Dênis E.C. Vargas, Afonso C.C. Lemonge, Helio J.C. Barbosa, and Heder S. Bernardino. An interactive reference-point-based method for incorporating user preferences in multi-objective structural optimization problems. *Applied Soft Computing*, 165:112106, 2024.
- [2] Cheng Gong, Yang Nan, Tianye Shu, Lie Meng Pang, Hisao Ishibuchi, and Qingfu Zhang. Interactive final solution selection in multi-objective optimization. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–9, 2024.
- [3] Arash Heidari, Sebastian Rojas Gonzalez, Tom Dhaene, and Ivo Couckuyt. Lower confidence bound for preference selection in interactive multi-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 339–342, 2024.
- [4] Babooshka Shavazipour, Jan H Kwakkel, and Kaisa Miettinen. Let decision-makers direct the search for robust solutions: An interactive framework for multiobjective robust optimization under deep uncertainty. *Available at SSRN 4782234*, 2024.
- [5] Alia Youssef Gebreel. An adaptive interactive multi-objective optimization approach based on decision neural network. *International Journal Scientific & Engineering Research (ISSN 2229-5518)*, 7(8):1178–1185, 2016.
- [6] Shubham Agrawal, Yogesh Dashora, Manoj Kumar Tiwari, and Young-Jun Son. Interactive particle swarm: a pareto-adaptive metaheuristic to multiobjective optimization. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, 38(2):258–277, 2008.
- [7] Hui Hou, Ziyin He, Muchao Xiang, Yanchao Lu, Jie Yang, Liang Huang, and Changjun Xie. Interval multi-objective optimization for low-carbon building energy management system upon deep reinforcement learning. *IEEE Transactions on Industry Applications*, pages 1–10, 2025.
- [8] Yating Wang, Zijun Wang, and Hui Wang. A human-machine integrated optimization method for long walkway space. *Journal of Asian Architecture and Building Engineering*, pages 1–24, 2025.
- [9] Ashfaq Ahmed Bhatti, Wei Liu, Mingze Li, Qingan Ma, Qian Xu, Jun Dai, and Xiaodong Zhang. Optimization of train grounding system in dual traction power supply system. *IEEE Transactions on Transportation Electrification*, 2025.
- [10] D.A.V. Veldhuizen. *Multiobjective Evolutionary Algorithms: Classifications, Analyses, and New Innovation*. PhD thesis, Department of Electrical Engineering and Computer Engineering, Airforce Institute of Technology, Ohio, 1999.
- [11] Qingfu Zhang and Hui Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *Evolutionary Computation, IEEE Transactions on*, 11:712 – 731, 01 2008.
- [12] Long Nguyen, Lam T Bui, and Hussein A Abbass. Dmea-ii: the direction-based multi-objective evolutionary algorithm-ii. *Soft Computing*, 18(11):2119–2134, 2014.
- [13] E. Zitzler, L. Thiele, and K. Deb. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation*, 8(1):173–195, 2000.
- [14] Long Nguyen and Lam Thu Bui. A ray based interactive method for direction based multi-objective evolutionary algorithm. In *Knowledge and Systems Engineering*, pages 173–184. Springer, 2014.
- [15] Long Nguyen and Lam Thu Bui. A multi-point interactive method for multi-objective evolutionary algorithms. In *2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 107–112, 2012.

A Balanced Genetic Algorithm for Earliness-Tardiness MS-RCPSP

1st Hao Nguyen Thi

Hanoi National University of Education, Ha Noi; Hanoi National University of Education Thuong Mai University, Ha Noi
Hung Vuong University, Phu Tho
Viet Nam
haont@hvu.edu.vn

2nd Loc Nguyen The*

Ha Noi, Viet Nam

*: Corresponding author

locnt@hnue.edu.vn

3rd Huu Dang Quoc

Ha Noi, Viet Nam

huudq@tmu.edu.vn

Abstract—Scheduling represents the systematic development of execution timetables for resource allocation across project tasks to achieve project objectives. Beyond task completion facilitation, the scheduling process encompasses ensuring adherence to predetermined project deadlines. Deadline compliance constitutes a fundamental requirement for preserving organizational reputation and stakeholder confidence while simultaneously reducing operational costs and mitigating potential risks. Premature completion may incur additional warehouse costs or necessitate waiting for other projects to synchronize. In contrast, excessive tardiness escalates project costs, constitutes contractual violations, and may generate other severe consequential impacts. Based on this reality, this research proposes the ETMS-RCPSP (Earliness Tardiness MS-RCPSP) formulation to minimize the total earliness and tardiness deviations of project tasks relative to their predetermined standard completion deadlines. To solve this problem, the Balanced Genetic Algorithm (BGA) is developed as an enhancement of the conventional Genetic Algorithm (GA) by incorporating a resource allocation strategy for tasks without precedence constraints in task execution. Experimental results reveal that BGA outperforms traditional GA methods, demonstrating the computational efficiency and accuracy of the proposed algorithm. The newly formulated problem and the proposed BGA algorithm will contribute significantly to the construction of execution schedules aimed at achieving project completion within established deadlines.

Index Terms—Earliness-Tardiness scheduling, Genetic Algorithm, Multi-skilled Resource Constrained Project Scheduling Problem.

I. INTRODUCTION

In real-world projects, particularly within manufacturing, logistics, and information technology sectors, beyond achieving project completion within the optimal timeframes and at minimal cost, ensuring on-time project delivery has become a fundamental success determinant. Early completion may result in additional warehousing costs or synchronization waiting periods. At the same time, excessive delays introduce severe risks, including contractual violations, supply chain disruptions, client trust deterioration, and escalated resource expenditures. Major infrastructure initiatives such as the Panama Canal [1], Crossrail railway [2], and Berlin Brandenburg Airport (BER) [3] all experienced multi-year delays and cost overruns exceeding billions of dollars, clearly substantiating the importance of effective project scheduling. Fig.1 below

illustrates the consequences of projects failing to complete within established deadlines.

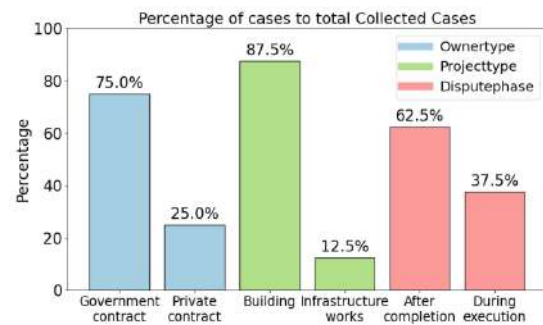


Fig. 1: Demographic analysis for collected cases [4]

The ETMS-RCPSP enables the identification of project execution approaches that minimize temporal deviations between actual completion times and standard completion schedules (planned completion times) through the allocation of multi-skill resources utilized for project tasks. The solution to the problem is a schedule for arranging resources to execute tasks in the project so that the actual completion time is on schedule. The objective of the ETMS-RCPSP problem is to bring many practical applications to solve manufacturing project management problems. Furthermore, it also contributes to improving resource utilization efficiency and minimizing additional costs in industries that require precision in completion timing.

This problem has been proven to be NP-hard, meaning that exact solutions cannot be obtained in polynomial time; therefore, approximation methods are required to find near-optimal solutions within acceptable computational time. This research proposes a solution approach based on the combination of Genetic Algorithm (GA) and Allocate Free Tasks (AFT) technique to generate schedules with optimal deadline adherence.

The primary contributions of this paper include: (i) Proposing a novel scheduling problem called ETMS-RCPSP and formulating ETMS-RCPSP's mathematical model. (ii) Proposing the BGA algorithm, enhanced from the conventional genetic algorithm, where BGA is designed to improve convergence and maintain diversity in the population through problem-

specific operators. (iii) Testing the BGA algorithm on standard iMOPSE benchmark datasets and comparison, statistical analysis, and evaluation to demonstrate the effectiveness and stability of the method.

The rest of the paper is structured as follows. Section II reviews the literature related to ETMS-RCPSP. Section III formulates the ETMS-RCPSP problem model. Section IV describes the proposed BGA algorithm. Section V presents experimental results, evaluation, and performance analysis of the BGA algorithm. Section VI concludes the study and outlines future research directions.

II. RELATED WORKS

In recent years, scheduling problems have attracted considerable attention from researchers worldwide. Vanhoucke and Coelho [5] proposed a method for narrowing the acceptable solution space through problem structure transformation, significantly reducing the number of solutions to be examined while ensuring that optimal solutions are not eliminated.

Some other approaches focus on improving algorithm stability and diversity. Liu et al. [6] proposed the Late-Mover Genetic Algorithm, in which low-quality individuals are retained to prevent premature convergence.

Myszkowski et al. [7]–[9] formulated the MS-RCPSP problem model and proved that the problem belongs to the NP-hard class, while simultaneously proposing metaheuristic algorithms to solve the problem. Furthermore, for algorithm evaluation, the research group also developed the iMOPSE benchmark dataset [10], which has been used in subsequent research studies.

Furthermore, several other studies have addressed the earliness and tardiness of project completion times to enhance the practicality of the problem. Vanhoucke et al. [11] proposed the RCPSPWET problem model as an extended variant of RCPSP. RCPSPWET aims to minimize the total weighted earliness-tardiness cost using search tree algorithms and branch-and-bound methods. Khoshjahan et al. [12] extended this approach and proposed the RCPSP-DET problem by incorporating time-discounted cost factors and solving the problem using metaheuristics. In the healthcare domain, Otten et al. [13] minimized earliness-tardiness costs in multi-machine surgical scheduling to reduce waiting times and overtime for medical staff. Akande [14], [15] provided a more comprehensive perspective by developing three RCPSP models with different E/T objectives: minimizing total tardiness cost, maximizing earliness while simultaneously minimizing tardiness, and finally minimizing a linear objective function combining both earliness-tardiness factors. However, Akande has not yet provided mathematical models for problems involving these two factors.

The studies mentioned above have established a solid theoretical and application foundation for solving scheduling problems with multi-skill resource constraints and earliness-tardiness considerations. However, there remains a lack of effective solutions that can simultaneously address multi-skill resource issues and the objective of temporal deviation from

predetermined standard completion times. This research gap serves as the motivation for our research group to develop the ETMS-RCPSP problem model and propose the BGA algorithm in this paper.

III. PROBLEM FORMULATION

A. Problem definition

In the ETMS-RCPSP problem, a project comprises a finite set of tasks, where each task requires one or more specific skills with defined skill levels. Each task may have precedence constraints, meaning that the task under consideration may need to be executed after one or several other tasks. Tasks that must be completed beforehand are called predecessor tasks or parent tasks, while tasks that can only be executed after their predecessor tasks are completed are called child tasks. Child tasks may have one or more parent tasks. Parent tasks may have one or more child tasks.

Skills considered in the problem include skill type and skill level. Each resource can possess one or more skills. At any given time, a resource can only be allocated to a single task. A resource can execute a task when it has the same skill type and a skill level higher than or equal to the required skill level of the task to be performed. Resources are reusable, meaning that after completing a task, a resource can perform another suitable task.

Unlike previous studies [16]–[18], ETMS-RCPSP provides a scheduling solution that minimizes the total earliness and tardiness deviations of all project tasks from predetermined standard completion deadlines.

The earliness E_i , tardiness T_i for each task i are formally defined as:

$$E_i = \max(0, f_i - h_i), T_i = \max(0, h_i - f_i) \quad (1)$$

where, f_i is the specified due date (deadline) for task i , h_i is the actual completion time of task i .

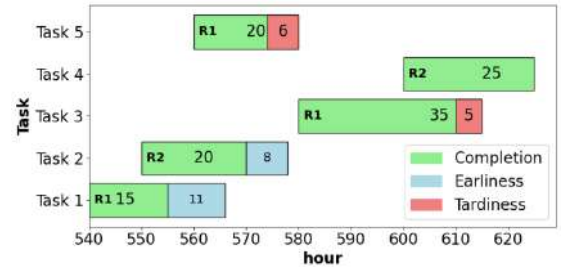


Fig. 2: Task completion with earliness and tardiness

Fig.2 illustrates a project comprising five tasks. Task 4 meets its deadline exactly, while Tasks 1 and 2 are completed early by 11 and 8 hours, respectively. Conversely, Tasks 3 and 5 are delayed by 5 and 6 hours. The total deviation, sum of earliness and tardiness, is 30 hours, representing the objective value to be minimized in earliness-tardiness-based scheduling.

B. Problem formulation

ETMS-RCPSP can be conceptually formulated based on the following notations:

- *Worker*: the set of resources; $Worker_i$: resource i^{th} ;

- $Worker^k$: contains resources that can perform task k ; $Worker^k \subseteq Worker$;
- $EXCE$: the project task set; $Exce_i$: task i^{th} ;
- $Exce^k$: contains tasks, tasks are performed by resource k , $Exce^k \subseteq EXCE$;
- Par_j : the set of the tasks, which need to be completed before the execution of task $Exce_j$ begins.
- $Skill$: the set of the skill; $Skill_i$: skill i^{th} .
- $Skill^i$: set of skills that the resource $Worker_i$ possesses. Clearly, $Skill^i \subseteq Skill$;
- t_{mn} : time for resource $Worker_n$ to perform task $Exce_m$.
- exe : execution time of the schedule
- Req^i : denotes the set of skills required by task $Exce_i$.
- $Beg(Exce_i), End(Exce_i)$: the time when the $Exce_i$ is started and finished;
- $A_{m,n}^j$: a binary decision variable that equals 1 if task $Exce_m$ is being executed by resource $Worker_n$ at time j ; otherwise, it equals 0.
- typ_i, lev_i : the type and required proficiency level of skill $Skill_i$;
- SC_{all} : denotes the set of all feasible schedules.
- $tcount$: number of tasks; $rcount$: number of resources.
- $Eval(SC)$: objective function, the value of this function is the cumulative deviation from the deadlines across all tasks with schedule SC .

$$Eval(SC) = \sum_{i \in EXCE} (e_i.E_i + t_i.T_i) \quad (2)$$

Where e_i and t_i are the penalty units for the earliness and tardiness times of task i , respectively. E_i, T_i are the earliness and the tardiness of task i , respectively and are defined in equation (1).

The goal of solving the ETMS-RCPSP problem is to find the schedule SC such that: $Eval(SC) \rightarrow \min$

A feasible schedule SC is required to satisfy the following set of constraints:

$$Skill^k \neq \emptyset, \forall Worker_k \in Worker \quad (3)$$

$$t_i \geq 0, \forall Exce_i \in EXCE \quad (4)$$

$$End(Exce_i) \geq 0, \forall Exce_i \in EXCE \quad (5)$$

$$End(Exce_i) \leq End(Exce_k) - t \geq 0, \quad (6)$$

$$\forall Exce_k \in EXCE, k \neq 1, Exce_i \in Par_k$$

$$\forall Exce_i \in Exce^k, \exists Skill_q \in Skill^k : \quad (7)$$

$$typ_{Skill_q} = typ_{Req^i}, lev_{Skill_q} \geq lev_{Req^i}$$

$$\forall Worker_k \in Worker, \forall q \in exe : \sum_{i=1}^n A_{i,k}^q \leq 1 \quad (8)$$

$$\forall Exce_i \in EXCE, \exists! q \in [0, exe], \quad (9)$$

$$\exists! Worker_k \in Worker : A_{i,k}^q = 1, with \frac{q}{i,k} \in \{0, 1\}$$

Specifically, the constraints are meant as follows:

- (3): Task assignment is only permitted for resources that hold at least one associated skill.
- (4, 5): The processing time of any task is required to be greater than or equal to zero.
- (6): Parent task i must be completed before its child task k begins.

- (7): For every task $Exce_i$ executable by resource $Worker_k$, there must exist at least one skill $Skill \in Skill^k$ such that: The resource's skill type aligns with the requirement of task $Exce_i$, and its proficiency level is no less than the task's required level.
- (8): A resource can handle at most one task at any time slot q ; thus, for each resource $Worker_k$, the sum of its assignments at time q must be binary: either 0 (idle) or 1 (assigned).
- (9): Every task must be assigned to one, and only one, resource, which is solely responsible for its execution over the entire processing duration.

IV. PROPOSED ALGORITHM

The BGA algorithm is an enhancement of the GA algorithm [19] by integrating a resource allocation strategy for tasks without parent tasks throughout the entire project schedule.

A. AFT technique

Within the ETMS-RCPSP problem formulation, certain tasks exist without parent dependencies, representing tasks free from precedence constraints. These tasks enable independent resource allocation, thereby offering opportunities to streamline the scheduling optimization procedure. The Allocate Free Tasks (AFT) technique is proposed as a pre-processing strategy to leverage tasks without parent tasks. The idea of the technique is to identify all tasks without parent tasks before optimization and allocate them to eligible resources, assuming full skill compatibility. This early resolution helps reduce the complexity of the search space, allowing the optimization algorithm to focus on more constrained parts of the schedule.

Example: Two resources, $Worker_1$ and $Worker_2$, are responsible for executing project tasks. Among these, tasks $Exce_1, Exce_2, Exce_3, Exce_7, Exce_8, Exce_{10}$, and, $Exce_{11}$ are identified as free tasks, meaning they have no predecessor constraints. These tasks are pre-assigned to resources $Worker_1$ and $Worker_2$ as follows: $Worker_1$ executes tasks $Exce_1, Exce_3, Exce_8, Exce_{11}$ and $Worker_2$ executes tasks $Exce_2, Exce_8, Exce_{10}$. Implementing the AFT technique significantly reduces the complexity of the initialization phase. The algorithm achieves faster convergence and improved scheduling quality, especially in contexts where efficient initial population generation is critical. Furthermore, it ensures balanced resource allocation to tasks or minimizes workload imbalance across available resources.

B. Balanced Genetic Algorithm

The Balanced Genetic Algorithm (BGA) represents a hybrid meta-heuristic approach that integrates the classical Genetic Algorithm (GA) with the AFT technique. BGA leverages the exploration strengths of the Genetic Algorithm (GA) while applying simplification strategies from AFT from the initial stage. Subsequently, the evolutionary process is carried out by integrating AFT into the crossover and mutation operations. As a result, BGA achieves superior search efficiency, enhances population quality, and accelerates convergence speed, making it particularly suitable for scheduling problems with multi-skill resource constraints.

The fundamental principle of BGA is to "preserve optimal solutions while evolving sub-optimal ones". After each generation, resource-to-task assignments are evaluated based on the discrepancy between actual completion times and predefined standard deadlines. Evaluation using this deviation assists the algorithm in preserving effective task allocations to resources and directing evolutionary search toward components requiring optimization.

Integrating the AFT technique into GA functions as a pre-processing improvement mechanism to enhance initial population quality and narrow the search space. This approach operates through three fundamental stages:

- Step 1: Calculate the temporal deviation between the actual completion time and the predetermined deadline for each task.
- Step 2: Sort all tasks in ascending order based on their deviation values (determined in Step 1).
- Step 3: Tapping 50% of the tasks with the lowest deviations for preservation in subsequent evolutionary processes, while the remaining tasks are subjected to continued optimization.

TABLE I: Resource allocation for free tasks

Task	Due date	Realized finish time
<i>Excel₁</i>	22	30
<i>Excel₂</i>	26	23
<i>Excel₃</i>	54	54
<i>Excel₇</i>	46	33
<i>Excel₈</i>	60	72
<i>Excel₁₀</i>	90	85
<i>Excel₁₁</i>	70	78

For illustration, Table I shows a subset of free tasks with their due dates and realized finish times. As shown in Fig. 3, tasks *Excel₃*, *Excel₂*, *Excel₁₀*, and *Excel₁* are retained for the next generation (flagged with 1), while the remaining tasks *Excel₁₁*, *Excel₈*, and *Excel₇* are subject to further evolutionary operations (flagged with 0).

Step 1		Step 2		Step 3
Task	Deviation	Task	Deviation	Tagging
<i>Exce₁</i>	8	<i>Exce₃</i>	0	1
<i>Exce₂</i>	3	<i>Exce₂</i>	3	1
<i>Exce₃</i>	0	<i>Exce₁₀</i>	5	1
<i>Exce₇</i>	13	<i>Exce₁</i>	8	1
<i>Exce₈</i>	12	<i>Exce₁₁</i>	8	0
<i>Exce₁₀</i>	5	<i>Exce₈</i>	12	0
<i>Exce₁₁</i>	8	<i>Exce₇</i>	13	0

Fig. 3: The processing steps applied to tasks exhibiting high deviation levels is presented

This paper proposes a BGA algorithm to solve the ETMS-RCPSp problem; its operating steps are presented in Fig.4.

A detailed description of the BGA algorithm is provided in Algorithm 1, where *Eval()*: objective function of ETMS-RCPSp; *CopyMPS()*: copy the individual; *randI(n)* returns a random integer from 0 to $n - 1$. In the proposed BGA algorithm for solving the ETMS-RCPSp problem, the AFT technique is employed to allocate resources to free tasks, while the evolutionary process is carried out only for non-free tasks that have not yet been assigned resources.

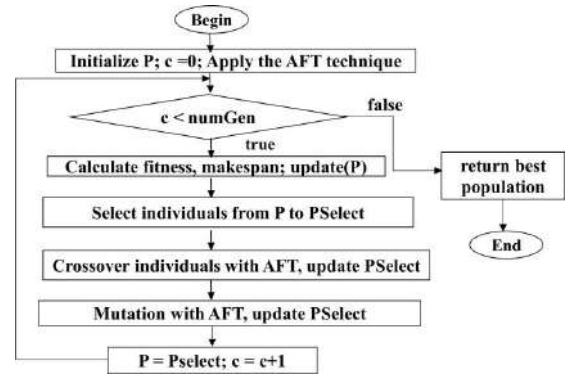


Fig. 4: BGA algorithm for ETMS-RCPSp

Algorithm 1: BGA algorithm

Data: *numGen*: number of generations, *tcount*: number of tasks, *Par_j*: is parent tasks of task *j*; *popsiz*: size of population

Result: the best population

Load and Valid dataset; $c = 0$; $P \leftarrow$ Initial population

for $i < tcount$ **do**

$$Fix[i] = \begin{cases} -1 & \text{if } parent_i \neq \emptyset, \\ \text{assigned resource} & \text{otherwise} \end{cases}$$

Update P base on Fix

while $c < numGen$ **do**

for $i < popsiz$ **do**

$fitness[i] = Eval(P[i])$

if $fitness[i] < oBest$ **then**

$oBest = fitness[i]$; $CopyMPS(P[i])$

for $i < popsiz$ **do**

$i1, i2 = randI(popsiz)$

$$PSelect[i] = \begin{cases} P[i1] & \text{if } fitness[i1] < fitness[i2] \\ P[i2] & \text{otherwise} \end{cases}$$

for $i < popsiz$ **do**

$cp = randI(tcount)$, $ic = randI(popsiz)$

$cathe = PSelect[i]$, $cathelai = PSelect[ic]$

for $k < tcount$ **do**

$$cathe[k] = \begin{cases} Fix[k] & \text{if } Fix[k] > -1 \\ cathelai[k] & \text{otherwise} \end{cases}$$

$PSelect[i] = cathe$

$mutR = rand(popsiz)$

for $i < mutR$ **do**

$icM = randI(popsiz)$, $ctM = PSelect[icM]$

$icTM = randI(tcount)$

for $k < icTM$ **do**

random a task $Exce_{it}$ and a resource
 $Worker_{ir} \in Worker^{it}$

$$ctM[it] = \begin{cases} Fix[it] & \text{if } Fix[it] > -1 \\ ir & \text{otherwise} \end{cases}$$

$PSelect[i] = ctM$

$P = PSelect$; $c = c + 1$

return best population

V. PERFORMANCE ANALYSIS

A. Experimental settings

The proposed algorithm for the ETMS-RCPSP problem was implemented and evaluated using Visual Studio 2022 with the C# programming language. All experimental computations were performed on a Windows 10 system featuring an Intel Core i5 processor operating at 2.4 GHz with 8 GB of RAM. Experiments were conducted using four different generation configurations: 1,000, 3,000, 5,000, and 10,000 generations. For each generation setting, each dataset was executed 10 times. To evaluate the performance and stability of the proposed algorithm, we conducted experiments on the iMOPSE dataset [13], a widely recognized benchmark in research studies for solving MS-RCPSP problems and their variants. iMOPSE's data is stored in the file .def, which includes the following information fields: set of tasks, set of resources, and set of priority relationships between tasks.

B. Experimental results

1) *BGA algorithm*: Experimental evaluation demonstrates that the BGA algorithm exhibits significant improvement throughout the evolutionary procedure. Fig. 5 presents the computational runtime of the BGA algorithm under four generation configurations. Statistical analysis indicates significant

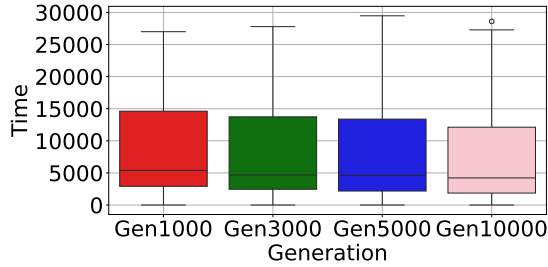


Fig. 5: Experimental results of the BGA algorithm on the iMOPSE dataset

differences in the objective function results of the ETMS-RCPSP problem across different generation configurations. Overall, the BGA algorithm demonstrates a decreasing trend in objective function values as the number of generations increases. The 1,000-generation configuration yields the highest median objective function value (approximately 5,400 time units), while the medians progressively decline to about 4,200 time units at 10,000 generations.

All four generation configurations exhibit considerable variation in results, as reflected by wide interquartile ranges (IQR). The 1,000-generation configuration has the widest IQR (about 11,700 units, from Q1 $\approx$ 2,900 to Q3 $\approx$ 14,600), indicating substantial variability. With higher generation counts, the IQR gradually narrows (to about 10,300 units at 10,000 generations), suggesting improved stability.

A comparative analysis between the 1,000-generation and 10,000-generation configurations reveals enhanced performance, with median values decreasing by approximately 22% (from 5,400 to 4,200) and the third quartile (Q3) decreasing

by about 17% (from 14,600 to 12,100). The first quartile (Q1) also decreases steadily (from about 2,900 to 1,900), reinforcing the observation that both the central tendency and the overall distribution of results improve as the number of generations increases. These findings suggest that the evolutionary optimization process not only enhances average performance but also reduces variability, resulting in more consistent solutions.

2) *Effectiveness Evaluation of BGA versus GA*: This section evaluates the experimental performance of the BGA algorithm against traditional GA. Fig. 6 displays box plot comparisons between BGA and GA results for the ETMS-RCPSP problem.

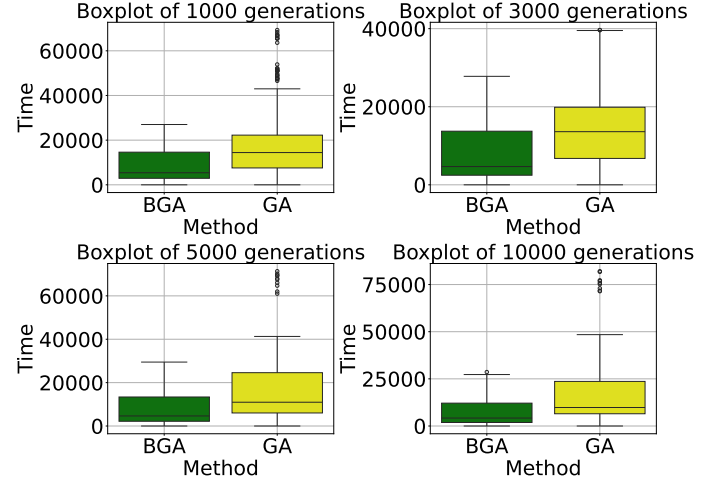


Fig. 6: Comparative plots of BGA and GA

The objective function results across the four generation configurations (1,000, 3,000, 5,000, 10,000) demonstrate that the BGA algorithm exhibits superior and stable performance with the median fluctuating consistently around 4,200–5,400 time units, which is 57–66% lower than the GA algorithm ($\approx$ 9,800–14,400). BGA achieves optimal performance and maintains stability with fewer outliers than GA. Particularly at the 10,000-generation configuration, BGA maintains stable performance with very few outliers, while GA exhibits high variability and poor convergence capability. Through experimentation across the four generation configurations, BGA has proven to be more effective and reliable in optimization processes.

Additionally, experimental evidence reveals stability and reliability properties in the BGA algorithm, displaying diminished variance and uniform performance across all generation configurations, with IQR values fluctuating within the 10,200–11,700 unit range. In contrast, the GA algorithm demonstrates higher variability with IQR extending from 13,100–18,600 units.

The BGA algorithm demonstrates higher predictability due to concentrated data distribution and fewer outliers. GA is unpredictable due to the frequent occurrence of outlier values. Although the median difference between the two algorithms ranges from 57% to 66% depending on the generation configuration, BGA consistently outperforms GA.

Analysis of experimental results across four generation configurations has proven that the BGA algorithm completely surpasses GA on all evaluation criteria. BGA achieves better performance with high stability and good predictability. Particularly, BGA attains optimal efficiency with significant computational resource savings. For optimization problems requiring high reliability and computational efficiency, especially in applications of multi-skill resource-constrained scheduling problems, the BGA algorithm holds advantages over GA in all aspects.

VI. CONCLUSION

This paper presents the novel ETMS-RCPSP problem, formulating ETMS-RCPSP's mathematical model, and proposes the BGA algorithm. To demonstrate the effectiveness of the BGA algorithm, this study conducted experiments on the iMOPSE dataset. Experimental results show that the algorithm achieves good effectiveness, specifically:

At the 1,000-generation configuration, BGA demonstrates a superior advantage with a median of approximately 5,400 time units, 63% lower than GA (approximately 14,400). BGA's interquartile range (2,900–14,600) is significantly narrower than GA's (7,500–22,200), indicating higher stability. Both algorithms have outliers, but GA exhibits more outliers. At the 3,000-generation configuration, BGA's superiority trend continues with a stable median around 4,700, while GA shows slight improvement to approximately 13,600 (a 6% reduction from 1,000 generations). BGA eliminates most outliers at this level, demonstrating high stability, while GA maintains a wide variation range. At the 5,000-generation configuration, BGA maintains stable performance with statistical indicators similar to previous levels. GA shows poor improvement with the median decreasing only to approximately 11,000 and outliers appearing in the 60,000–70,000 range, indicating instability in the convergence process. At the highest generation configuration, BGA exhibits optimal performance with Q1 decreasing to approximately 1,900 and maintaining a median around 4,200. Conversely, GA shows minimal improvement and outliers reaching 77,000–82,000 time units.

The research results evaluate the advantages of BGA over conventional evolutionary methods, establishing new research directions in optimization theory. Simultaneously, the specific formulation of the ETMS-RCPSP problem model has high practical applicability in production management and logistics operations. Projects focus on schedule optimization to achieve on-time completion, reduce incurred costs, and minimize risks. Furthermore, the research findings serve as a foundation for researchers to continue exploring and developing more extensively, including both improved problem models and advanced solution methods. In the future, we will continue researching and developing algorithms for the ETMS-RCPSP problem.

In future work, we will extend our study by comparing BGA with other metaheuristics such as PSO, ACO, Tabu Search, and recent GA variants. We also plan to apply the proposed ETMS-RCPSP model and solution approach to real-world industrial case studies in manufacturing, logistics, and IT

projects, in order to better demonstrate both the effectiveness of the algorithm and the practical relevance of the proposed problem formulation.

REFERENCES

- [1] Sandle P, "London's \$24 billion Crossrail finally opens", 2022. <https://www.reuters.com/world/uk/londons-24-billion-crossrail-finally-opens-2022-05-23/>, Reuters. Accessed June 6, 2025.
- [2] Pretel E, Moreno E, "Panama opens canal extension amid growth risks, cost battle", 2016. <https://www.reuters.com/article/business/panama-opens-canal-extension-amid-growth-risks-cost-battle-idUSKCN0ZC0Q8/>, Reuters. Accessed June 6, 2025.
- [3] Bowlby C, "Berlin Brandenburg: The airport with half a million faults". 2019. <https://www.bbc.com/news/world-48527308>, BBC News. Accessed June 6, 2025.
- [4] Alrasheed KA, Soliman EM, Al-Bader HB, "Delay Dispute Cases: Comparative Analysis and Claimed Value Prediction Model", *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 2024;16(1):04523048. doi: <https://doi.org/10.1061/JLADAH.LADR-1038>
- [5] M. Vanhoucke, J. Coelho, "Reducing the feasible solution space of resource-constrained project instances", *Computers & Operations Research*, vol. 165, 2024.
- [6] Y. Liu, L. Huang, X. Liu, G. Ji, X. Cheng, E. Onstein, "A late-mover genetic algorithm for resource-constrained project-scheduling problems", *Information Sciences*, vol. 642, 2023
- [7] M. Antkiewicz, P.B. Myszkowski, K.Gmyrek, A.Krzeminski, J.L.Calvo-Rolle, "Efficiency of Specialized Genetic Operators in Non-dominated Tournament Genetic Algorithm (NTGA2) Applied to Multi-objective Multi-skill Resource Constrained Project Scheduling Problem", *Advances in Computational Collective Intelligence. ICCCI 2024. Communications in Computer and Information Science*, vol 2166. Springer, Cham, 2024
- [8] M. Antkiewicz, P.B. Myszkowski, "Balancing Pareto Front exploration of Non-dominated Tournament Genetic Algorithm (B-NTGA) in solving multi-objective NP-hard problems with constraints", *Information Sciences*, vol. 667, 2024.
- [9] P.B. Myszkowski, M. Laszczyk, "Investigation of benchmark dataset for many-objective Multi-Skill Resource Constrained Project Scheduling Problem", *Applied Soft Computing*, 2022
- [10] P.B. Myszkowski, M. Laszczyk, I. Nikulin, M. Skowro, "iMOPSE: a library for bicriteria optimization in Multi-Skill Resource-Constrained Project Scheduling Problem", *Soft Computing Journal*, vol. 23, 2019.
- [11] Mario Vanhoucke, Erik Demeulemeester And Willy Herroelen, "An Exact Procedure for the Resource-Constrained Weighted Earliness-Tardiness Project Scheduling Problem", Kluwer Academic Publishers. Manufactured in The Netherlands, 2001.
- [12] Y. Khoshjahan, A.A. Najafi, B.A. Nadjafi (2013), "Resource constrained project scheduling problem with discounted earliness-tardiness penalties: Mathematical modeling and solving pro-cedure", *Computers & Industrial Engineering*, 2013.
- [13] Maarten Otten, Aleida Braaksma, Richard J. Boucherie, "Minimizing Earliness/Tardiness costs on multiple machines with an application to surgery scheduling", *Operations Research for Health Care*, vol. 22, 2019
- [14] S. Akande, G. Ajisegiri, "Three classes of Earliness-Tardiness (E/T) scheduling prob-lems", *Fuel Communications*, vol. 10, 2022.
- [15] S. Akande, Ajisegiri GO (2021), "Single machine multi-criteria scheduling problem:total completion time, maximum lateness, and maximum earliness performance measures", *International Journal of Planning and Scheduling*, vol. 3, 2021.
- [16] A. Christian, S. Demasse, E. Néron, "Resource Constrained Project Scheduling: Models, Algorithms, Extensions and Applications", ISBN 978-1-84821-034-9, 2008.
- [17] A.H. Hosseinian, V. Baradaran, "Detecting communities of workforces for the multi-skill resource-constrained project scheduling problem: A dandelion solution approach", *Journal of Industrial and Systems Engineering*, pp. 72-99, 12.2019.
- [18] A.H. Hosseinian, V. Baradaran, "P-GWO and MOFA: two new algorithms for the MSRCPS with the deterioration effect and financial constraints (case study of a gas treating company)", *Applied Intelligence*, 50, pp. 2151-2176, 2020.
- [19] Binh HTT, Thanh PD (2024), "Evolutionary Computation. Bach Khoa publishing house", ISBN: 978-604-488-368-7.

An integrated workforce planning with LLM support for blood bank operations optimization.

Le Tuong Nguyen

HCMC University of Technology and Education.

International University, VNU-HCMC.

Ho Chi Minh city, Vietnam.

ORCID: 0009-0008-1277-2274

Nguyen Van-Hop

International University, VNU-HCMC

Ho Chi Minh city, Vietnam.

nvhop@hcmiu.edu.vn

Abstract—Efficient staff rostering in blood banks is critical to ensuring continuous, high-quality service and dynamic operational constraints. This study proposes an intelligent scheduling framework that integrates Genetic Algorithms (GA) with adaptive penalty weighting guided by a Large Language Model (LLM). Three variants including plain GA, GA with fixed weights, and weighted GA with LLM support, are evaluated across five realistic scenarios. Experimental results from 20 independent runs per method demonstrate that the proposed approach significantly reduces soft constraint violations, improves fairness, and minimizes average overtime. Paired t-tests confirm the statistical significance of these improvements ($p < 0.05$). The findings highlight the potential of integrating LLM-driven reasoning into evolutionary optimization for robust, context-aware workforce scheduling in healthcare environments.

Keywords—workforce planning, staff rostering, genetic algorithm, LLM, blood bank.

I. INTRODUCTION

Blood is an extremely rare source to serve emergency treatment and save lives. electronic versions of their papers. Blood and blood products, being a fresh product, is a perishable product, play a critical role in healthcare industry. are provided. The process of blood collection, transportation and storage is very important to ensure the best possible blood quantity and quality for transfusion to patients in need. Effective workforce planning for blood banks is inherently challenging due to the significant uncertainty and variability associated with blood supply from voluntary donors and fluctuating hospital demands. This dynamic environment requires staffing to be continually adjusted, informed by accurate forecasts of blood donations and hospital orders. Furthermore, processing blood into specific components involves stringent regulatory compliance, quality assurance protocols, and specialized skills, adding complexity to workforce management, especially during peak periods like weekends or emergencies. Currently, rostering, scheduling, and task assignments rely primarily on managerial experience, making rapid responses to unforeseen changes difficult. Recent studies highlight traditional methods' limitations: exact optimization methods offer accuracy but face scalability issues, [1], [2], [3] and heuristics provide faster results but compromise optimality [4], [5]. Reinforcement Learning approaches offer adaptability and sophisticated decision-making capabilities but require extensive data and computational resources [6], [7]. Large Language Models (LLMs) have recently emerged as supportive tools capable of enhancing heuristics by providing intelligent, language-based guidance. However, LLMs also introduce reliability concerns due to sensitivity to prompts, limited domain-specific knowledge, and their black-box nature [8], [9], [10]. To

overcome these challenges, this study makes the following contributions:

- 1) Develops an advanced optimization approach that integrates heuristic Genetic Algorithms (GA) enhanced by LLM-based intelligent guidance, significantly improving solution quality, computational efficiency, and ensuring responsive practicality under real-time, uncertain, and rapidly changing operational conditions.
- 2) Proposes a dynamic mechanism leveraging LLMs to adjust weights for fitness evaluation, improving the effectiveness of roster selection processes.

The proposed integrated framework effectively addresses uncertainty, variability, and real-time dynamics in blood bank staffing, delivering robust solutions crucial for maintaining efficient and timely blood processing and distribution operations.

The remainder of this paper is structured as follows. Section II presents a comprehensive literature review on workforce planning approaches. Section III details the methodology. Section IV provides the experimental results and discussion. Finally, Section V concludes the paper with key findings and directions for future research.

II. LITERATURE REVIEW

Workforce planning and allocation has been extensively studied due to their critical role in various domains such as healthcare, manufacturing, transportation, etc. Existing literature commonly addresses these problems through either sequential approach, where rostering, scheduling, and assignment are solved in distinct stages [5], [7], [11], [12], or integrated approaches that model all components simultaneously. While sequential methods offer computational simplicity and modular design, integrated frameworks capture interdependencies more effectively, often leading to higher solution quality. Recent advancements in staff rostering have explored a diverse set of methodologies, ranging from exact optimization techniques to heuristics, reinforcement learning (RL), and, more recently, the integration of large language models (LLMs) as intelligent assistants. Among exact methods, Kraul et al. [1] addressed the physician scheduling problem in a German training hospital by developing an integrated mixed-integer programming (MIP) framework that jointly considers rostering, assignment, and scheduling. By embedding a Dantzig-Wolfe decomposition into a branch-and-price algorithm, the model can generate optimal solutions while incorporating realistic constraints such as overnight shifts, break times, and labor regulations. However, despite its precision and practicality, this approach may struggle with computational efficiency when applied to large-scale or real-

time settings due to the complexity of the branch-and-price technique and scenario-based modeling.

Similarly, Cappanera et al. [2] proposed an exact MIP-based approach for the integrated rostering and scheduling of airport ground staff, developed in collaboration with a major airline. The model simultaneously handles shift assignment, task sequencing, and regulatory compliance while incorporating overlapping duties and individual preferences. Tested on real-world data from a large international airport, the method demonstrated solid performance on moderately large instances. Nevertheless, its assumption of deterministic inputs limits its ability to handle real-time disruptions or uncertainties in staff availability and task durations. In contrast, Abdullah et al. [3] focused exclusively on the rostering component by formulating a practical MILP model for anaesthetist scheduling, prioritizing fairness and workload balance. While effective for small- to medium-scale instances, the model does not incorporate scheduling or task assignment, nor does it address dynamic or uncertain conditions.

Heuristic approaches have gained popularity for their ability to handle flexibility and scalability. Schwemmer et al. [4] presented a simulation-based heuristic using genetic algorithms and repair operators to create adaptive workforce rosters in decentralized Industry 4.0 production systems. The method prioritized worker-preferred hours and dynamically fluctuating demand, reflecting the shift toward worker-centric and flexible scheduling frameworks. Similarly, Turhan and Bilgen [5] developed a mat-heuristic for the extended nurse rostering problem that combines integer programming for feasible shift generation with discrete particle swarm optimization to refine shift allocations. In this approach, unit assignments are determined only after shift patterns are fixed, highlighting a two-stage hybrid design.

Reinforcement learning methods have also emerged as promising tools for addressing large-scale, real-time rostering problems. Xu et al. [6] tackled the Empty Container Repositioning (ECR) problem using a pure RL framework enhanced with a multi-weight adaptive reward function, enabling real-time trade-off handling between conflicting objectives. Platten et al. [7] utilized reinforcement learning with graph neural networks (GNNs) to address personnel scheduling as a combinatorial optimization task, though they approached shift matching and roster repair in separate stages. Isakov et al. [8] further advanced RL-based scheduling by introducing an explainable multi-agent framework in healthcare settings, incorporating human-in-the-loop feedback to dynamically guide heuristics for real-time patient-centric scheduling.

Building upon the flexibility of learning-based approaches, recent studies have begun integrating LLMs as supportive agents in workforce planning. Wan et al. [9] demonstrated that LLMs, when combined with prompt engineering and real-time data retrieval, can significantly accelerate surgery scheduling and staff allocation while adapting to unpredictable patient flows without expensive re-optimization. In a different domain, Tang et al. [10] proposed an LLM-assisted reinforcement learning approach for task scheduling in multi-cloud systems, where LLMs generate informed action candidates to guide RL agents, thereby improving adaptability and decision efficiency in large-scale, dynamic environments.

Collectively, these studies illustrate a clear evolution in staff rostering methodologies from exact formulations ensuring optimality, to heuristic and learning-based strategies that embrace flexibility and adaptability. Notably, the emerging use of LLMs highlights a new paradigm in which AI-powered guidance enhances traditional methods, allowing for dynamic response to uncertainty, real-time adjustments, and reduced modeling burden in complex workforce planning scenarios.

III. METHODOLOGY

The methodology section describes the structured approach utilized in addressing workforce planning within a blood bank setting. Given the complexity of staffing operations, characterized by dynamic supply and demand, regulatory compliance requirements, and the necessity for flexibility in real-time and an integrated rostering model is essential.

A. Problem description

The rostering problem examined in this study centers on a blood bank's operational complexities, specifically focusing on the blood processing and distribution department. Effective workforce planning is critical to ensure that appropriately qualified personnel are available at all times to handle essential tasks, including donor screening, blood collection, component processing, storage management, warehouse cleaning and timely distribution to hospitals. Given the stringent medical safety standards, irregular and unpredictable demand patterns, emergency scenarios, and regulatory limits on working hours, the scheduling process requires meticulous balancing to maintain operational continuity and staff well-being.

The department under study consists of 31 employees, encompassing department heads, senior technicians, administrative personnel, quality controllers, and specialist technical staff. Operational workflows involve multiple shifts with differing start, lunch, and end times, including special on-call duties, overtime, and weekend shifts to manage fluctuating blood volumes. Specific tasks are divided among various roles, with clear delineations for responsibilities such as blood reception and processing, irradiation, inspection and labeling, allocation, blood transport, packaging, warehouse cleaning, and dispatch operations. Specially, high-volume tasks such as company dispatches and monthly warehouse cleaning demand significant, coordinated manpower efforts. Scheduling must adhere strictly to several constraints, including mandatory compensatory rest periods after overnight shifts, maternity leave accommodations, balanced overtime allocations, and pre-scheduled departmental activities such as training and meetings. The challenge lies in efficiently integrating these constraints with daily operational requirements and demand forecasts. Consequently, a sophisticated scheduling mechanism is necessary, capable of responding swiftly to real-time changes and uncertainties, thereby ensuring a consistent and reliable supply chain without compromising regulatory compliance or staff health.

B. Proposed model

The proposed framework aims to comprehensively address staff rostering challenges by ensuring adequate, flexible, and robust personnel allocation capable of

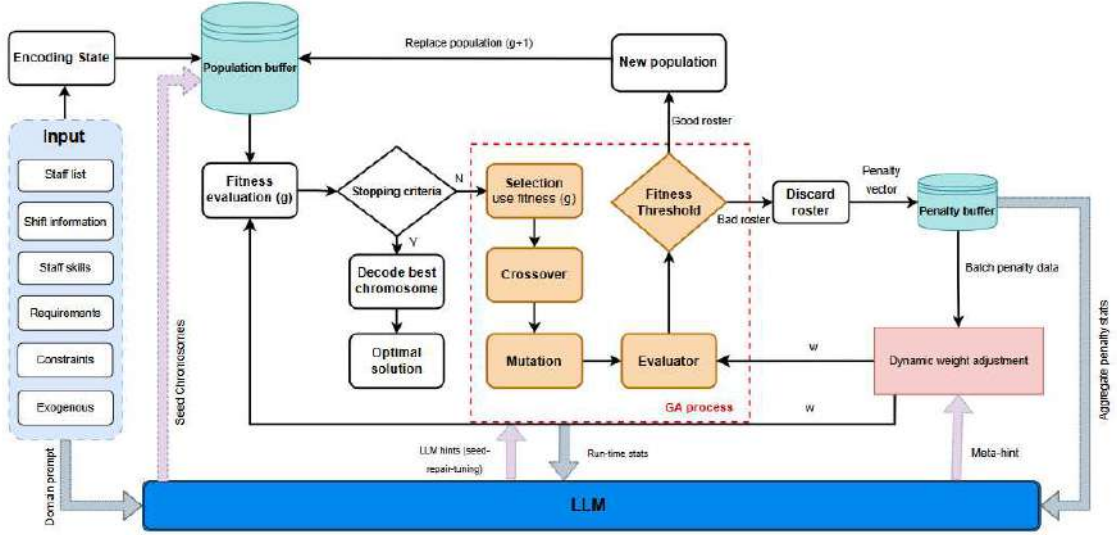


Fig. 1. Overall framework for integrated staff rostering using Genetic Algorithm and LLM support

responding efficiently to real-time dynamics and uncertainty as Fig. 1. At its core, the model begins by encoding input data, including staff lists, shift details, skillsets, requirements, constraints, and exogenous variables, into a structured state representation. This initial encoding populates a buffer, providing the initial set of candidate solutions (chromosomes). A Genetic Algorithm (GA) process is employed to evolve these solutions iteratively. Within this GA loop, chromosomes undergo selection based on fitness criteria, crossover operations to combine advantageous traits, and mutation to introduce diversity and prevent premature convergence. Fitness evaluation continuously assesses candidate solutions, discarding poor-quality rosters based on a dynamically determined fitness threshold, ensuring computational resources are efficiently allocated toward promising solutions. Significantly, a novel feature in this framework is the integration of a Dynamic Weight Adjustment (DWA), which utilizes aggregated penalty data from previously discarded rosters stored in a penalty buffer. This function dynamically adjusts the weights used in evaluating the rosters, enhancing the selection of optimal solutions by capturing and leveraging historical performance data.

The inclusion of Large Language Models (LLMs) provides multi-dimensional support by supplying domain-specific prompts to guide initial chromosome generation, offering runtime statistical insights to assist the GA's fitness evaluations, and delivering meta-hints that continuously update the DWA. By leveraging the interpretative and adaptive capabilities of LLMs, the proposed model achieves a higher level of responsiveness, swiftly accommodating real-time updates, unexpected shifts in operational requirements, and uncertainties inherent in workforce scheduling.

C. Problem formalization

a) Sets and indices:

E: Set of employees, $j \in E$
S: Set of shifts, $s \in S$
T: Set of tasks, $t \in T$
D: Set of planning days, $d \in D$
Q: Set of qualified task-staff pairs
G: Staff gender, $g_j \in \{M, F\}$
 $E_{admin}, E_{tech}, E_{qc}$: Subsets of staff by role/qualification

b) Parameters:

$R_{t,d,s}$: Required number of staff for task t on day d , shift s
 $Cap_{j,t}$: Capacity of staff j to perform task t
 $MaxH_j$: Maximum working hours allowed for staff j in planning period
 OT_j : Overtime limit for staff j
 G_t : Indicator if task t is considered heavy (for gender constraint)
Delayable: Boolean: task t can be delayed within a 2-day window.
Critical: Boolean: task t is critical and must be prioritized.
 $CleanDays \subseteq D$: Days where warehouse cleaning is required.
 $ShipDays \subseteq D$: Days with product shipment.
W: Set of weeks in the planning period, indexed by w .

c) Decision variables:

$x_{j,t,d,s} \in \{0,1\}$: 1 if staff j is assigned to task t on day d , shift s ; 0 otherwise.
 $y_{j,d} \in \{0,1\}$: 1 if staff j is active on day d ; 0 otherwise
 $h_j \in \mathbb{R}^+$: Total working hours assigned to staff j in the planning period.
 $z_j \in \mathbb{R}^+$: Overtime hours assigned to staff j

2) Objective function.

$$\min \sum_{j \in E} \sum_{t \in T} \sum_{d \in D} Penalty_{j,t,d} \cdot Z_{j,t,d} + \sum_k \alpha_k \cdot P_k \quad (1)$$

Objective function is minimize: (1) Penalizes any uncovered shifts to ensure that all required shifts are staffed according to daily demand; and (2) promotes fairness by minimizing the difference between the maximum and minimum total working hours among all employees, thereby distributing workload equitably.

3) Constraints

To ensure the feasibility and realism of generated rosters, both hard and soft constraints were encoded based on operational regulations and domain-specific knowledge.

a) Hard constraints:

C1. Each staff assigned to at most one shift per day

$$\sum_{t,s} x_{j,t,s,d} \leq 1, \forall j, d \quad (2)$$

C2. Task-staff compatibility: $x_{j,t,d,s} = 0$ if $(j, t) \notin Q$ (3)

C3. Minimum staffing per task-shift-day:

$$\sum_j x_{j,t,d,s} \geq R_{t,d,s} \quad \forall t, d, s \quad (4)$$

C4. Rest after 24h shift (S4): no assignment for 1 day after a full-day shift

C5. Task assignment requires activation.

If $x_{j,t,d,s} = 1$, then $y_{j,d} = 1$.

$$\text{Formally: } x_{j,t,d,s} \leq y_{j,d} \quad \forall j, t, d, s \quad (5)$$

C6. Fixed-role task assignment.

C7. Only one early morning shift allowed per day:

$$\sum_j x_{j,T8,d,1} = 1 \quad \forall d \quad (6)$$

C8. ≥ 8 staff during cleaning days:

$$\sum_{j,t,s} x_{j,t,d,s} \geq 8 \quad \forall d \in \text{CleanDays} \quad (7)$$

C9. Reduce QC staff on heavy handover/cleaning days:

$$\sum_{j \in E_{qc}} x_{j,t,d,s} \leq \theta_{qc} \quad \forall d \in \text{CleanDays} \quad (8)$$

C10. Reduce QC staff on heavy handover/cleaning days:

$$\sum_{j,t,s} x_{j,t,d,s} \geq \gamma \quad \forall d \in \text{CleanDays} \quad (9)$$

C11. On-call constraint: specific staff must appear once every 7 or 15 days.

C12. Only cleaning tasks allowed during warehouse cleaning (18-20th)

$$x_{j,t,d,s} = 0 \quad \forall t \notin \text{CleanTasks}, d \in \{18,19,20\} \quad (10)$$

C13. On holidays: only 4 staff allowed, and only for allocation/distribution:

$$\sum_{j,t,s} x_{j,t,d,s} \leq 4, \quad \forall d \in \text{HolidayDays} \text{ and } x_{j,t,d,s} = 0, \forall t \notin \{\text{Allocation, Distribution}\} \quad (11)$$

C14.1. Exactly 6 staff do max 2 evening shifts/week (shit 2):

$$\sum_{d \in \text{Week}(w)} \sum_t x_{j,t,d,2} \leq 2, \forall j \in E_{\text{evening}6} \quad (12)$$

C14.2. Other staff do max 1 evening shift/week:

$$\sum_{d \in \text{Week}(w)} \sum_t x_{j,t,d,2} \leq 1, \forall j \notin E_{\text{evening}6} \quad (13)$$

C15. Weekend: minimum staff by shift based on blood input:

$$\sum_{j,t} x_{j,t,d,s} \geq \phi_{d,s}, \forall d \in \text{Weekend}, s \quad (14)$$

C16. Overtime only allowed for specific staff:

$$z_j = \max(0, h_j - \text{Max}H_j) \quad (15)$$

$$z_j \leq OT_j, \forall j \in E_{OT} \quad (16)$$

$$z_j = 0, \forall j \notin E_{OT} \quad (17)$$

C17. Task capacity requirement_Pt,d: production requirement for task t on day d.

$$\sum_j \text{Cap}_{j,t} \cdot x_{j,t,d,s} \geq P_{t,d}, \forall t, d \quad (18)$$

C18. Task must finish before cleaning/shipping days:

$$\sum_j x_{j,t,d,s} = 0, \forall t \in T, d \in D_{\text{before cleaning/shipping}} \quad (19)$$

C19. Warehouse cleaning constraints (male-only, reduces QC, fixed capacity):

- Only male staff for heavy cleaning tasks
 - Reduce QC capacity
 - Fixed min/max cleaning staff during cleaning days
- $$x_{j,t,d,s} = 0 \text{ if } g_j = F, t \in \text{HeavyClean}, d \in \text{CleanDays} \quad (20)$$

b) Soft constraints:

SC1: Penalty for assigning female staff to heavy physical tasks

SC2: Penalty for exceeding legal overtime thresholds

SC3: Penalty for large deviation in total working hours between staff

SC4: Penalty if low-skilled staff are assigned to critical tasks

SC5: Penalty for assigning packaging tasks to staff not in volunteer list.

SC6: Penalty for scheduling staff on too many weekend shifts (Sat–Sun)

SC7: Penalty for uneven distribution of tasks and shifts over time

D. Model parameters

1) Chromosome representation

Each chromosome encodes a full-week schedule for all staff. Its structure can be represented as a 3D matrix:

$$\text{Chromosome}_{j,t,d} \in \{0,1\} \quad (21)$$

Each element indicates whether staff j is assigned to task t on day d.

2) Fitness function

The fitness function is designed to minimize violations and promote equitable, feasible rosters:

$$\text{Fitness} = -(\alpha \cdot V_h + \sum_{i=1}^{n_{sc}} \omega_i \cdot V_{sci}) \quad (22)$$

where:

V_h : number of hard constraint violations,

V_{sci} : violation magnitude of soft constraint I,

ω_i : penalty weight for soft constraint I,

α : large constant to heavily penalize hard violations.

3) Scenario definition

To test real-world robustness, five dynamic operational contexts were modeled:

Flu_outbreak	High absenteeism due to health issues
Audit_mode	Increased compliance and workload transparency
High_demand	Surge in blood processing and distribution
Staff_leave	Reduced staff availability
Female protection	Policy enforcing gender sensitivity on tasks

These scenarios directly influence how LLM reweights soft constraint importance, simulating adaptive human reasoning.

4) Context-sensitive re-weighting with LLM

To facilitate dynamic optimization, each operational context is associated with a unique vector of soft constraint weights, generated by a LLM based on natural-language descriptions. For example, under the flu outbreak context, the model emphasizes volunteer task fulfillment (SC5) and weekend coverage (SC6) and, whereas in audit mode, fairness (SC3) and compliance with overtime regulations (SC2) are prioritized. This context-to-weight mapping allows the GA to optimize schedules in a way that is interpretable, explainable, and aligned with real-world operational goals.

The LLM module receives a structured context prompt, and returns a modified vector of weights:

$$\text{Weight vector}_{\text{context}} = \text{LLM}(\text{context}) \quad (23)$$

5) Processing pipeline

To evaluate the effectiveness of the proposed rostering framework, three distinct methods were compared: (1) Plain GA, a baseline Genetic Algorithm without consideration of penalty weights; (2) GA+Weight, a Genetic Algorithm incorporating a fixed vector of penalty weights for addressing soft constraint violations; and (3) GA+Weight+LLM, a GA enhanced by adaptive, context-sensitive penalty weights dynamically guided by a LLM. The Plain GA serves as a baseline approach, where fixed weights are assigned to soft constraints, and solutions are evaluated using these fixed penalty values throughout. Standard genetic operators, including selection, crossover, and mutation, are applied iteratively over 50 to 100 generations, after which the best-

performing roster is selected as the final solution. Then, the GA+Weight method enhances this approach by incorporating a manually tuned vector of penalty weights $[w_1, w_2, \dots, w_7]$ to better address soft constraint priorities. Apart from this adjustment, it maintains the same procedural steps as the Plain GA. In contrast, the GA+Weight+LLM method integrates LLM to dynamically adjust soft constraint weights according to specific operational scenarios. At the beginning of each run, the LLM module analyzes the given scenario context and returns tailored penalty weights. These scenario-specific weights guide the genetic algorithm's evolutionary process, allowing the final solution to more accurately reflect operational priorities and context-sensitive requirements.

All experiments were executed using Python 3.11 on a standard desktop (Intel i7, 64GB RAM), with OpenAI GPT-3.5 Turbo API integrated for automated weight generation in the LLM-assisted method.

IV. RESULTS AND DISCUSSION

With described processing pipeline, each method underwent evaluation over 20 independent runs within five realistic operational scenarios. Performance was quantitatively assessed through four key metrics: (1) Soft Violations, representing the cumulative penalty scores from unmet soft constraints (with lower values indicating better performance); (2) Hard Violations, reflecting the frequency of breaches in mandatory constraints (ideally zero); (3) Fairness, measured as the standard deviation of workload distribution among staff members; and (4) Average overtime, indicating the mean overtime hours per staff member.

TABLE 1. SUMMARY OF AVERAGE PERFORMANCE ACROSS METHODS OVER 20 INDEPENDENT RUNS

Method	Hard violations mean	Soft violations mean	Average overtime mean	Fitness mean
Plain GA	51	0	65.89	9.61
GA+Weight	0	518.2	67.49	8.62
GA+Weight+LLM	1.1	485.05	51.73	8.25

As shown in Table 1, Plain GA tends to violate hard constraints despite achieving zero soft penalties, whereas GA+Weight and GA+Weight+LLM almost eliminate hard violations by trading off soft penalties. GA+Weight+LLM further reduces average overtime and achieves the lowest fitness score, indicating superior overall solution quality. In terms of stability, Plain GA and GA+Weight exhibit low variability across 20 runs (standard deviations are typically under 5% relative to the mean), while GA+Weight+LLM shows higher variance, particularly in fairness and overtime

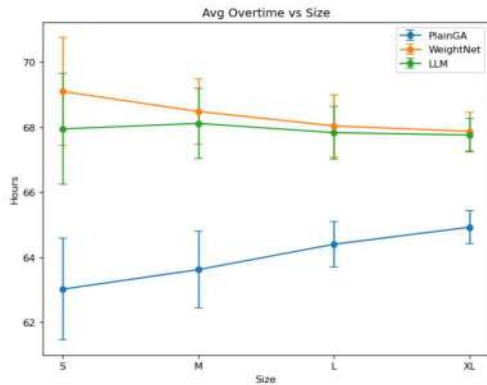


Fig.2. Average overtime comparison across methods

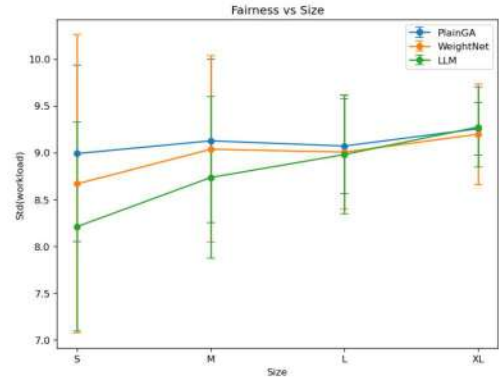


Fig.3. Fairness comparison across methods

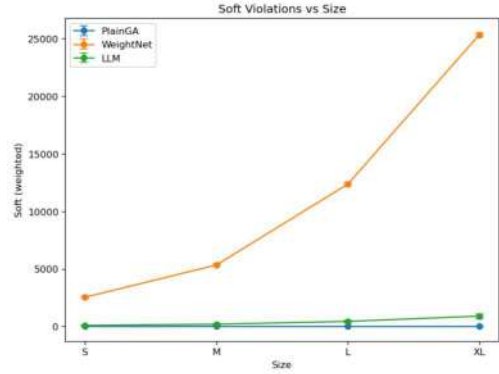


Fig.4. Soft violation comparison across methods

(approximation 1–10%). This reflects the stochastic nature of context-driven weight generation that although GA+Weight+LLM improves average performance, individual runs may differ more strongly, suggesting that robustness could be enhanced through ensemble runs or multiple-seed averaging.

To further assess scalability, we scaled the problem size by multiplying staff and task counts while keeping the day/shift grid unchanged. Four levels were considered: S (baseline), M ($\sim 1.5\times$ staff/tasks), L ($2\times$), and XL ($3\times$). Each configuration was also evaluated over 20 independent runs, allowing us to observe how solution quality and runtime evolve as the problem grows. In terms of overtime in Fig.2, Plain GA consistently achieves the lowest values across all sizes, whereas GA+Weight and GA+Weight+LLM incur 3–5% higher. Conversely, in terms of fairness in Fig.3, GA+Weight and GA+Weight+LLM outperform Plain GA substantially at small and medium sizes (S, M), producing more balanced workloads. At larger sizes (L, XL), the fairness

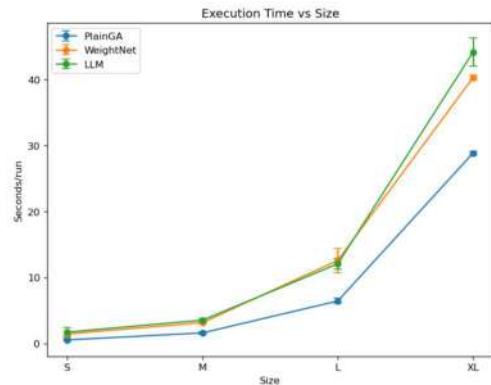


Fig.5. Execution time comparison across methods

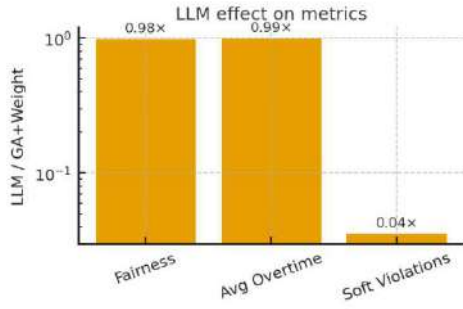


Fig. 6. Effect of LLM-driven weights relative to GA+Weight (ratio LLM/Weight; lower is better)

gap narrows and methods converge, suggesting that the GA search space smooths out as scale increases. Fig. 4 reports the trend of soft violations with increasing problem size. Plain GA consistently records zero soft penalties, but only by tolerating frequent hard violations as discussed earlier in Table 1. GA+Weight, in contrast, incurs substantial soft penalties that grow sharply with scale, indicating that fixed weight vectors amplify certain preference-related terms under larger workloads. GA+Weight+LLM, however, maintains much lower and more stable soft penalties across all sizes, reflecting its ability to contextually down-weight non-critical soft terms when they conflict with feasibility or fairness. Combined with the fact that hard violations remain near zero for both GA+Weight and GA+LLM (hard violations are not plotted, as they remain close to zero across all methods and sizes), these results confirm that the LLM mechanism achieves feasibility while effectively suppressing soft penalty accumulation at scale. Finally, Fig. 5 presents runtime growth. All methods show near-linear scaling with problem size. Plain GA is consistently fastest, GA+Weight requires $\sim 1.4\text{--}2\times$ more time, and GA+Weight+LLM incurs the highest overhead ($\sim 1.5\times$ Plain GA at XL) due to weight generation. Nevertheless, runtimes remain operationally acceptable for practical deployment.

In terms of complexity, all GA variants are dominated by fitness evaluations. With population size P and number of generations G , the overall cost scales as $O(P \times G)$. Dynamic weights add only a minor $O(P)$ overhead per generation, while GA+Weight+LLM introduces a constant-factor increase from weight generation ($\approx 1.5\times$ at XL), without altering the asymptotic complexity. Thus, runtimes remain practically feasible. Beyond runtime, Fig. 6 shows that LLM-driven weighting reduces soft penalties to only $0.04\times$ of GA+Weight, while fairness ($0.98\times$) and overtime ($0.99\times$) remain essentially unchanged. This demonstrates that the LLM reallocates weights to suppress non-critical soft terms without affecting core performance.

Statistical analysis using paired t-tests at $\alpha = 0.05$ confirmed significant performance differences among the rostering methods. The GA+Weight+LLM method outperformed others across all key metrics, reducing soft violations by 425.2 points ($p < 0.00001$), improving fairness and average overtime ($p < 0.005$), and maintaining feasibility with minimal hard violations. These results validate the effectiveness of using dynamic, scenario-specific weights via the LLM module to enhance roster quality.

V. CONCLUSION

This study presents a context-aware rostering framework that integrates genetic algorithms with adaptive penalty weighting

guided by OpenAI LLM via real-time API calls. Across multiple realistic operational scenarios, the GA+Weight+LLM method significantly outperformed baseline approaches in soft constraint compliance, fairness, and workload balancing, while maintaining feasibility.

By using the actual OpenAI API, this research demonstrates the feasibility of embedding large language models into operational scheduling workflows. Nonetheless, potential limitations remain, such as response latency, API stability, and dependency on prompt formulation. Future work may explore prompt optimization, caching strategies to reduce overhead, and integrating user feedback to refine LLM outputs over time, further enhancing the robustness and adaptability of the system.

ACKNOWLEDGMENT

The authors acknowledge the support of HCMC University of Technology and Education, and Intenational University, VNU-HCMC for this study.

REFERENCES

- [1] Sebastian Kraul, Melanie Erhard, Jens O. Brunner, "Optimizing physician schedules with resilient break assignment," *Omega* 129 – 103154, 2024.
- [2] Paola Capanera, Leonardo Di Gangi, Matteo Lapucci, Giulia Pellegrini, Marco Roma, Fabio Schoen, Alessio Sortino, "Integrated task scheduling and personnel rostering of airports ground staff: A case study," *Expert Systems With Applications* 238 – 121953, 2024.
- [3] Norizal Abdullah, Masri Ayob, Meng Chun Lam, Nasser R. Sabar, Graham Kendall, and Liu Chian Yong, "Advancing Anaesthetist Rostering Quality: A Practical Approach Toward Fairness and Efficiency," *IEEE Access*, Volume 13, 2025.
- [4] Schwemmer, Julia Günsel, Christian Kühn, Mathias Schmidt, Thorsten, "Workforce rostering for decentrally controlled production systems: A simulation-based optimization framework using a genetic algorithm," *Logistics Research*, ISSN 1865-0368, Bundesvereinigung Logistik (BVL), Bremen, Vol. 16, Iss. 1, pp. 1-26, 2023.
- [5] Aykut Melih Turhan, Bilge Bilgen, "A mat-heuristic based solution approach for an extended nurse rostering problem with skills and units," *Socio-Economic Planning Sciences* 82 – 101300, 2022.
- [6] Xiaofeng Xu, Xinru Huang, Lianju Wang, "Empty container repositioning problem using a reinforcement learning framework with multi-weight adaptive reward function," *Maritime Policy & Management*, 2024.
- [7] Benjamin Platten, Matthew Macfarlane, David Graus and Sepideh Mesbah, "Automated Personnel Scheduling with Reinforcement Learning and Graph Neural Networks," *The 2nd Workshop on Recommender Systems for Human Resources*, in conjunction with the 16th ACM Conference on Recommender Systems, September 18–23, 2022, Seattle, USA, 2022.
- [8] Artem Isakov, Danil Peregorodiev, Ivan Tomilov, Chuyang Ye, Natalia Gusarova, Aleksandra Vatian and Alexander Boukhanovsky, "Real-Time Scheduling with Independent Evaluators: Explainable Multi-Agent Approach," *Technologies*, 12, 259, 2024.
- [9] Fang Wan, Kezhi Wang, Tao Wang, Hu Qin, Julien Fondreville, Antoine Duclos, "Enhancing healthcare resource allocation through large language models," *Swarm and Evolutionary Computation* 94 – 101859, 2025.
- [10] Xuhao Tang, Fagui Liu, Dishu Xu, Jun Jiang, Quan Tang, Bin Wang, Qingbo Wu, and C.L. Philip Chen, "Assisted Reinforcement Learning: Leveraging Lightweight Large Language Model Capabilities for Efficient Task Scheduling in Multi-Cloud Environment," *IEEE Transactions on Consumer Electronics*, 2024.
- [11] Philippe Racette, Frédéric Quesnel, Andrea Lodi, François Soumis, "Gaining insight into crew rostering instances through ML-based sequential assignment," *Springer, TOP* 32:537–578, 2024.
- [12] Aymen Ben Said and Malek Mouhoub, "Machine Learning and Constraint Programming for Efficient Healthcare Scheduling," *arXiv:2409.07547v1 [cs.AI]*.

Efficient Mining of Frequent High Average-Utility Occupancy Sequences

Tien Hoang

Faculty of Information
Technology, University of
Science, Ho Chi Minh City
Vietnam National University,
Ho Chi Minh City, Vietnam
Department of Mathematics
and Computer Science, Dalat
University, Vietnam
tienhm@dlu.edu.vn

Hai Duong*

Department of Mathematics
and Computer Science, Dalat
University, Vietnam
haidv@dlu.edu.vn

Tin Truong

Department of Mathematics
and Computer Science, Dalat
University, Vietnam
tintc@dlu.edu.vn

Bac Le

Faculty of Information
Technology, University of
Science, Ho Chi Minh City
Vietnam National University,
Ho Chi Minh City, Vietnam
lhbac@fit.hcmus.edu.vn

Abstract—Frequent High Utility Occupancy Sequence Mining (FHUOSM), which integrates both frequency and utility occupancy, has recently become a key research focus with broad real-world applicability. However, existing FHUOSM approaches often generate excessively large result sets, hindering effective analysis and decision-making. To overcome this limitation, the Average-Utility Occupancy (AUO) function, accounting for pattern length, has been introduced as a more informative criterion. This paper explores the problem of discovering frequent high average-utility occupancy patterns. (FHAUOSM). A key difficulty arises from the fact that the AUO function does not possess the anti-monotonicity (AM) property, which is essential for effective search-space pruning. To address this issue, two new bounds are introduced, each formulated to satisfy the AM property and enable efficient pruning mechanisms. Using these theoretical contributions, we design a new algorithm called FHAUOS-Miner to discover FHAUOSs efficiently. Experimental evaluations on both real-world and synthetic quantitative sequence datasets show that FHAUOS-Miner notably surpasses the baseline approach in runtime efficiency, memory usage, and scalability.

Keywords—high utility sequence, Utility mining, weak upper bound, upper bound, pruning strategy.

I. INTRODUCTION

High Utility Pattern Mining (HUPM) over Quantitative Sequence Databases (QSDBs) has become a prominent research topic in recent years, driven by its applicability across various domains such as web usage analysis [1], mobile commerce [2], healthcare cost-event analysis [3], and gene regulation [4]. In HUPM, each item has a *quantity* and a *quality* value, which together determine the item's utility. A central challenge in HUPM on QSDBs is that a sequence can occur multiple times within a single sequence, at different positions and with varying utilities. Consequently, different utility aggregation methods have been proposed, including the maximum utility function, the minimum utility function, and the balanced utility function [5][6].

Related work. Unlike traditional frequency-based measures, utility measures do not satisfy the AM property, an essential characteristic often leveraged to effectively discard the unpromising candidates. To overcome this limitation, researchers have utilized upper bounds (UBs) that preserve anti-monotonicity and enable efficient pruning. For the *maximum utility function*, commonly used bounds include SWU [5], PEU, and SEU [7]. For the *minimum utility function*, UBs such as RBU and LRU have been proposed [8]. More recently, the *balanced utility function* has been supported by the development of UBs in [6], designed to facilitate efficient mining under this newer utility measure.

One notable limitation of HUPM is its tendency to generate a large number of overly long patterns. This occurs because extensions of high utility patterns often also yield high utility,

resulting in verbose outputs that are difficult for users to interpret and analyze. To overcome this limitation and introduce a more balanced utility evaluation, researchers have introduced the *average-utility* (AU) measure, which accounts for pattern length. Specifically, the AU value of a sequence is obtained by dividing its total utility by the number of items it contains. This refined measure has been applied to sequences in QSDBs [9], leading to the formulation of the problem of mining *High AU Patterns* (HAUPM). However, as with conventional utility measures, the AU function fails to satisfy AM. As a result, several UBs have been developed to enable effective pruning, leveraging anti-monotone-like or anti-monotonic properties, as demonstrated in [6].

While HAUPM offers notable advantages over HUPM, one key limitation is that the discovered patterns do not inherently capture the utility or average utility *occupancy* of a pattern within each supporting sequence. To address this gap, the concept of *occupancy* was introduced [10]. UO measures the proportion of a pattern's utility relative to the total utility of the transaction in which it appears, thereby indicating the pattern's significance and representativeness within individual sequences. This concept has demonstrated its usefulness in a variety of real-world applications [10]. In HUPM, utility occupancy has been leveraged in several extensions, including the discovery of high utility occupancy patterns [11]. More recently, Kumar et al. proposed the notion of average-utility occupancy (AUO), which combines AU with occupancy, leading to the definition of the Frequent High AU Occupancy Itemset discovering problem [12].

However, to the best of our understanding, no prior work has addressed *High Average-Utility Occupancy Sequence Mining* (HAUOSM) in QSDBs. To address this research gap, the current study presents new theoretical foundations, and proposes an efficient algorithm for HAUOSM.

Contributions. The main contributions of this study are outlined as follows. Firstly, it proposes a novel Average-Utility Occupancy (AUO) measure for sequential patterns, effectively overcoming the major shortcomings of existing high utility occupancy sequence mining approaches. Secondly, it proposes two theoretical bounds, an upper bound (*ubauos*) and a weak upper bound (*wubauos*), accompanied by effective techniques to early discard unpromising candidates. Thirdly, it presents an efficient algorithm, FHAUOS-Miner, specifically designed to exploit frequent high AU occupancy sequences. Comprehensive experiments on several datasets demonstrate that the proposed algorithm consistently surpasses the baseline method with respect to runtime, memory efficiency, and the number of generated candidates.

The paper is structured in the following manner. Section II introduces the essential preliminaries and definitions. In Section III, the proposed UBs, pruning mechanisms, and the FHAUOS-Miner framework are elaborated. Section IV presents and interprets the experimental results. Lastly, Section V offers the conclusion and highlights prospective research directions.

II. PRELIMINARY CONCEPTS

For $\mathcal{A} = \{a_1, a_2, \dots, a_M\}$ being a collection of items, then an event (or itemset) $E \subseteq \mathcal{A}$ is written in lexicographical order $<$ (e.g., $\{a, b, d\}$ as abd). A sequence $\alpha = \langle E_1, E_2, \dots, E_p \rangle$ is referred to as an ordered collection of events $E_k, k = 1, \dots, p$. Each item a has an external utility $p(a) \in R_+$. A pair (a, q) is called a quantitative-item (denoted as q -item), where q is internal utility. A q -itemset E' based on E is a collection of q -items, $E' \stackrel{\text{def}}{=} \{(a_i, q_i) \mid a_i \in E, q_i \in R_+\}$, written as $[(a_i, q_i) \mid a_i \in E]$ with $E = \text{proj}(E')$ as its projected itemset. For brevity, a single q -itemset $[(a_i, q_i)]$ is simply written as (a_i, q_i) . A q -sequence $\alpha' = \langle E'_1, E'_2, \dots, E'_p \rangle$, has length $\text{length}(\alpha') \stackrel{\text{def}}{=} \sum_{k=1..p} |E'_k|$ and size $\text{size}(\alpha') \stackrel{\text{def}}{=} p$, where the element count of E'_k is represented by $|E'_k|$. The null q -sequence, denoted as $<>$, satisfies $\text{size}(\alpha') = 0$. A quantitative sequence database (QSDB), denoted as $\mathcal{D}' = \{\Psi'_i, i = 1, \dots, N\}$, includes a set of input q -sequences, each associated with a unique sequence identifier (SID). The projected sequence of α' is $\alpha = \text{proj}(\alpha') \stackrel{\text{def}}{=} \langle \text{proj}(E'_1), \text{proj}(E'_2), \dots, \text{proj}(E'_p) \rangle$.

Definition 1 (Average utility and Utility [13], [14]).

- The utilities of a q -item (a, q) , q -itemset $E' = \langle (a_{i_1}, q_{i_1}), \dots, (a_{i_m}, q_{i_m}) \rangle$, q -sequence $\alpha' = \langle E'_1, E'_2, \dots, E'_p \rangle$ and $\mathcal{D}'$ are referred to as $u((a, q)) \stackrel{\text{def}}{=} p(a) \times q$, $u(E') \stackrel{\text{def}}{=} \sum_{j=1..m} u((a_{i_j}, q_{i_j}))$, $u(\alpha') \stackrel{\text{def}}{=} \sum_{i=1..p} u(E'_i)$, $u(\mathcal{D}') \stackrel{\text{def}}{=} \sum_{\Psi' \in \mathcal{D}'} u(\Psi')$, respectively.
- The average-utility of α' is computed as $au(\alpha') \stackrel{\text{def}}{=} u(\alpha') / \text{length}(\alpha')$, where $\text{length}(\alpha')$ denotes the total number of q -items in α' .

To avoid redundant utility computations, the utility can be computed once using $u((a, q)) = p(a) \times q$, after which q is replaced by this value. This transformation yields an integrated QSDB derived from $\mathcal{D}'$. For brevity, we will refer to this integrated form simply as $\mathcal{D}'$ throughout the paper. As an example, Table 1 provides an integrated QSDB used as the illustrative example throughout the paper.

For any two q -sequences $\alpha' = \langle E'_1, E'_2, \dots, E'_p \rangle$ and $\beta' = \langle F'_1, F'_2, \dots, F'_q \rangle$, then $\alpha = \langle E_1, E_2, \dots, E_p \rangle$ and $\beta = \langle F_1, F_2, \dots, F_q \rangle$ are respectively projected sequences of α' and β' .

TABLE 1. INTEGRATED QSDB $\mathcal{D}'$.

	Input q -sequence
Ψ'_1	$<[(a,11)(b,7)],(c,15),(f,35),(g,30),[(b,8)(c,6),(e,18)]>$
Ψ'_2	$<(b,8), [(c,18)(e,10)], (a,40), [(a,20)(d,15)]>$
Ψ'_3	$<[(c,74)(d,64)],[(b,8)(e,45)],[(b,7)(e,12)],[(b,35)(c,18)(d,6)(e,20)]>$

Definition 2 (Extensions [13], [14]). The s -extension and i -

extension of a sequence α with β are defined respectively as $\alpha \diamond_s \beta \stackrel{\text{def}}{=} \langle E_1, E_2, \dots, E_p, F_1, F_2, \dots, F_q \rangle$ and $\alpha \diamond_i \beta = \langle E_1, E_2, \dots, (E_p \cup F_1), F_2, \dots, F_q \rangle$, where for each element a in E_p and b in F_1 , it holds that $a < b$. A forward extension (FE) of α by β , $\gamma = \alpha \diamond \beta$, may correspond to be $\alpha \diamond_s \beta$ or $\alpha \diamond_i \beta$.

Definition 3 (Partial order relations). Let $E' = [(a_{i_1}, q_{i_1}), \dots, (a_{i_m}, q_{i_m})]$ and $F' = [(a_{j_1}, q_{j_1}), \dots, (a_{j_n}, q_{j_n})]$, where $m \leq n$. We say that E' is a sub q -itemset of F' , written as $E' \subseteq F'$, if there exist positive integers $1 \leq k_1 < k_2 < \dots < k_m \leq n$ such that for every $l \in \{1, \dots, m\}$, both $q_{i_l} = q_{j_{k_l}}$ and $a_{i_l} = a_{j_{k_l}}$ hold. Based on this definition, for any two q -sequences α' and β' , α' is included in β' , written as $\alpha' \sqsubseteq \beta'$ (or equivalently, β' is a super q -sequence of α'), if $p \leq q$, and there exist indices, $1 \leq j_1 < j_2 < \dots < j_p \leq q$ satisfying $E'_k \subseteq F'_{j_k}$, for all $k = 1, \dots, p$. Moreover, $\alpha' \subset \beta' \Leftrightarrow (\alpha' \sqsubseteq \beta' \wedge \alpha' \neq \beta')$. Similarly, α is considered to be included in β , written as $\alpha \sqsubseteq \beta$, if there exist p positive integers, $1 \leq j_1 < j_2 < \dots < j_p \leq q$ such that $E_k \subseteq F_{j_k}$, for all $k = 1, \dots, p$. When $\alpha \sqsubseteq \beta$ and $\alpha \neq \beta$, we denote $\alpha \subset \beta$. β' contains α (or equivalently, β' is a super q -sequence of α), if $\text{proj}(\beta') \supseteq \alpha$; this relation is denoted by $\alpha \sqsubseteq \beta'$. If $\text{proj}(\beta') \supset \alpha$, then $\beta' \supset \alpha$. The collection of input q -sequences that include α is defined as $\rho(\alpha) = \{\Psi' \in \mathcal{D}' \mid \Psi' \supset \alpha\}$. Consequently, the support of α is computed as $\text{supp}(\alpha) = |\rho(\alpha)|$.

Example 1. For $\alpha = \langle b, e \rangle$ and $\Psi_2 = \text{proj}(\Psi'_2) = \langle b, ce, a, ad \rangle$, since $\Psi_2 \supset \alpha$, so $\Psi'_2 \supset \alpha$. It follows that $\Psi'_1 \supset \alpha$ and $\Psi'_3 \supset \alpha$, thus $\rho(\alpha) = \{\Psi'_1, \Psi'_2, \Psi'_3\} = \mathcal{D}$. Therefore, the support of α is $\text{supp}(\alpha) = 3$. Moreover, α appears three times in Ψ'_3 , since all three subsequences $\langle (b, 8), (e, 12) \rangle \supseteq \alpha$, $\langle (b, 8), (e, 20) \rangle \supseteq \alpha$ and $\langle (b, 7), (e, 20) \rangle \supseteq \alpha$. The utility values of these three occurrences are 20, 28 and 27, respectively.

For any $\Psi' \in \rho(\alpha)$, let $\mathcal{O}(\alpha, \Psi') \stackrel{\text{def}}{=} \{\alpha' \mid \alpha' \sqsubseteq \Psi' \wedge \text{proj}(\alpha') = \alpha\}$ denote as the collection of occurrences α' of α in Ψ' . Since it can include multiple occurrences, various approaches can be used to determine the utility value of α in the input sequence Ψ' . For instance, for $\Psi'_3 \in \rho(\alpha)$, then $\mathcal{O}(\alpha, \Psi'_3) = \{\langle (b, 8), (e, 12) \rangle, \langle (b, 8), (e, 20) \rangle, \langle (b, 7), (e, 20) \rangle\}$. Thus, the utility of α within Ψ' is determined as either the *minimum* or *maximum utility* among all occurrences of α in Ψ' . In this work, we adopt the maximum form defined as below.

Definition 4 (Maximum utility function). The maximum utility value of sequence α in Ψ'_i [5] is calculated as $u_{\max}(\alpha, \Psi'_i) \stackrel{\text{def}}{=} \max\{u(\alpha') \mid \alpha' \in \mathcal{O}(\alpha, \Psi'_i)\}$, and its maximum utility in the dataset $\mathcal{D}'$ is $u_{\max}(\alpha) \stackrel{\text{def}}{=} \sum_{\Psi'_i \in \rho(\alpha)} u_{\max}(\alpha, \Psi'_i)$.

Example 2. Consider $\alpha = \langle b, e \rangle$ as in Example 1, we have $\mathcal{O}(\alpha, \Psi'_3) = \{\langle (b, 8), (e, 12) \rangle, \langle (b, 8), (e, 20) \rangle, \langle (b, 7), (e, 20) \rangle\}$. Thus, $u_{\max}(\alpha, \Psi'_3) = 8 + 20 = 28$.

For brevity, from now on, whenever the utility of a sequence is mentioned, it means that the maximum utility of that sequence is used.

Definition 5 (Average-utility occupancy of sequence).

For any sequence α , we denote $u(\alpha, \Psi') =$

$\max(u(\alpha', \Psi') | \alpha' \in \mathcal{O}(\alpha, \Psi'))$ and $au(\alpha, \Psi') = \frac{u(\alpha, \Psi')}{|\alpha|}$ as its utility and AU in Ψ' , respectively. The AU occupancy of α in Ψ' is determined as $auos(\alpha, \Psi') \stackrel{\text{def}}{=} \frac{au(\alpha, \Psi')}{SU(\Psi')}$, where $SU(\Psi')$ is the total utility of all item utilities in Ψ' . Then, the average utility occupancy of α in $\mathcal{D}'$ is defined as $auos(\alpha) = \frac{1}{\text{supp}(\alpha)} \sum_{\Psi' \in \rho(\alpha)} auos(\alpha, \Psi')$.

Example 3. Consider sequence $\alpha = \langle b, e \rangle$, the utility and AU of α in Ψ'_3 are respectively $u(\alpha, \Psi'_3) = \max(u(\alpha', \Psi'_3) | \alpha' \in \mathcal{O}(\alpha, \Psi'_3)) = \max(20; 28; 27) = 28$ and $au(\alpha, \Psi'_3) = \frac{u(\alpha, \Psi'_3)}{|\alpha|} = \frac{28}{2} = 14$. Since the input q-sequence utility of Ψ'_3 is $SU(\Psi'_3) = 289$, the utility occupancy of α in Ψ'_3 is $uos(\alpha, \Psi'_3) = \frac{u(\alpha, \Psi'_3)}{SU(\Psi'_3)} = \frac{28}{289} = 0.0969$. Thus, the average utility occupancy of α (with $\text{length}(\alpha) = 2$) in Ψ'_3 , $auos(\alpha, \Psi'_3) = \frac{0.0969}{2} = 0.0485 = (\frac{au(\alpha, \Psi'_3)}{SU(\Psi'_3)}) = \frac{14}{289} = 0.0485$. Similarly, by considering all input q-sequence containing α , $\rho(\alpha) = \{\Psi'_1, \Psi'_2, \Psi'_3\}$, the average utility occupancy of α in $\mathcal{D}'$ is determined as follows: $auos(\alpha) = \frac{(\frac{25}{130} + \frac{18}{111}) + 0.0485}{3} = \frac{(\frac{0.1923 + 0.1621}{2}) + 0.0485}{3} = 0.0752$.

Problem Definition. Given two user-specified thresholds, a minimum support $ms > 0$ and a minimum AU occupancy $mo \in (0, 1]$, a sequence α is regarded as frequent if $\text{supp}(\alpha) \geq ms$. If $auos(\alpha) \geq mo$, then α is termed a high AU occupancy sequence (HAUOS); otherwise, it is called a low AU occupancy sequence (LAUOS). A sequence α that satisfies both these conditions, i.e., $\text{supp}(\alpha) \geq ms$ and $auos(\alpha) \geq mo$, is defined as a frequent high average utility occupancy sequence (FHAUOS). The FHAUOS mining problem (FHAUOSM) aims to exploit the complete collection of such sequences, formally defined as:

$$\text{FHAUOS} \stackrel{\text{def}}{=} \{\alpha \mid auos(\alpha) \geq mo \text{ and } \text{supp}(\alpha) \geq ms\}.$$

A key challenge in addressing the FHAUOSM problem lies in the exponential growth of the search space, particularly due to the involvement of the average utility occupancy function. To address this challenge, it is essential to exploit anti-monotonic property either of $auos$ or of its suitable upper bounds to early discard unpromising candidates. Formally, a measure f defined over sequences satisfies the anti-monotonic ($\mathcal{AM}$) property if it holds that $f(\beta) \leq f(\alpha)$, for any α and its super-sequence β . This property is particularly important in pattern mining, as it allows the elimination of entire branches of the search space without loss of completeness.

Unfortunately, the function $auos$ does not satisfy this property. For example, consider $\alpha = \langle b, e \rangle$ and its super-sequence $\beta = \langle b, e, a \rangle$. We observe that $auos(\alpha) = 0.07524 < auos(\beta) = 0.1741$, thus violating the anti-monotonicity condition.

To tackle this challenge, the subsequent sections introduce

newly formulated upper bounds (UBs) for the $auos$ function that exhibit anti-monotone-like property, enabling an efficient reduction of the search space during the mining process.

III. MINING FHAUOSS

A. Novel upper bounds and pruning strategies

Definition 6 (Weak upper bound and upper bound). For any functions f , ub and wub , then ub is referred to as an upper bound (UB) of f if the condition $f(\alpha) \leq ub(\alpha)$ holds for any sequence α . wub is referred to as a weak UB (WUB) of f , if the condition $f(\beta) \leq wub(\alpha)$ holds for any proper FE β of α , $\beta = \alpha \diamond \delta \supset \alpha$ with $\delta \neq \langle \rangle$.

Let $\mathcal{AM}(ub)$ denote as any UB ub that satisfies the $\mathcal{AM}$ property.

To effectively narrow the search space, we begin by proposing a new UB on $auos$ and then develop a matching pruning strategy.

For any q -sequence $\Psi' \in \rho(\alpha)$, after sorting the utility values of items in Ψ' in descending order, the new series $U(\Psi') \stackrel{\text{def}}{=} \{y_1, \dots, y_N\}$ is obtained with N being the total count of all items in Ψ' . Let define $\tau(\Psi', k)$, the AU of the top- k highest values in $U(\Psi')$, $\tau(\Psi', k) \stackrel{\text{def}}{=} \frac{1}{k} \sum_{1 \leq i \leq k} y_i$, for $k = 1, 2, \dots, N$.

Definition 7 (UB ubaos on auos). For any frequent sequence α and an input q -sequence $\Psi' \in \rho(\alpha)$,

- The AU occupancy UB of α in Ψ' is determined as $ubaos(\alpha, \Psi') \stackrel{\text{def}}{=} \frac{\tau(\Psi', l)}{SU(\Psi')} = \frac{\frac{1}{l} \sum_{1 \leq i \leq l} y_i}{SU(\Psi')}$, where $l = \text{length}(\alpha)$ and $SU(\Psi')$ is the sum of all item utilities in Ψ' , i.e. $SU(\Psi') = \sum_{1 \leq i \leq N} y_i$.
- The AU occupancy UB of α in $\mathcal{D}$ is defined as $ubaos(\alpha) \stackrel{\text{def}}{=} \frac{1}{ms} \sum_{1 \leq i \leq ms} v_i$, where $V = \{v_1, v_2, \dots, v_n\}$ that is formed by sorting the series $\{ubaos(\alpha, \Psi'), \Psi' \in \rho(\alpha)\}$ in descending order, and $n = |\rho(\alpha)|$.

Example 4. Consider $\alpha = \langle b, e \rangle$ and $\rho(\alpha) = \{\Psi'_1, \Psi'_2, \Psi'_3\}$. For Ψ'_1 , after sorting item utilities of Ψ'_1 in descending order, the new series $U(\Psi'_1) = \{35; 30; 18; 15; 11; 8; 7; 6\}$ is obtained and $(\tau(\Psi'_1, k), k = 1, 2, \dots, 8) = (35; 32.5; 27.67; 24.5; 21.8; 19.5; 17.71; 16.25)$ since $\tau(\Psi'_1, 1) = 35$, $\tau(\Psi'_1, 2) = (35 + 30)/2 = 32.5$, $\tau(\Psi'_1, 3) = (35 + 30 + 18)/3 = 27.67$, and so on. Therefore, $ubaos(\alpha, \Psi'_1) = \frac{\tau(\Psi'_1, 2)}{SU(\Psi'_1)} = \frac{32.5}{130} = 0.25$. Similarly, we have $ubaos(\alpha, \Psi'_2) = 0.27$ and $ubaos(\alpha, \Psi'_3) = 0.24$. For $ms = 2$ and $mo = 0.2$, after sorting the series $\{0.25, 0.27, 0.24\}$ in descending order to create the new series $V = \{v_1, v_2, v_3\} = \{0.27, 0.25, 0.24\}$, then $ubaos(\alpha) \stackrel{\text{def}}{=} \frac{1}{ms} \sum_{1 \leq i \leq ms} v_i = \frac{1}{2} (0.27 + 0.25) = 0.26 \geq auos(\alpha) = 0.075$.

The $ubaos$ is a UB on $auos$ and has the $\mathcal{AM}$ property as presented in Proposition 1. Its proof is not provided in the paper because of the space constraints.

Proposition 1 (The $\mathcal{AM}$ property of $ubauos$).

- a) $ubauos$ is a UB on $auos$, i.e. $auos(\alpha) \leq ubauos(\alpha), \forall \alpha$.
- b) The $\mathcal{AM}$ property holds for $ubauos$, i.e. $ubauos(\beta) \leq ubauos(\alpha), \forall \beta \supseteq \alpha$.

Example 5. For $\alpha = \langle b, c \rangle$ and its super-sequence $\beta = \langle b, c, f \rangle$, then $\rho(\beta) = \{\Psi'_1\}$, $\rho(\beta) \subseteq \rho(\alpha) = \{\Psi'_1, \Psi'_2, \Psi'_3\}$, $auos(\alpha) = 0.08$, $ubauos(\alpha) = 0.26$ and $ubauos(\beta) = 0.213$. Hence, $auos(\alpha) < ubauos(\beta) < ubauos(\alpha)$.

Let $WPC(\alpha) \stackrel{\text{def}}{=} (ubauos(\alpha) < mo \text{ or } supp(\alpha) < ms)$ be the width pruning condition. For brevity, we denote $branch(\alpha)$ (or $pbranch(\alpha)$) as the set including extensions of α , comprising (or excluding) α itself.

Theorem 1 (Pruning strategies based on ms and $ubauos$).

- a) (Pruning strategy 1 – PS1) For any sequence α , if $WPC(\alpha)$ holds, then $auos(\beta) < mo$ or $supp(\beta) < ms$ for super-sequences β of α . Then, the entire branch $branch(\alpha)$ can be pruned early.
- b) (Pruning strategy 2 – PS2) For any item y , if $WPC(y)$ holds, y may be discarded from the QSDB $\mathcal{D}'$. Then, the $ubauos$ values of all sequences in $\mathcal{D}'$ are often reduced.

The proof of this theorem is not provided in the paper due to the limitation of the space.

Pruning Strategy 2 reduces the database at the start, while Pruning Strategy 1 prunes branches $branch(\alpha)$ during mining. Since $ubauos$ is usually very large, branch pruning rarely occurs; instead, it is mainly applied in Strategy 2 to remove items for reducing the dataset. To enable more effective branch pruning, a tighter upper bound is required, as introduced below.

The collection of all input q -sequences that contain at least one proper FE of α is denoted as $PE(\alpha)$. Let $\mathcal{O}^*(\alpha, \Psi') \stackrel{\text{def}}{=} \{\alpha' \in \mathcal{O}(\alpha, \Psi') \mid rem(\alpha', \Psi') \neq \langle \rangle\}$.

Definition 8 (The largest average utility of proper forward extensions). For any $\alpha, \Psi' \in PE(\alpha)$ and $\alpha' \in \mathcal{O}^*(\alpha, \Psi')$, the largest average utility of proper FEs of α' in Ψ' is determined as follows:

$$lau(\alpha', \Psi') = \begin{cases} f(r), & \text{if } mru(\alpha', \Psi') \geq au(\alpha', \Psi') \\ f(1), & \text{if } mru(\alpha', \Psi') < au(\alpha', \Psi') \end{cases}$$

where $mru(\alpha', \Psi') = \max\{u(y, \Psi') \mid \text{item } y \in rem(\alpha', \Psi')\}$, $r = |rem(\alpha', \Psi')|$, $l = length(\alpha)$ and $f(k) \stackrel{\text{def}}{=} \frac{l \cdot au(\alpha', \Psi') + k \cdot mru(\alpha', \Psi')}{l + k}$ is an intermediate value between $au(\alpha', \Psi')$ and $mru(\alpha', \Psi')$, $\forall k = 1 \dots r$.

Then, a WUB on $auos$ can be designed using the definition of lau .

Definition 9 (Novel WUB on $auos$). For any sequence α with $l = length(\alpha)$ and $\Psi' \in PE(\alpha)$,

- a) The WUB of average utility occupancy of α in Ψ' for $auos$ is defined as

$$wubauos(\alpha, \Psi') \stackrel{\text{def}}{=} \frac{1}{SU(\Psi')} \max\{lau(\alpha', \Psi') \mid \alpha' \in \mathcal{O}^*(\alpha, \Psi')\}.$$

- b) After sorting the series $\{wubauos(\alpha, \Psi') \mid \Psi' \in PE(\alpha)\}$ in descending order to create the new series $T = \{t_1, t_2, \dots, t_n\}$,

where $n = |PE(\alpha)|$, the WUB $wubauos$ for $auos$ on the entire dataset is defined as $wubauos(\alpha) \stackrel{\text{def}}{=} \frac{1}{ms} \sum_{1 \leq i \leq ms} t_i$.

Example 6. For $ms = 2$, $\alpha = \langle b, c \rangle$, and $PE(\alpha) = \{\Psi'_1, \Psi'_2, \Psi'_3\}$, then α has two endings of 2 and 5 in Ψ'_1 , and $\mathcal{O}^*(\alpha, \Psi'_1) = \{\alpha'_1 = \langle (b, 7), (c, 15) \rangle, \alpha'_2 = \langle (b, 7), (c, 6) \rangle\}$. Then, $rem(\alpha'_1, \Psi'_1) = \langle (f, 35), (g, 30), [(b, 8)(c, 6), (e, 18)] \rangle$ and $rem(\alpha'_2, \Psi'_1) = \langle (e, 18) \rangle$. Therefore, for q -sequence α'_1 , we have $mru(\alpha'_1, \Psi'_1) = 35$, $r = 5$ and $au(\alpha'_1, \Psi'_1) = 11$. Since $mru(\alpha'_1, \Psi'_1) > au(\alpha'_1, \Psi'_1)$, so $lau(\alpha'_1, \Psi'_1) = f(r) = \frac{2 \cdot 11 + 5 \cdot 35}{2 + 5} = 28.14$. For q -sequence α'_2 , then $mru(\alpha'_2, \Psi'_1) = 18$, $r = 1$ and $au(\alpha'_2, \Psi'_1) = 6.5$. We have $lau(\alpha'_2, \Psi'_1) = f(1) = \frac{2 \cdot 6.5 + 18}{2 + 1} = 10.33$, since $mru(\alpha'_1, \Psi'_1) > au(\alpha'_1, \Psi'_1)$. Thus, $wubauos(\alpha, \Psi'_1) = \frac{1}{130} \max(28.14, 10.33) = 0.216$. Similarly, $wubauos(\alpha, \Psi'_2) = 0.28$ and $wubauos(\alpha, \Psi'_3) = 0.057$. Hence, $T = \{t_1, t_2, t_3\} = \{0.28, 0.216, 0.057\}$, and $wubauos(\alpha) = \frac{1}{2} (0.28 + 0.216) = 0.248$.

The next proposition establishes that $wubauos$ serves as the WUB for $auos$. Its proof is omitted here due to space constraints.

Proposition 2 (Property of the WUB $wubauos$). $wubauos$ is a WUB on $auos$, i.e. $auos(\beta) \leq wubauos(\alpha)$ with any proper forward extension β of α .

From the characteristics of $wubauos$, the following theorem can be immediately derived. The detailed proof is omitted here because of space constraints.

Theorem 2 (Pruning strategy 3 – PS3). For any sequence α , if $wubauos(\alpha) < mo$, we can prune the proper branch $pbranch(\alpha)$. Additionally, if $auos(\alpha) < mo$, the entire branch $branch(\alpha)$ is safely eliminated.

This theorem enables the early elimination of unpromising branches from the prefix search tree during the exploration process.

B. Proposed FHAUOS-Miner Algorithm

Building upon the theoretical foundations presented above, we design a novel algorithm named FHAUOS-Miner to identify all FHAUOSs. To enhance mining efficiency, the algorithm incorporates pruning techniques derived from Theorems 1 and 2, effectively narrowing the search space.

Fig. 1 shows the pseudo-code of FHAUOS-Miner, which accepts as input a dataset $\mathcal{D}'$ together with the minimum threshold values mo and ms . Initially, FHAUOS-Miner scans $\mathcal{D}'$ to identify the set $S := \{x \in \mathcal{A} \mid \text{not}(WPC(x))\}$, which includes only those items that qualify as FHAUO candidates based on their $ubauos$ and $supp$ values (line 1). In lines 2–5, it discards all unpromising items in $\mathcal{A} \setminus S$ from $\mathcal{D}'$, and then updates (and potentially reduces) the $ubauos$ of the remaining relevant items in S . Finally, for each item $x \in S$, the recursive procedure Search-FHAUOS is invoked (line 7).

The Search-FHAUOS procedure, illustrated in Fig. 2, operates on a given sequence α , and two sets I and S for candidate items, along with the thresholds mo and ms . Initially, if the conditions specified at line 1 are met, the sequence α is added to the result set. Following that, a depth pruning strategy is applied at line 2, leveraging the $wubauos$

WUB described in Theorem 2. The algorithm then proceeds with a tightening strategy (lines 4–9), which uses the *ubauos* UB to guide the extension of α efficiently. Finally, Search-FHAUOS is recursively executed for each item in the *IE* and *SE* extension sets of α (lines 10–13).

FHAUOS-Miner ($\mathcal{D}', mo, ms$)
Input: A dataset $\mathcal{D}'$ and values of mo and ms .
Output: the *FHAUOS* set of all FHAUOSs.
1. Calculate $S := \{x \in \mathcal{A} \mid \text{not}(WPC(x))\}$; //Theorem 1
2. **for each** item $y \in \mathcal{A}$ **do**
3. **if** $WPC(y)$ **then** // Theorem 1
4. Eliminate irrelevant item y from $\mathcal{D}'$;
5. Update the WUBs for items in S .
6. **for each** item $x \in S$ **do**
7. **Search-FHAUOS** (x, S, S, mo, ms);

Fig. 1. The FHAUOS-Miner Algorithm.

Search-FHAUOS (α, I, S, mo, ms)
Input: α , two candidate sets I and S , and values of mo and ms .
Output: sequence α if it is an FHAUOS.
1. **if** ($auo(\alpha) \geq mo$ and $supp(\alpha) \geq ms$) **then** output FHAUOS α ;
2. **if** ($wubauos(\alpha) < mo$ or $supp(\alpha) < ms$) **then return**;
3. $IE := SE := \emptyset$; $x := \text{lastItem}(\alpha)$;
4. **for each** item $y \in I$ such that $y > x$ **do** //Theorems 1 and 2
5. **if** ($ubauos(\alpha \circ_i y) \geq mo$ and $supp(\alpha \circ_i y) \geq ms$) **then**
6. $IE := IE \cup \{y\}$;
7. **for each** item $y \in S$ **do** //Theorems 1 and 2
8. **if** ($ubauos(\alpha \circ_s y) \geq mo$ and $supp(\alpha \circ_s y) \geq ms$) **then**
9. $SE := SE \cup \{y\}$;
10. **for each** item $y \in IE$ **do**
11. **Search-FHAUOS** ($\alpha \circ_i y, IE, SE, mo, ms$);
12. **for each** item $y \in SE$ **do**
13. **Search-FHAUOS** ($\alpha \circ_s y, SE, SE, mo, ms$);

Fig. 2. The Search-FHAUOS Procedure.

IV. EXPERIMENTAL EVALUATION

All experiments were conducted on a computing system equipped with an Intel(R) Core(TM) i5-2400U processor running at 3.1 GHz and 16 GB of RAM. The algorithms used for comparison were all implemented in Java version 1.8. Owing to limited space, the performance assessment focused on two datasets, one real-life dataset (Gazelle [15]) and one synthetic dataset (D4C7T5N5S6I4). Their main characteristics are outlined in Table 2. Utility values were randomly created utilizing the IBM Quest data generator [15]. As far as our review indicates, no prior work has addressed the mining of FHAUOSs. Consequently, the proposed method is compared with a baseline version that omits the integration of the proposed UBs and WUBs.

A. Impact of the minimum support parameter

The first experiment evaluates the impact of ms on runtime, memory consumption, and candidate count in the QSDBs shown in Table 2. The results are shown in Fig. 3. It is observed that as ms decreases, runtime, memory consumption, and the candidate count (#Candidates) increase significantly. This is because a lower ms value leads to a substantial growth in #Candidates. From the figure, it can be seen that FHAUOS-Miner consistently outperforms the Baseline algorithm in terms of both runtime and memory consumption. For example, on the QSDB Gazelle, the average runtime of FHAUOS-Miner is four times smaller than that of the Baseline algorithm. This improvement is due to

FHAUOS-Miner employing effective strategies based on the proposed bounds to early eliminate unpromising candidates, resulting in a much smaller number of candidates being evaluated compared to the Baseline algorithm, as shown in Fig. 3.

TABLE 2. CHARACTERISTICS OF DATABASES

Database	#seq	#item	Avg. length	Data type
Gazelle	59,601	497	2.51	Web click stream
D4C7T5N5S6I4	4,000	5,000	28.7	Synthetic

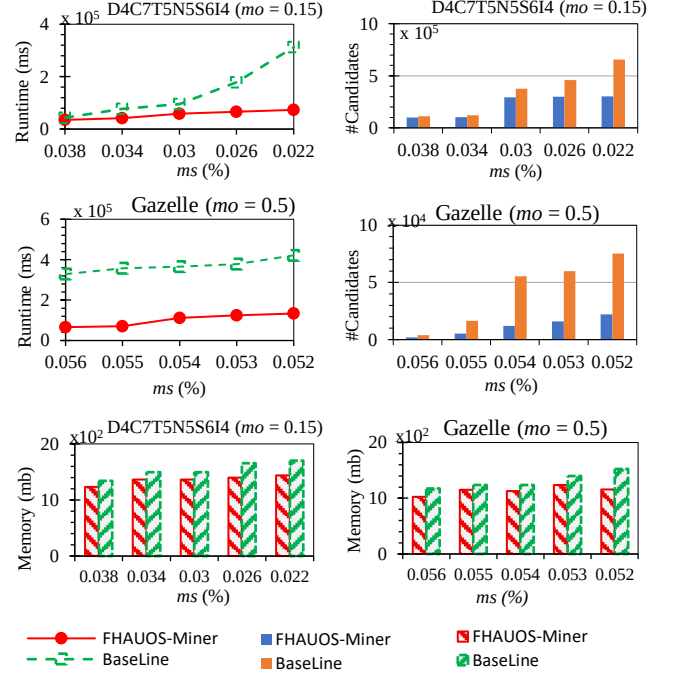


Fig. 3. Comparing two algorithms for fixed mo .

B. Impact of the mo parameter

In the second experiment, we examine the effect of varying the mo value while keeping ms fixed, using the QSDBs in Table 2. Similarly, as mo decreases, runtime, memory consumption, and the candidate count increase, since lower mo values allow more patterns to meet the threshold. Across all datasets, the proposed FHAUOS-Miner consistently outperforms the Baseline algorithm, achieving shorter runtimes and lower memory consumption. For instance, on the Gazelle QSDB, FHAUOS-Miner runs about nine times faster than the Baseline by leveraging proposed pruning strategies to discard unpromising candidates early, thereby greatly reducing the number of candidates considered.

Overall, the experimental results demonstrate that both ms and mo significantly affect execution time, memory consumption, and the candidate count. In all cases, FHAUOS-Miner consistently outperforms the Baseline algorithm by maintaining lower computational costs. This advantage is mainly attributed to its effective upper-bound-based pruning strategies, which drastically reduce the search space.

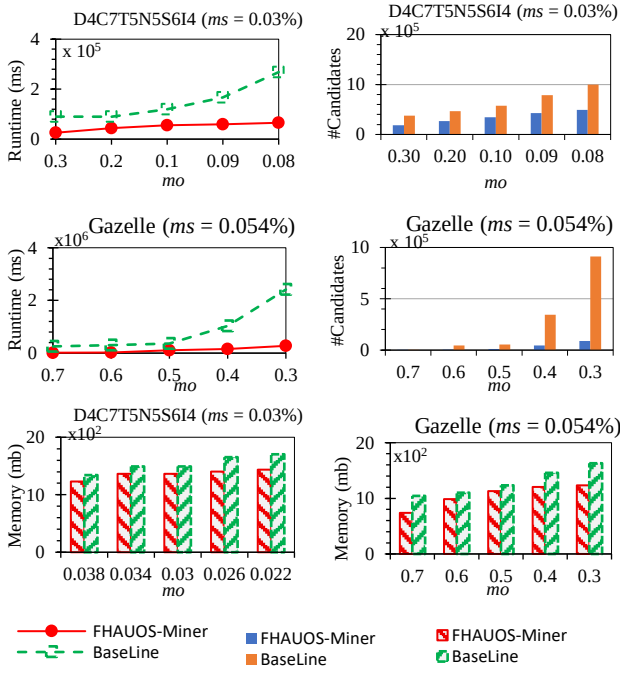


Fig. 4. Comparing two algorithms for fixed ms .

V. CONCLUSION

This paper presented the novel problem of mining FHAUOSs, which is highly relevant to numerous real-world applications. To address this problem efficiently, we proposed two new bounds - $ubauos$ (an upper bound) and $wubauos$ (a weak upper bound) - along with corresponding pruning techniques to early discard unpromising candidates during the mining process. Building on these theoretical foundations, we developed a new algorithm, FHAUOS-Miner, designed to effectively discover FHAUOSs. Experimental results demonstrate that our algorithm is efficient in terms of execution time, memory consumption, and candidate count.

For future work, we aim to develop tighter UBs and apply parallel processing to improve FHAUOS mining on large datasets.

ACKNOWLEDGMENT. This research is funded by Dalat University, Vietnam under decision number 198/QĐ-ĐHDL, 2025.

REFERENCES

- [1] C. F. Ahmed, S. K. Tanbeer, and B. S. Jeong, "Mining high utility web access sequences in dynamic web log data," in *Proceedings of 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD2010*, 2016, pp. 76–81.
- [2] B. E. Shie, J. H. Cheng, K. T. Chuang, and V. S. Tseng, "A one-phase method for mining high utility mobile sequential patterns in mobile commerce environments," In: Jiang H., Ding W., Ali M., Wu X. (eds) *Advanced Research in Applied Artificial Intelligence. IEA/AIE 2012. Lecture Notes in Computer Science*, 2012, pp. 616–626.

- [3] B. Dalmás, P. Fournier-Viger, and S. Norre, "TWINCLE: A Constrained Sequential Rule Mining Algorithm for Event Logs," in *Proceedings of 9th International KES Conference (IDT-KES 2017)*, 2017, pp. 205–214.
- [4] C. F. Ahmed, S. K. Tanbeer, and B. S. Jeong, "A novel approach for mining high-utility sequential patterns in sequence databases," *ETRI Journal*, vol. 32, no. 5, pp. 676–686, 2016.
- [5] J. Yin, Z. Zheng, and L. Cao, "USpan: An efficient algorithm for mining high utility sequential patterns," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 660–668.
- [6] H. Duong, T. Truong, B. Le, and P. Fournier-Viger, "Mining Interesting Sequential Patterns using a Novel Balanced Utility Measure," *Knowl Based Syst*, vol. 294, p. 111796, 2024.
- [7] W. Gan, J. C.-W. Lin, J. Zhang, P. Fournier-Viger, H. Chao, and P. S. Yu, "Fast Utility Mining on Complex Sequences," *CoRR*, vol. 1904, no. 2, pp. 1–15, 2019.
- [8] T. Truong, H. Duong, B. Le, and P. Fournier-Viger, "FMaxCloHUSM: An efficient algorithm for mining frequent closed and maximal high utility sequences," *Eng Appl Artif Intell*, vol. 85, no. 1, pp. 1–20, 2019.
- [9] M. Thilagu and R. Nadarajan, "Efficiently Mining of Effective Web Traversal Patterns with Average Utility," *Proc. Technol.*, vol. 6, no. 1, pp. 444–451, 2012.
- [10] W. Gan, J. C. W. Lin, P. Fournier-Viger, H. C. Chao, and P. S. Yu, "HUOPM: High-Utility Occupancy Pattern Mining," *IEEE Trans Cybern*, vol. 50, no. 3, pp. 1195–1208, Mar. 2020.
- [11] G. Huang, W. Gan, and P. S. Yu, "Towards Sequence Utility Maximization under Utility Occupancy Measure," *Journal of the ACM*, vol. 1, no. 1, pp. 1–24, 2022.
- [12] M. J. K. Kumar and D. Rana, "HAUOPM: High Average Utility Occupancy Pattern Mining," *Arab J Sci Eng*, vol. 49, pp. 3397–3416, 2024.
- [13] T. Truong, H. Duong, B. Le, and P. Fournier-Viger, "EHAUSM: An efficient algorithm for high average utility sequence mining," *Inf Sci (N Y)*, vol. 515, no. 1, pp. 302–323, 2020.
- [14] H. Duong, T. Truong, T. Hoang, and B. Le, "U-HPAUSM: Mining high probability average utility sequences in uncertain quantitative sequential databases," *Eng Appl Artif Intell*, vol. 141, Feb. 2025.
- [15] P. Fournier-Viger *et al.*, "SPMF: a Java Open-Source Pattern Mining Library Version 2," *Machine Learning Research*, vol. 15, no. 1, pp. 3389–3393, 2014.

EvoPort: An Evolutionary Framework for Portfolio Optimization via Randomized Alpha Discovery and Ensemble-Based Allocation

Nguyen Van Thanh

*Faculty of Information Technology
University of Engineering and Technology
Vietnam National University
Hanoi, Vietnam
23025090@vnu.edu.vn*

Nguyen Thi Hau*

*Faculty of Information Technology
University of Engineering and Technology
Vietnam National University
Hanoi, Vietnam
nguyenhau@vnu.edu.vn*

Abstract—Portfolio optimization is a core problem in quantitative finance, aiming to balance return, risk, and stability across market regimes. Traditional methods often depend on fixed models or limited signals, which fail to adapt to changing environments. We introduce EvoPort, an evolutionary framework that combines randomized alpha discovery, ensemble-based model selection, and stochastic weight allocation. Inspired by evolutionary search, EvoPort iteratively refines candidate signals through mutation and selection, ensuring adaptability to shifting markets. From raw equity data, EvoPort generates diverse trading signals using symbolic, time-series, and logical operators, optimized via randomized hill climbing guided by the Sharpe ratio. Predictive models are selected through a stochastic ensemble of machine learning and reinforcement learning techniques. Portfolio weights are dynamically assigned based on risk-adjusted return, turnover, and alpha diversity. Our best configuration achieves an out-of-sample Sharpe ratio of 2.70 and return of 17.5%, outperforming benchmarks such as SPY of Yahoo Finance (1.20, 17%), KMRF of Piotr Pomorski and Denise Gorse, 2023 (1.44, 14.4%), and DRL-PPO of Srijan Sood et al., 2023 (1.16, 12.1%). EvoPort also surpasses traditional weighting rules like Equal Weight and Low Correlation in robustness and stability.

Index Terms—optimize portfolio, random hill climbing, ensemble model, weight allocation

I. INTRODUCTION

Portfolio optimization is a core challenge in quantitative finance, aiming to allocate capital across assets to maximize return for a given level of risk [15]. Its importance spans asset management, hedge funds, pension systems, and robo-advisory platforms. Markowitz’s Modern Portfolio Theory (MPT) [15] pioneered the return–risk framework via variance-based modeling and the efficient frontier. However, its assumptions—normal return distributions, stable covariance, and quadratic utility—are often violated in real markets, which exhibit nonstationarity, heavy tails, and abrupt regime shifts. These flaws make traditional portfolios vulnerable to estimation error, especially in high-dimensional settings.

To address this, three major research directions have emerged. First, Bayesian and robust optimization methods

enhance stability under model uncertainty [3, 16, 2]. Second, machine learning (ML) and deep learning (DL) approaches uncover complex patterns in financial data without strong assumptions [8, 6]. Third, reinforcement learning (RL) treats portfolio construction as a sequential decision problem, allowing dynamic adaptation through policies learned from market data [19, 20]. While powerful, these methods often decouple the investment pipeline—splitting alpha generation, model learning, and allocation into isolated steps—limiting their adaptability and overall performance [1]. Many also rely on narrow, hand-crafted features [10], and even advanced RL models can suffer from overfitting and large drawdowns in volatile markets [9].

To overcome these challenges, we propose **EvoPort**, an evolutionary framework that unifies alpha discovery, model selection, and portfolio allocation into a single adaptive process. Inspired by evolutionary search, EvoPort iteratively refines candidate signals through variation and selection, enabling continuous adaptation to shifting market regimes. EvoPort builds diverse trading signals through symbolic and time-series transformations, refined via randomized hill climbing based on Sharpe ratio, return, and drawdown. These signals feed a stochastic ensemble of ML and RL models for prediction. Final portfolio weights are optimized using stochastic sampling strategies that consider alpha quality, turnover, and market risk.

By integrating the full investment stack, EvoPort produces stable, adaptive portfolios that generalize across regimes. Its modular architecture allows for flexible extension to multi-asset, multi-frequency, or constraint-aware scenarios, offering a practical and scalable solution for modern portfolio optimization.

II. RELATED WORK

A. Feature Engineering and Alpha Search

Early studies on quantitative investing devoted most of their attention to feature engineering, by transforming raw data into predictive factors. Common transformations include rolling means and standard deviations, z-scores and a wide range

*Corresponding author: nguyenhau@vnu.edu.vn

of technical indicators such as RSI, MACD, and Bollinger Bands [22]. Although these hand-crafted features are easy to interpret, they are typically selected through heuristic trial-and-error, which makes them vulnerable to data-snooping bias and regime shifts. More recent work has explored automated construction of features with principal-component analysis, auto-encoders, and other representation-learning tools, yet these methods often sacrifice transparency and still require a separate stage for alpha discovery.

Alpha search aims to combine those primitives into executable trading rules whose performance is verified by back-testing. Genetic Programming remains the most visible approach, evolving tree-structured expressions to maximise out-of-sample Sharpe ratios but rarely addressing turnover or drawdown, and its best reported Sharpe of 1.1 [7] leaves room for improvement. Ant Colony Optimisation [13] and Tabu Search augmented with Simulated Annealing [14] broaden the exploration of the search space but incur high computational cost. Randomised Hill Climbing [4], by contrast, offers a lightweight mechanism that quickly exploits local neighbourhood structure, making it attractive when frequent re-evaluation of candidates is expensive; its main drawback is the tendency to settle in local optima unless diversification mechanisms are added. EvoPort retains this efficiency while embedding guided mutation and adaptive stopping to mitigate premature convergence.

B. Portfolio Optimization and Model Training

Research on portfolio construction has likewise advanced from classical mean-variance theory to frameworks that learn to adapt to market regimes and nonlinear dynamics. Pomorski and Gorse (2023) propose a regime-prediction overlay that shifts allocations ahead of structural breaks, demonstrating materially higher risk-adjusted returns than static optimisation [19]. In parallel, Sood *et al.* (2023) employ a PPO agent to learn trading actions directly from price trajectories, outperforming traditional baselines in turbulent markets [20]. Weighting schemes such as Equal Weighting [18], Low Correlation Weighting [17], and explicit Low Turnover portfolios [21] address specific practical concerns but treat feature generation, signal learning, and allocation as isolated steps, often leading to sub-optimal interplay between alpha quality and execution cost.

By coupling symbolic alpha discovery with ensemble model learning and by embedding turnover and drawdown directly in its objective, EvoPort unifies these stages into a single optimisation loop. As the back-tests on Yahoo Finance data [23] and the head-to-head comparison against the regime model of Pomorski-Gorse and the DRL-PPO agent of Sood *et al.* (reported in Table III) show, such integration yields portfolios that remain resilient across shifting regimes while preserving implementability through controlled trading intensity.

III. METHODOLOGY

The EvoPort architecture is designed as a four-stage pipeline integrating evaluate data, feature generation, model training,

and portfolio optimization, as illustrated in Fig 1. The four stages are linked by a shared randomized up-hill climbing mechanism such that exploration of the solution space via systematic searches can be employed to construct varied and robust investment portfolios.

A. Evaluate Data and Operator Declaration

We conduct an extensive data evaluation step to characterize the statistical nature of every feature and inform downstream operator and lookback parameter selection. Table I provides the metrics considered in the evaluation:

TABLE I
DATA EVALUATION METRICS

Method	Formula
Coverage	$coverage\_ratio = \frac{count\_non\_null}{total\_expected}$
Frequency	$frequency\_ratio = \frac{count\_changes}{N-1}$
Outlier Detection	$outlier\_ratio = \frac{count(data > X)}{total}$
Median Deviation	$deviation\_from\_expected = \frac{ median - expected }{expected}$
Distribution Shape	skewness, kurtosis
Missingness	missing_ratio
Duplication	duplicate_ratio
Continuity	max_gap
Volatility	volatility_ratio

To comprehensively decide the quality of the dataset, we employ a variety of measurement metrics that capture the different aspects of the data. We use Coverage to ascertain the percentage of missing values by calculating the fraction of values present to values that ought to be present. Frequency is used to measure the rate of updates by calculating the fraction of change within a particular window. Anomaly detection identifies outliers as a function of the ratio of values exceeding a threshold. Distribution shape is analyzed through skewness and kurtosis to detect asymmetry and unusual tails in the data. We adopt these metrics following the data quality framework outlined in Ehrlinger *et al.* (2019) [5], which provides a comprehensive review of data quality assessment methods. We categorize a varied assortment of operators into 10 operational groups. In addition to standard mathematical, logical, cross-sectional, and time-series transforms, we consider advanced modules such as machine learning models, risk control operators, deep learning conversions, and ensemble statistics. Examples of operators appear in Table II. This taxonomy is inspired by the categorization frameworks proposed in He *et al.* (2019) [11], which highlight the role of diverse operator types in automated machine learning pipelines and quantitative modeling.

B. Feature Generation and Alpha Signal Construction

A symbolic feature generation mechanism is employed in Fig 2, in which operators from categories such as mathematical transformations, time-series filters, cross-sectional ranks, and logical conditions are randomly combined with raw data inputs. To effectively explore the vast space of potential alpha signals, EvoPort employs a randomized hill-climbing optimization process. This heuristic optimization

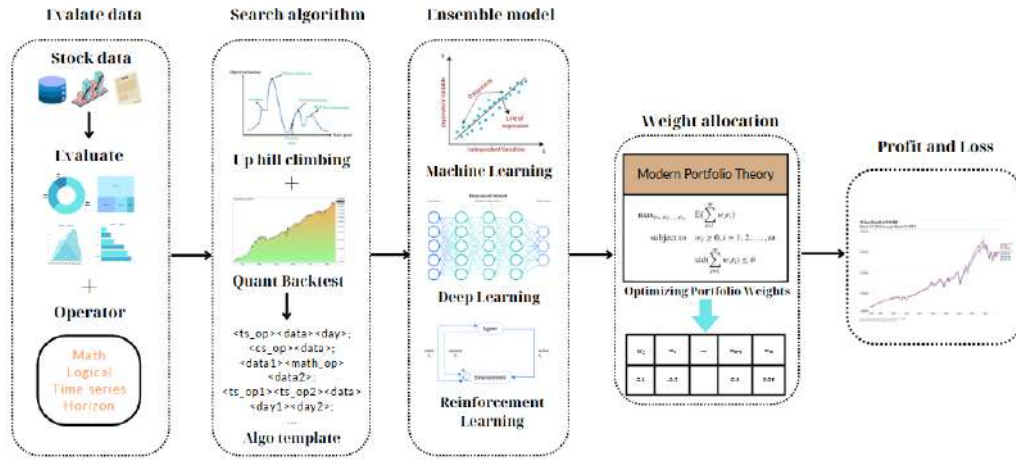


Fig. 1. EvoPort framework

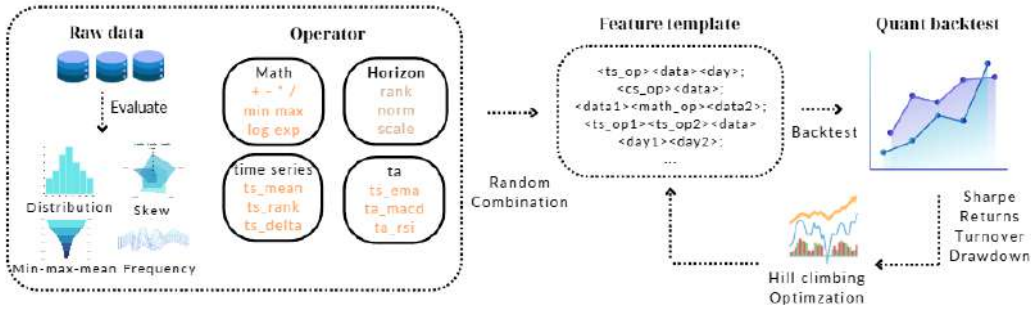


Fig. 2. Finding trading signals

TABLE II
REPRESENTATIVE OPERATORS

No	Category	Example Operators
1	Mathematical	add, mean, max, log
2	Horizontal	rank, quantile, norm
3	Group-Based	group_rank, group_mean
4	Logical	and, or, equal, not
5	Risk Metrics	ts_beta, ts_sharpe
6	Time-Series	ts_mean, ts_rank
7	Transformations	sin, cos, tail
8	Classifier	svm, logistic_regression
9	Clustering	kmeans_group, dbscan_group
10	Technical Indicators	ts_macd, ta_rsi

technique, adapted from the approach developed by Sathiyaraj Chinnasamy (2022) [4], enables efficient search in high-dimensional, discrete feature spaces common in financial modeling. After initial generation of candidate features via symbolic composition, each alpha is evaluated using a backtesting-based fitness function that jointly considers Sharpe ratio, turnover, and drawdown, allowing for more robust and practical signal discovery. In addition to conventional performance metrics, we further incorporate higher-moment and tail-sensitive measures to better capture the distributional characteristics of alpha returns. Specifically, Skew quantifies the asymmetry of the return distribution, indicating whether

gains or losses dominate in the tails. Kurt (kurtosis) measures the “peakedness” and tail heaviness, highlighting the likelihood of extreme outcomes compared to a normal distribution. Tail risk captures the downside behavior under adverse market conditions, reflecting the probability and severity of extreme losses. Together, these metrics provide a more comprehensive assessment of the robustness and practical implementability of discovered alphas beyond mean-variance statistics. To further improve efficiency and prevent overfitting, EvoPort incorporates an adaptive early stopping mechanism that halts the search once validation performance stabilizes. At each iteration, the hill-climbing mechanism introduces a random perturbation to the current alpha expression—either by modifying operators, switching input variables, or adjusting lookback periods. If the new variant improves validation performance, it replaces the original; otherwise, it is discarded.

For example, the process begins with a simple expression such as `rank(returns)`. If validation metrics (e.g., $\text{Sharpe} < 1$, $\text{return/drawdown} < 1$) are unsatisfactory, the ranking function is randomly replaced by alternatives (`quantile`, `scale`, `zscore`) to increase diversity. Each strategy is constrained to ≤ 5 operators and ≤ 5 data inputs, with a maximum of 10 iterations and early stopping once validation performance plateaus. This evolutionary refinement discards suboptimal variants and gradually converges to robust alpha templates.

C. Ensemble Model Selection using Randomized Optimization

In the third stage, the selected alpha signals are fed into a library of candidate models encompassing classical machine learning, statistical regression, deep learning, and reinforcement learning architectures. Machine learning models such as support vector machines, logistic regression, decision trees are employed to identify linear and time-dependent relations. For sequential modeling, deep learning structures such as multi-layer perceptrons, LSTM, BiLSTM, GRU, and their CNN-augmented counterparts are used. Structured radial basis function networks and transformer-based models are used to learn complex interdependencies and latent dynamics. In parallel, model-free reinforcement learning agents such as Deep Q-Network, Proximal Policy Optimization [20], Deep Deterministic Policy Gradient, and Soft Actor-Critic are trained to learn best decision policies in trading environments, with generalizations to multi-agent and hybrid meta-policy setups. Instead of having one model class choice, EvoPort has a stochastic ensemble formation procedure. Stochastic up-hill climbing constructs and optimizes collections of models ensembles sequentially to make progressively increasing improvements according to gains on the validation measure. Every ensemble member contributes an input to be fed into some algorithm that calculates output, and standard ensemble methodologies such as stacking, bagging, and weighted vote are exploited dynamically.

D. Portfolio Weight Optimization and Backtest Evaluation

In the final stage, the ensemble-generated alpha signals are transformed into portfolio allocations through stochastic weight optimization. EvoPort employs extensions of the classical Markowitz mean-variance framework [15]. Rather than deterministically solving for the efficient frontier, the framework samples from multiple weight assignment strategies, including equal weighting, inverse volatility, cardinality-constrained optimization, and probabilistic frontier-based sampling. The resulting portfolios are subjected to rigorous backtesting over historical market data. Key evaluation metrics include cumulative return, Sharpe ratio, maximum drawdown, turnover, and margin utilization. Portfolios that consistently perform well across these metrics, particularly in the out-of-sample window, are retained as candidates for live deployment or further ensemble construction.

Portfolio weights are assigned based on stochastic optimization methods inspired by Modern Portfolio Theory [15]. The weighting strategies include:

- $ir \times e^{-turnover}$: Favors high-IR, low-turnover alphas
- $sigmoid(momentum)$: Favors positive momentum
- $max(expected\_pnl, 0)$: Favors positive expected PnL
- $\frac{ir + \frac{1}{1+volatility} + sigmoid(momentum)}{3}$: Composite score

IV. EXPERIMENTS AND RESULTS

A. Data Preparation

To ensure robustness and fairness in evaluation, we design three distinct experimental scenarios, each tailored to serve a specific objective in our study of the EvoPort framework.

Scenario 1: Full-period optimization for EvoPort. This scenario aims to identify the most effective combinations of feature generation and alpha modeling in EvoPort. We use daily S&P 500 data from January 2016 to March 2025, incorporating price volume, fundamental, option, and macroeconomic indicators. After preprocessing, 1,265 features are generated. The data is split into a training set (2016–2024) and a test set (from 2024 onward).

Scenario 2: Benchmark comparison with prior methods. To ensure a fair and rigorous comparison, we align EvoPort’s evaluation with the test periods and data scopes used in previous studies. For the DRL-PPO baseline from Sood et al., we evaluate EvoPort on S&P 500 from 2012 to 2021. For the KMRF benchmark from Pomorski and Gorse, we adopt S&P 500 and apply EvoPort over the same four-year test window ending April 29, 2022. In both cases, we use the top-performing EvoPort configuration identified in Scenario 1—XGBoost - GRU - PPO—chosen for its strong out-of-sample Sharpe ratio and robust performance. This setup ensures that all models are compared on matching datasets and timelines, enabling a direct, meaningful performance assessment.

Scenario 3: Evaluation of portfolio weighting schemes. Here, we test different portfolio weighting strategies using the fixed model configuration from Scenario 1. Using S&P 500 data from 2016 to March-2025 (train: up to 2024, test: 2024 onward), we apply various rule-based and stochastic weighting formulas (see Table IV) on top of generated signals. This scenario isolates the impact of weighting logic on portfolio returns, Sharpe ratio, and drawdown, while keeping the alpha engine constant. In addition, we compare the performance against three common portfolio optimization approaches: equal weight, low correlation, and low turnover.

All experiments were conducted on a GIGABYTE GAMING A16 CMH laptop, equipped with a 13th Gen Intel Core i7 13620H processor and 16 GB of physical RAM, running a 64 bit Windows 11 environment. The system provided 27.6 GB of total virtual memory and 12 GB of page file space.

B. Alpha Discovery Results

EvoPort successfully discovered 1,947 alphas satisfying stringent performance criteria: Sharpe ratio above 1.5, return-to-drawdown ratio over 2, diverse levels of turnover and minimum margin exceeding 5 basis points. These metrics, widely adopted in quantitative finance and highlighted in Kakushadze’s work [12], reflect both statistical strength and real-world implementability of alpha signals. Compared to related approaches, EvoPort demonstrates a clear quantitative advantage. Genetic Programming [7], under identical data and constraints, produced only 624 qualifying alphas, primarily due to its tendency to overfit and lack of local refinement. Tabu Search [14] generated 487 alphas, but often converged to low-quality regions due to limited diversification in its memory-guided search. Ant Colony Optimization [13], more suited for discrete problems, discovered just 352 alphas with the same thresholds, struggling in continuous symbolic expression

spaces. In contrast, EvoPort’s semantic-aware hill climbing not only scales better with dimensionality but also maintains higher discovery rates with more stable outputs. This confirms EvoPort’s superiority in balancing exploration and precision within complex alpha landscapes. Beyond discovery counts, EvoPort also shows clear advantages in runtime and scalability. On identical hardware, it completed full discovery runs in nearly half the time of Genetic Programming and Tabu Search, and scaled quasi-linearly as feature dictionaries and backtest horizons expanded, whereas baselines exhibited superlinear slowdowns. This efficiency ensures EvoPort remains practical for large-scale alpha mining tasks where both throughput and robustness are critical.

C. Optimize Portfolio Results

The updated experimental findings presented in Table III reveal several important insights into the effectiveness and robustness of various model combinations within the EvoPort framework. **The most effective configuration remains the combination of XGBoost + GRU + PPO**, achieving an out-of-sample Sharpe ratio of 2.70, alongside a robust annual return of 17.50% and a controlled drawdown of 7.00%. This result highlights the powerful synergy between tree-based ensemble learners and sequence models, while leveraging PPO’s policy-based reinforcement learning for dynamic portfolio allocation. While XGBoost + GRU + PPO does not exhibit the highest skew (2.15), kurtosis (22.73), or tail risk (13.44), these values remain within acceptable ranges, it stands as the most reliable choice for further study.

Importantly, the inclusion of related benchmark strategies further strengthens the empirical validation. The SPY market index [23], representing a passive investment benchmark, records an out-of-sample Sharpe ratio of 1.20, with an annual return of 17.00% and a drawdown of 8.00%, reflecting solid market performance under traditional indexing strategies. In contrast, KMRF, a regime-based allocation model leveraging adaptive indicators and random forests [19], achieves a Sharpe ratio of 1.44 with an estimated return of 14.40% and a moderate drawdown of 9.50%, indicating decent performance but limited upside in turbulent markets. Meanwhile, DRL-PPO [20], which employs deep reinforcement learning to directly learn portfolio policies, reports a Sharpe ratio of 1.16 with a return of 12.10%, but suffers from a substantial drawdown of 33.00% and relatively high turnover of 8.00%, reflecting potential instability under volatile conditions. By comparison, EvoPort achieves a higher Sharpe ratio of 2.2, a return of 16.30%, and maintains a low drawdown of only 6.00% with turnover of 5.00% during the 2018–2022 period, consistently outperforming both baselines in terms of robustness, risk control, and efficiency.

The results in Table IV illustrate the performance of ten portfolio-weighting schemes, each embedding a distinct view of risk, trading cost, and market behaviour. The formulas are applied on top of the same alpha ensemble; only the weighting logic changes the final position sizing. Among all rules,

the composite score $\frac{ir + \frac{1}{1+volatility} + \text{sigmoid}(\text{momentum})}{3}$ remains a top performer, delivering the second-highest out-of-sample Sharpe of 2.90, a returns of 22.20%, and a drawdown of 11.30%. Blending information ratio, volatility control, and momentum strength into a single score yields a holistic assessment of alpha quality and offsets the weaknesses of any individual metric. The best result overall comes from the exponential penalty formulation $ir \cdot e^{-turnover}$, which achieves the highest out-of-sample Sharpe of 3.00, return of 26.70%, and drawdown of 11.90%, validating the advantage of explicitly accounting for transaction cost when sizing exposures. To contextualize performance, we compare against three classic weighting approaches from prior literature: Equal Weight, Low Correlation and Low Turnover. While these benchmarks are widely used, they underperform across all metrics in our setup. Equal Weight achieves an out-of-sample Sharpe of 2.80, while Low Corr and Low Turnover reach only 2.50 and 2.30, respectively—significantly lower than the top adaptive schemes. This confirms that static rules, although interpretable and robust in theory, fail to match the effectiveness of our risk-aware, cost-sensitive weighting functions in practice.

Beyond conventional metrics, the additional distributional measures (Skew, Kurt, and Tail) provide deeper insights into the robustness of portfolio strategies. As reported in Tables III and IV, EvoPort configurations generally maintain moderate skewness (2.5–3.2), indicating balanced upside asymmetry rather than excessive reliance on rare positive jumps. Kurtosis values, while elevated (16–24), are systematically lower than several benchmark rules, suggesting that EvoPort portfolios avoid excessive concentration in extreme outliers. Most importantly, Tail risk remains consistently contained—best configurations such as Random Forest + GRU + PPO achieve the highest Sharpe ratios while limiting extreme downside to manageable levels.

V. CONCLUSION

In this paper, we introduced EvoPort, a modular portfolio optimization framework that unifies randomized alpha discovery, ensemble model selection, and stochastic weighting into a scalable pipeline. Our experiments show consistent outperformance, with out-of-sample Sharpe ratios above 2.0 across diverse model combinations and weighting schemes. The use of randomized ensembles and dynamic weighting plays a critical role in enhancing robustness and mitigating overfitting. EvoPort’s flexible architecture supports the integration of machine learning, deep learning, and reinforcement learning, reinforcing the value of hybrid approaches in quantitative finance. As a result, EvoPort serves as a solid foundation for developing adaptive, scalable, and data-driven portfolio solutions. In future work, we aim to extend EvoPort to multi-asset and multi-frequency settings. Finally, applying EvoPort in live or simulated trading environments with transaction costs and real-world constraints will be essential to evaluate its practical viability.

TABLE III
PORTFOLIO PERFORMANCE ACROSS MODEL COMBINATIONS

Method	Dataset (S&P500)	IS Sharpe	OS Sharpe	Returns	Drawdown	Turnover	Kurt	Skew	Tail
XGBoost + GRU + PPO	2016–2025	3.50	2.70	17.50%	7.00%	9.00%	14.24	2.98	15.14
SVM + CNN-LSTM + SAC		3.00	2.20	14.80%	6.00%	10.00%	19.62	3.15	13.45
Random Forest + VMD-BiLSTM		2.90	2.20	14.00%	8.00%	8.00%	16.51	2.70	11.76
Decision Tree + LSTM + DDPG		2.70	2.10	13.00%	7.00%	11.00%	18.43	2.12	14.24
Logistic Regression + BiLSTM + MAPPO		3.20	2.50	16.20%	6.00%	9.00%	21.52	3.01	17.32
Ridge Regression + CNN-LSTM + PPO		3.10	2.40	15.80%	7.00%	8.00%	12.84	2.59	13.12
OLS + GRU		2.80	2.00	12.50%	7.00%	7.00%	14.99	2.44	11.03
XGBoost + VMD-BiLSTM + SAC		3.30	2.60	16.50%	6.00%	9.00%	20.32	3.23	18.04
SVM + LSTM + DDPG		2.90	2.20	13.80%	7.00%	8.00%	19.57	2.81	16.77
Random Forest + GRU + PPO		3.10	2.50	15.90%	6.00%	10.00%	22.44	2.91	19.68
SPY (Benchmark)		0.80	1.20	17.00%	8.00%	9.00%	9.12	3.24	12.34
EvoPort	2018–2022	1.90	2.20	16.30%	6.00%	5.00%	15.64	3.02	10.72
KMRF [19]		1.35	1.44	14.40%	9.50%	7.50%	13.23	2.90	11.98
EvoPort	2012–2021	2.00	2.30	16.10%	7.00%	6.00%	19.25	2.82	14.23
DRL-PPO [20]		1.20	1.16	12.10%	33.00%	8.00%	12.43	3.15	10.68

TABLE IV
PERFORMANCE OF DIFFERENT PORTFOLIO WEIGHTING SCHEMES

Method	Weighting Formula	IS Sharpe	OS Sharpe	Returns	Drawdown	Turnover	Kurt	Skew	Tail
EvoPort	$ir \cdot e^{-turnover}$	3.80	3.00	26.70%	11.90%	7.00%	22.73	2.15	13.44
	$\frac{1}{1+avg\_corr}$	3.10	2.50	23.10%	10.80%	8.00%	19.41	3.04	15.09
	$\frac{1}{1+volatility}$	3.00	2.40	22.70%	11.80%	6.00%	18.37	2.56	14.21
	$sigmoid(momentum)$	2.90	2.30	21.30%	12.40%	9.00%	16.88	1.72	12.67
	$1 - 0.5 - long_short_ratio $	3.50	2.80	21.00%	10.10%	7.00%	20.14	2.40	10.94
	$if\_else(zscore < 2, 1, 0)$	3.30	2.70	23.80%	10.30%	7.00%	17.92	1.99	14.35
	$max(expected\_pnl, 0)$	3.00	2.20	19.90%	15.00%	6.00%	23.58	2.65	18.16
	$\frac{ir + \frac{1}{1+volatility} + sigmoid(momentum)}{3}$	3.60	2.90	22.20%	11.30%	8.00%	21.34	2.89	16.22
	$\frac{ir\_short + ir\_long}{2}$	3.20	2.40	25.80%	8.40%	7.00%	24.12	2.52	19.78
	$\frac{1}{1+drawdown}$	3.40	2.80	25.10%	8.60%	8.00%	22.81	3.15	17.49
Benchmark	Equal Weight	3.50	2.70	17.50%	7.00%	9.00%	19.65	2.99	15.92
	Low Corr	3.30	2.50	21.00%	10.00%	9.00%	22.18	3.06	16.53
	Low Turnover	3.10	2.30	19.00%	12.00%	4.00%	14.77	1.87	13.28

REFERENCES

- [1] David H. Bailey, Marcos Lopez de Prado, and Q. Malik. "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance". In: *Notices of the AMS* 61.5 (2014), pp. 458–471.
- [2] Dimitris Bertsimas and Despina A. Pachamanova. "Robust portfolio optimization and management". In: *Operations Research* 57.5 (2009), pp. 1120–1139.
- [3] Fischer Black. "Global portfolio optimization". In: *Financial Analysts Journal* 48.5 (1992), pp. 28–43.
- [4] Sathiyaraj Chinnasamy et al. "A review on hill climbing optimization methodology". In: *Recent Trends in Management and Commerce* 3.1 (2022), pp. 1–7.
- [5] Lisa Ehrlinger, Elisa Rusz, and Wolfram WöB. "A survey of data quality measurement and monitoring tools". In: *arXiv preprint arXiv:1907.08138* (2019).
- [6] Yuxing Feng, Qing He, and Patrick Orbanz. "Deep learning for financial applications: A survey". In: *arXiv preprint arXiv:1808.09940* (2018).
- [7] Diandra Chika Fransisca et al. "Robust Portfolio Mean-Variance Optimization for Capital Allocation in Stock Investment Using the Genetic Algorithm: A Systematic Literature Review". In: *Computation* 12.8 (2024), p. 166.
- [8] Shuai Gu, Bryan Kelly, and Dacheng Xiu. "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5 (2020), pp. 2223–2273.
- [9] Ben Hambly and Lilian Yang. "Recent Advances in Reinforcement Learning in Finance". In: *arXiv preprint arXiv:2112.04553* (2021).
- [10] Campbell R. Harvey, Yan Liu, and He Zhu. "... and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1 (2016), pp. 5–68.
- [11] Xin He, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A survey of the state-of-the-art". In: *Knowledge-Based Systems* 212 (2021), p. 106622.
- [12] Zura Kakushadze. "101 formulaic alphas". In: *Wilmott* 2016.84 (2016), pp. 72–81.
- [13] Chiranun Kamolsin and Porawat Visutsak. "Solving Portfolio Optimization Problem for Long-term Stocks Investment using Ant Colony Optimization". In: *Proceedings of the 2024 9th International Conference on Intelligent Information Technology*. 2024.
- [14] Woo Sik Lee. "A Study on Portfolios Using Simulated Annealing and Tabu Search Algorithms". In: *Journal of The Korean Society of Industry Convergence* 27.2_2 (2024), pp. 467–473.
- [15] Harry Max Markowitz. "Modern portfolio theory". In: *Journal of Finance* 7.11 (1952), pp. 77–91.
- [16] Attilio Meucci. "Bayesian techniques for improving portfolio analysis". In: *Risk* 18.10 (2005), pp. 68–72.
- [17] Szilárd Pafka and Imre Kondor. "Estimated correlation matrices and portfolio optimization". In: *Physica A: Statistical Mechanics and its Applications* 343 (2004), pp. 623–634.
- [18] Yuliya Plyakha, Raman Uppal, and Grigory Vilkov. "Why does an equal-weighted portfolio outperform value-and price-weighted portfolios?" In: *Available at SSRN* 2724535 (2012).
- [19] Piotr Pomorski and Denise Gorse. "Improving Portfolio Performance Using a Novel Method for Predicting Financial Regimes". In: *International Conference on Machine Learning, Optimization, and Data Science*. Springer Nature Switzerland, 2023.
- [20] Srijan Sood. "Deep Reinforcement Learning for Optimal Portfolio Allocation: A Comparative Study with Mean-Variance Optimization". In: *FinPlan* 2023. 2023, p. 21.
- [21] Srijan Sood et al. "Information horizon, portfolio turnover, and optimal alpha models". In: *The Journal of Portfolio Management* 34.1 (2007), pp. 27–40.
- [22] The Investopedia Team. *7 Technical Indicators To Build a Trading Tool Kit*. Investopedia. Jan. 2025. URL: <https://www.investopedia.com/top-7-technical-analysis-tools-4773275> (visited on 07/13/2025).
- [23] Yahoo Finance. *SPY Historical Data*. <https://finance.yahoo.com/quote/SPY/history>. Accessed: 2025-06-24.

Simulation-Optimization for Healthcare Supplier Selection Problem under Intuitionistic Fuzzy Environment

Nguyen Ngoc Kim Ngan
Industrial Systems Engineering
Department
International University-National
University HoChiMinh City, VietNam
Email: nnkngan.w@gmail.com

Ha Thi Xuan Chi
Industrial Systems Engineering
Department
International University-National
University HoChiMinh City, VietNam
Email: httxchi@hcmiu.edu.vn

Cao Thanh Truc
Industrial Systems Engineering
Department
International University-National
University HoChiMinh City, VietNam
Email: cthanhtruc911@gmail.com

Abstract— The global healthcare sector has expanded rapidly in recent years, fueled by rising public awareness, globalization, and population growth. While many studies have utilized Multi-Criteria Decision-Making (MCDM) methods for supplier selection, most fail to adequately address the complexities of evaluating multi-sourcing impacts. This study proposes a three-stage methodology to address these limitations. First, Data Envelopment Analysis (DEA) is used to identify efficient suppliers. Second, an MCDM method is applied within an intuitionistic fuzzy environment to score the shortlisted suppliers and manage uncertainties in performance evaluations. Finally, a simulation-optimization model is used to enhance decision-making under uncertain conditions, such as supply disruptions, delivery delays, and variable quality. The ultimate objective is to identify the most suitable suppliers to strengthen the healthcare supply chain, thereby improving delivery service, enhancing patient satisfaction, and achieving cost savings.

Keywords— Healthcare Supplier Selection, Data Envelopment Analysis, Intuitionistic Fuzzy Environment, Multi-Criteria Decision-Making, Genetic Algorithm, Simulation-Optimization.

I. INTRODUCTION

In today's healthcare systems, the availability of essential medical products on time is important to save lives and enhance population welfare [1]. Especially in Covid-19 pandemic, it has shown how fragile and critical the healthcare supply chain is in the face of disruption [2]. According to Ivanov & Dolgui [3], Delivery delays, shortages or suppliers out of business can have caused chain effects in the quality of treatment and the resilience of the system. Consequently, optimizing suppliers' selection is becoming a concerning strategy in the medical sector where high product variability, short life cycles and unpredicted demand create more complexity [4]. In this context, Supplier selection refers to the systematic procedure of evaluating and selecting suppliers based on multiple conflicting criteria, including cost, delivery reliability, quality, and risk factors, often under conditions of uncertainty [5].

Recent studies on Intuitionistic Fuzzy Sets (IFS) have gained attention as they allow for modeling hesitation in human decision-making by including degrees of different criteria [7]. Additionally, simulation-based approaches have been employed to represent supply chain disruptions and analyze their impacts on procurement and inventory decisions [8]. Despite these advances, however, most existing frameworks fail to integrate disruption-aware supplier evaluation with inventory management in a unified, continuous innovation context [8]. According to Hamdan &

Cheaitou [4], traditional MCDM approaches are not sufficiently equipped to handle supply disruptions, fluctuating lead times, or practical demand changes. Furthermore, simulation-optimization techniques are rarely linked to earlier stages of supplier evaluation, limiting their ability to guide sourcing decisions in the volatile healthcare environment [6].

This study proposes a hybrid decision-making framework that integrates Data Envelopment Analysis (DEA), Intuitionistic Fuzzy MCDM (IF-CoCoSo), and simulation-optimization to address the pharmaceutical supplier selection problem under uncertainty. The first stage employs DEA to screen suppliers based on financial efficiency. The second stage applies IF-MCDM to evaluate qualitative supplier performance under uncertain group decision-making. Finally, a simulation-optimization model is developed to incorporate stochastic disruptions, such as lead time variability and demand uncertainty, while optimizing order allocation and inventory costs. The model is implemented to a case study involving 25 Indian pharmaceutical suppliers using real financial data and expert input. This approach contributes to more effectiveness and resilience to disruption supplier selection in the healthcare sector.

II. MATHEMATICS

A. Assumption

- Inventory management component adheres to a (Q, R) policy.
- Each supplier operates with a dedicated TL vehicle of capacity u_j .
- Every batch ordered Q_{ij} may contain a proportion k_j of imperfect items.
- Imperfect items are stored separately and only released for sale in the next procurement cycle.
- Screening of items is assumed to be 100% effective and instantaneous; screening time is ignored.
- Plants incur external failure costs a_i for any defective product that reaches customers.
- Unsatisfied demand due to delayed replenishment or disruption leads to lost sales.
- 8 hours per day and 250 days per year.
- The distance between a plant and a supplier was computed using the Euclidean distance.

B. Mathematical model

Phase 1: Determining supplier scores

Data Envelopment Analysis

Data Envelopment Analysis (DEA) employs a non-parametric linear programming framework to evaluate how effectively decision-making units (DMUs) utilize multiple inputs to produce outputs. The method generates an efficiency frontier from actual performance, enabling the identification of efficient and inefficient DMUs.

Based on Chakraborty et al. (2024) [10], the LP models are formulated by treating criterion CCC as the sole input (non-beneficial) and the remaining criteria as outputs (beneficial). Among twenty-five anonymous healthcare suppliers, those who had the efficiency scores at 100% were considered the most efficient performers and were selected for scoring and ranking in the next step to identify the best option in an uncertain setting using five qualitative evaluation criteria.

IF-MCDM

According Chakraborty et al. (2024) [10], IF-CoCoSo, an extension of the original CoCoSo (Combined Compromise Solution) method, is aggregated with intuitionistic fuzzy (IF) sets to better resolve uncertainty and imprecision in decision-making processes. The following is an expression for an IFS in a finite set X:

$$A = \{ \langle x, \mu_A(x), \nu_A(x), \pi_A(x) \rangle \mid x \in X \}$$

where $\mu_A(x)$: membership function; $\nu_A(x)$: non-membership function and $\pi_A(x)$: hesitancy function:

$X \rightarrow [0,1]$. Then, it can be expressed that:

$$0 \leq \mu_A(x) + \nu_A(x) \leq 1; \pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$$

$$0 \leq \pi_A(x) \leq 1$$

IF-CoCoSo method is applied following each step:

Step 1: Gather the relative importance of evaluation criteria and the performance of alternatives with respect to each criterion, expressed in linguistic terms.

Step 2: Create an aggregated IF decision matrix from the language rating matrix for criteria and decision makers.

Step 3: Based on the corresponding IFNs (IF numbers) and using the below equation to obtain the corresponding relative weights of each decision maker.

$$\lambda_k = \frac{\left(\mu_k + \pi_k \left(\frac{\mu_k}{\mu_k + \nu_k} \right) \right)}{\sum_{k=1}^l \left(\mu_k + \pi_k \left(\frac{\mu_k}{\mu_k + \nu_k} \right) \right)} \text{ with } \sum_{k=1}^l \lambda_k = 1 \quad (1)$$

Step 4: Calculate the criteria's IF weights using the associated language ratings

$$\tilde{w} = \{ \tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_j, \dots, \tilde{w}_n \}$$

where

$$\begin{aligned} \tilde{w}_j &= IFWA_j (w_j^{(1)}, w_j^{(2)}, \dots, w_j^{(l)}) \\ &= \lambda_1 w_j^{(1)} \oplus \lambda_2 w_j^{(2)} \oplus \dots \oplus \lambda_l w_j^{(l)} \end{aligned}$$

$$= \left[1 - \prod_{k=1}^l (1 - \mu_j^{(k)})^{\lambda_k}, \prod_{k=1}^l (\nu_j^{(k)})^{\lambda_k}, \prod_{k=1}^l (1 - \mu_j^{(k)})^{\lambda_k} - \prod_{k=1}^l (\nu_j^{(k)})^{\lambda_k} \right] \quad (2)$$

Step 5: Convert linguistic rating matrix of the evaluation of each alternative in terms of each criterion by decision makers (DMs) into aggregated IF decision matrix:

$$\tilde{A} = \{ \tilde{r}_{ij} \}$$

where

$$\begin{aligned} \tilde{r}_{ij} &= IFWA_{\lambda} (r_{ij}^{(1)}, r_{ij}^{(2)}, \dots, r_{ij}^{(l)}) \\ &= \lambda_1 r_{ij}^{(1)} \oplus \lambda_2 r_{ij}^{(2)} \oplus \dots \oplus \lambda_l r_{ij}^{(l)} \\ &= \left[1 - \prod_{k=1}^l (1 - \mu_{ij}^{(k)})^{\lambda_k}, \prod_{k=1}^l (\nu_{ij}^{(k)})^{\lambda_k}, \prod_{k=1}^l (1 - \mu_{ij}^{(k)})^{\lambda_k} - \prod_{k=1}^l (\nu_{ij}^{(k)})^{\lambda_k} \right] \end{aligned} \quad (3)$$

Step 6: Combine the criteria weights with the alternative ratings to derive the weighted aggregate.

$$\tilde{A}' = \{ \tilde{x}'_{ij} \}$$

where $\tilde{x}'_{ij} = \tilde{w}_j \otimes \tilde{r}_{ij} = \langle \mu'_{ij}, \nu'_{ij}, \pi'_{ij} \rangle$

By using the below equation

$$A \otimes B = \{ \mu_A(x) \cdot \mu_B(x), \nu_A(x) + \nu_B(x) - \nu_A(x) \cdot \nu_B(x) \mid x \in X \} \quad (4)$$

Step 7: Compute the weighted sum model (WSM) value of each alternative:

$$\tilde{S}_i = \sum_{j=1}^n \tilde{x}'_{ij} \quad (1)$$

Step 8: Calculate weighted product model (WPM) for each alternative:

$$\tilde{P}_i = \prod_{j=1}^n \tilde{x}'_{ij} \quad (6)$$

Step 9: Compute each $\tilde{S}_i$ and $\tilde{P}_i$ (score function value)

$$s_i = \frac{\mu_{\tilde{S}_i} - \nu_{\tilde{S}_i} + 1}{2} \quad (7)$$

$$p_i = \frac{\mu_{\tilde{P}_i} - \nu_{\tilde{P}_i} + 1}{2} \quad (8)$$

Step 10: Use the three assessment score procedures listed below to estimate the alternatives' relative weights:

$$k_{ia} = \frac{s_i + p_i}{\sum_{i=1}^m (s_i + p_i)} \quad (9)$$

$$k_{ib} = \frac{s_i}{\min_i s_i} + \frac{p_i}{\min_i p_i} \quad (10)$$

$$k_{ic} = \frac{\gamma s_i + (1-\gamma) p_i}{\gamma \times \max_i s_i + (1-\gamma) \times \max_i p_i} \quad (11)$$

($0 \leq \gamma \leq 1$)

where

k_{ia} : The mean value obtained by averaging the WSM and WPM scores of each alternative.

k_{ib} : Sum of relative WSM and WPM scores compared to the best.

k_{ic} : Balanced compromise of WSM and WPM scores

Step 11: Compute the assessment value (k_i) of each alternative:

$$k_i = (k_{ia} + k_{ib} + k_{ic})^{\frac{1}{3}} + \frac{(k_{ia} + k_{ib} + k_{ic})}{3} \quad (12)$$

Step 12: Rank the alternatives in which the best one has the maximum k_i value

In this stage, supplier performance is evaluated by using this technique to determine the score of shortlisted suppliers. Based on five qualitative criteria (service, delivery time, flexibility, cost, and quality), three decision-makers (DMs) who have diverse experience and knowledge will evaluate selected suppliers. The three decision-makers include:

- DM₁: a 52-year-old female physician who has been working in a government healthcare unit for over 20 years.
- DM₂: a 45-year-old male who has extensive experience in the pharmacy department, especially 15 years of experience in order placement, fulfillment, stock keeping, supplier liaison, and quality checks
- DM₃: a 39-year-old male has the same experience with DM₂ with experience for 10 years.

The evaluations from participating DMs are classified as very important (VI), important (I), and medium (M), corresponding to λ_1 , λ_2 , and λ_3 respectively.

TABLE I. DECISION MAKERS' LINGUISTIC RATINGS FOR THE CRITERIA

Linguistic term	Symbol	μ	ν	π
Very Important	VI	0.88	0.07	0.03
Important	I	0.70	0.25	0.05
Medium	M	0.50	0.45	0.05
Unimportant	U	0.25	0.70	0.05
Very Unimportant	VU	0.07	0.88	0.03

The linguistic ratings in Table 1 are employed to evaluate supplier performance under the five criteria: service, delivery time, flexibility, cost, and quality.

TABLE II. THE ALTERNATIVES' LINGUISTIC RATING

Linguistic term	Symbol	μ	ν	π
Extremely good	EG	1.00	0.00	0.00
Very very good	VVG	0.85	0.10	0.05
Very good	VG	0.80	0.15	0.05
Good	G	0.70	0.20	0.10
Medium good	MG	0.60	0.30	0.10
Fair	F	0.5	0.4	0.1
Medium bad	MB	0.40	0.50	0.10
Bad	B	0.25	0.60	0.15
Very bad	VB	0.10	0.75	0.15
Very very bad	VVB	0.10	0.90	0.00

Through the mathematical process of IF-MCDM techniques, these linguistic ratings of suppliers' performance then are converted to IFNs, combined and formed the aggregated decision matrix.

TABLE III. CONSIDERED CRITERIA WITH LINGUISTIC RATINGS

Decision Maker	Cost	Delivery Time	Service	Quality	Flexibility
DM1	VI	I	I	I	M
DM2	VI	I	M	I	M
DM3	I	VI	I	VI	M

After this stage, the aggregated criteria weights calculated are multiplied by the aggregated decision matrix to obtain the value of the weighted aggregated decision matrix - which is considered as the input of the IF-MCDM methods. The given tables below are DMs' evaluations of the criteria and alternatives' performance with linguistics terms.

TABLE IV. ALTERNATIVES AGAINST DIFFERENT CRITERIA WITH LINGUISTIC RATINGS

Decision Maker	Healthcare Supplier	Criteria				
		Cost	Delivery Time	Service	Quality	Flexibility
DM1	A1	MG	MG	G	F	F
	A2	MG	G	MB	F	B
	A3	VVG	VG	VG	G	G
	A4	G	MG	G	MG	MG
DM2	A1	MG	MG	G	F	MB
	A2	G	G	F	F	MB
	A3	VG	G	VG	VVG	G
	A4	VG	MG	G	G	MG
DM3	A1	MG	G	F	F	F
	A2	G	MG	MB	F	MB
	A3	VG	G	G	MG	MG
	A4	G	G	MG	MG	VVG

Following the IF-CoCoSo procedure, the weighted aggregated decision matrix from step 5 is used to compute the Weighted Sum Model (WSM) and Weighted Product Model (WPM) values. Then, the related score function value for each WSM and WPM are subsequently computed. Through three evaluation scoring strategies, the alternatives' relative weights are determined, and then all final evaluation values for shortlisted healthcare suppliers are computed. These values are used as parameters in the simulation-optimization stage, specifically in the first objective - Z_1 .

Phase 2: Simulation-Optimization

This study adopts the mathematical model of Saputro et al. (2020) [9], with the following parameters and decision variables used to formulate the model

Indices:

i: plant, $i = 1, 2, \dots, n$

j: supplier, $j = 1, 2, \dots, n$

Parameters:

- Plant i

$E[D_i]$: Expected demand

o_i : Setup costs

s_i : Shortage costs (per uni)

a_i : imperfect items' external failure costs (per unit)

h_i : perfect items' holding costs (per unit)

h_i' : imperfect items' holding costs (per unit)

- Supplier j

$E[k_j]$: Expected imperfect quality (rate)

f_j : Fixed contractual costs (annual)

c_j : Purchasing costs (per unit)

w_j : Capacity (annual)

u_j : Vehicle capacity

θ_j : Disruption frequency (rate)

v_j : Disruption duration

SS_j : Supplier j 's score (assessment value k_i)

$E[LTD_{ij}]$: Expected lead time demand plant i & supplier j
 $\eta[LTD_{ij}, R_{ij}]$: Standardised loss function plant i & supplier j
 p_{ij} : Fixed transportation costs per replenishment from supplier j to plant i
 r_{ij} : Transportation costs per mile and per replenishment from supplier j to plant i
 d_{ij} : Distance from plant i to supplier j
 l_{ij} : Lead time plant i & supplier j

Decision variables:

X_j : $\begin{cases} 1, & \text{if supplier } j \\ 0, & \text{otherwise} \end{cases}$

Y_{ij} : Purchase amount allocated by plant i to supplier j

Q_{ij} : Order quantity of plant i to supplier j

R_{ij} : Reorder point of plant i to supplier j

Model formulation:

The objective function is expressed as:

- Total value of purchasing:

$$\text{Max } Z_1(TVP) = \sum_i^n \sum_j^m SS_j Y_{ij} \quad (13)$$

- Total cost:

$$\text{Min } Z_2(TC) = \sum_{j=1}^m f_j X_j + \sum_{i=1}^n \sum_{j=1}^m c_j Y_{ij} \quad (14a)$$

$$+ \sum_{i=1}^n \sum_{j=1}^m \frac{o_i Y_{ij}}{Q_{ij}(1-E[k_j])} + \sum_{i=1}^n \sum_{j=1}^m h_i \left(\frac{Q_{ij}(1-E[k_j])}{2} + R_{ij} - E[LTD_{ij}] \right) \quad (14b-1)$$

$$+ \sum_{i=1}^n \sum_{j=1}^m s_i \eta(LTD_{ij}, R_{ij}) \frac{Y_{ij}}{Q_{ij}(1-E[k_j])} \quad (14b-1)$$

$$+ \sum_{i=1}^n \sum_{j=1}^m \frac{(p_{ij} + r_{ij} d_{ij}) \left(\frac{Q_{ij}}{u_j} \right) Y_{ij}}{Q_{ij}(1-E[k_j])} \quad (14c)$$

$$+ \sum_{i=1}^n \sum_{j=1}^m a_i Y_{ij} E[k_j] + \sum_{i=1}^n \sum_{j=1}^m h'_j Q_{ij} E[k_j] \quad (14d)$$

Subject to:

$$\sum_{j=1}^m Y_{ij} = E[D_i], \quad \forall i \in I \quad (15)$$

$$\sum_{i=1}^n Y_{ij} \leq b_j X_j, \quad \forall j \in J \quad (16)$$

$$Y_{ij} \geq 0, Q_{ij} \geq 0, X_j = 0 \text{ or } 1, \forall i \in I, \forall j \in J \quad (17)$$

Objective function (13) incentivizes allocations that favor suppliers with superior performance metrics while considering the actual distribution of procurement quantities across all plant-supplier pairs. This objective (14a) captures the total procurement expenses, comprising two primary components: (i) the fixed annual contractual fee incurred upon selecting a supplier, and (ii) the variable purchasing cost tied to the unit price of each item ordered. Moving to (14b-1), this portion of the objective accounts for setup costs per replenishment cycle, along with inventory holding costs, which are computed based on expected inventory levels

adjusted for imperfect quality rates. The second part of (14b), shown as (14b-2), reflects the penalty for shortages, representing the cost of unmet demand due to stockouts over time. Equation (14c) quantifies the transportation expenses, which are divided into fixed logistics costs such as loading and unloading per shipment and variable costs dependent on delivery distance between suppliers and plants. Finally, (14d) addresses the costs associated with imperfect quality, including both the external failure penalties (e.g., customer returns, complaint handling) and the additional holding costs of storing imperfect items separately from usable stock.

To solve the objective function, certain conditions are required to hold, which are specified through the following set of constraints. Equation (15) satisfies plant demand via allocations Y_{ij} . Moreover, equation (16) to ensure the order allocation Y_{ij} within the suppliers' capacity b_j . Lastly, equation (17) non-negativity and binary decision variables.

According to Saputro et al. (2021), this research conducts multi-objective approach by each step as bellow:

Step 1: The model with multiple objectives is separated into two single-objective sub-problems.

The first is determined using the objective function in Eq. (13) and constraints in Eqs. (15), (16), and (17)

Subject to the same constraints, the objective function in Eqs. (14a)–(14d) makes up the second sub-problem.

An S-O technique is used to solve these two sub-problems independently to determine optimal solutions, $Z_{1\max}$ and $Z_{2\min}$.

Step 2: The deviation (e) of the objective function from the ideal solution Z^* ($Z^* = Z_{\max}$ and $Z^* = Z_{\min}$) is calculated using the distance method.

$$f_1 = \frac{Z_{1\max} - Z_1}{Z_{1\max}} \quad (18)$$

$$f_2 = \frac{Z_2 - Z_{2\min}}{Z_{2\min}} \quad (19)$$

Finally, the two objective functions are combined into one by applying the Weighted Comprehensive Criterion Method (WCCM), aiming to minimize the total deviation (e). Each objective function is assigned the importance weights (α_1, α_2). The following is the expression for a single objective function:

$$\text{Min } e = \alpha_1 \cdot f_1 + \alpha_2 \cdot f_2 \quad (20)$$

$$\text{and } \alpha_1 + \alpha_2 = 1$$

Subject to:

Equation (15), (16), (17)

Simulation – Optimization Proposed Approach:

The multi-objective model is divided into two single-objective sub-problems and each of them is solved individually to get their best ideal solutions. For a given X , the order allocation Y is obtained using Vogel's Approximation Method (VAM) combined with the Modified Distribution Method (MODI). Once the values of Y are determined, order quantity (Q) and reorder point (R) are also computed. Selected supplier selection (X) as well as inventory decisions (Q, R) are then subsequently input into the next steps to simulate the uncertain demand and potential disruptions including AME method for the result of second objective function Z_2 .

III. COMPUTATIONAL RESULTS

Data Envelopment Analysis result

TABLE V. THE RESULT TABLE OF DEA

Supplier	Efficiency	Supplier	Efficiency
1	73.2493	14	70.1257
2	81.8867	15	32.8518
3	60.7663	16	100
4	72.0248	17	100
5	100	18	48.6448
6	100	19	70.6225
7	34.9802	20	37.6268
8	75.1580	21	87.2227
9	77.7357	22	55.4754
10	32.9134	23	67.3718
11	9.0632	24	66.6373
12	84.6500	25	20.1933
13	81.4943		

With the application of the DEA method, evaluate 25 suppliers' financial aspects and choose those that have the efficiency score at 100%. Based on the result table V, suppliers 5, 6, 16 and 17 are those who have the score at 100%, meaning that they will be selected for the MCDM assessment step to get the score and prepare for the number of suppliers in the simulation-optimization approach, which is 4 suppliers

Multi-criteria Decision-Making result

The MCDM method – CoCoSo integrated with IFS (Intuitionistic Fuzzy Set) is used to determine the suitable pharmaceutical supplier with their assessment from the four shortlisted efficient choices. As previously mentioned, 3 DMs with different knowledge and experience will evaluate shortlisted healthcare suppliers based on 5 criteria, including service, delivery time, flexibility, cost, and quality. Through the very first step, the relative weights of 3 DMs are estimated using a computation procedure with the results as shown in table VI.

TABLE VI. RELATIVE WEIGHTS OF DMs

	DM1	DM2	DM3
λ	0.4182	0.3394	0.2424

Based on the commutated relative weights of those DMs and their assessment of all the five qualitative importance, the aggregated criteria weights are estimated.

TABLE VII. AGGREGATED CRITERIA WEIGHTS

Criteria	μ	ν	π
Cost	0.846	0.109	0.045
Delivery Time	0.791	0.16	0.049
Service	0.684	0.263	0.053
Quality	0.626	0.321	0.054
Flexibility	0.5	0.45	0.05

The linguistic ratings of each DM for 4 healthcare suppliers will be transformed to IFNs and the aggregated decision matrix is then formed by that combination as the table below.

TABLE VIII. AGGREGATED DECISION MATRIX

Healthcare Supplier	Cost	Delivery Time	Service
A1	(0.6, 0.3, 0.1)	(0.627, 0.272, 0.101)	(0.66, 0.237, 0.103)
A2	(0.662, 0.237, 0.101)	(0.678, 0.221, 0.101)	(0.436, 0.464, 0.1)
A3	(0.823, 0.127, 0.051)	(0.747, 0.177, 0.076)	(0.779, 0.161, 0.06)
A4	(0.739, 0.181, 0.08)	(0.627, 0.272, 0.101)	(0.678, 0.221, 0.101)
Healthcare Supplier	Quality	Flexibility	
A1	(0.5, 0.4, 0.1)	(0.468, 0.431, 0.1)	
A2	(0.5, 0.4, 0.1)	(0.341, 0.54, 0.119)	
A3	(0.746, 0.174, 0.08)	(0.678, 0.221, 0.101)	
A4	(0.637, 0.261, 0.101)	(0.685, 0.23, 0.085)	

After computation, the aggregated decision matrix is multiplied by the criteria weights to form the weighted matrix, which serves as the input for the MCDM method.

TABLE IX. WEIGHTED AGGREGATED DECISION MATRIX

Healthcare Supplier	Cost	Delivery Time	Service
A1	(0.508, 0.376, 0.116)	(0.496, 0.388, 0.116)	(0.451, 0.438, 0.111)
A2	(0.56, 0.32, 0.12)	(0.536, 0.346, 0.118)	(0.298, 0.605, 0.097)
A3	(0.696, 0.222, 0.082)	(0.591, 0.309, 0.1)	(0.533, 0.382, 0.085)
A4	(0.625, 0.27, 0.105)	(0.496, 0.388, 0.116)	(0.464, 0.426, 0.11)
Healthcare Supplier	Quality	Flexibility	
A1	(0.313, 0.593, 0.094)	(0.234, 0.687, 0.079)	
A2	(0.313, 0.593, 0.094)	(0.171, 0.747, 0.082)	
A3	(0.467, 0.439, 0.094)	(0.339, 0.572, 0.089)	
A4	(0.399, 0.498, 0.103)	(0.343, 0.577, 0.08)	

Finally, the evaluation values for all the shortlisted potential healthcare suppliers are assessed. It can be observed from Table X and Table XI, A₃ is the best, followed by A₄.

TABLE X. RESULTS OF WSM, WPM, SCORE FUNCTION VALUES

Healthcare Supplier	S i	s i	P i	p i
A1	(0.928, 0.026, 0.046)	0.951	(0.008, 0.973, 0.019)	0.018
A2	(0.918, 0.03, 0.052)	0.944	(0.005, 0.982, 0.013)	0.012
A3	(0.98, 0.007, 0.014)	0.986	(0.035, 0.92, 0.045)	0.057
A4	(0.96, 0.013, 0.027)	0.974	(0.02, 0.946, 0.035)	0.037

TABLE XI. SUPPLIERS SCORE BASED ON THE MCDM METHOD

Healthcare Supplier	k ia	k ib	k ic	k i	Rank
A1	0.244	2.507	0.929	2.055	3
A2	0.24	2	0.917	1.813	4
A3	0.262	5.794	1	3.501	1
A4	0.254	4.115	0.969	2.784	2

Simulation-Optimization result

This approach (Simulation-Optimization) is employed to solve the final decision-making result for the selection of healthcare suppliers. As previously indicated, in order to get the best solution approach, a metaheuristic – Genetic Algorithm will be combined with Simulation. The data employed in this section are based on Saputro et al. (2020) [12].

By integrating the suppliers' score from the first phase, and running GA with the figure as below, we can get the Z_{1max} value is 55,127.673 and supplier A₃ and A₄ were selected. While the best value for Z_{2min} is 135,618.986, selected suppliers include A₁ and A₃.

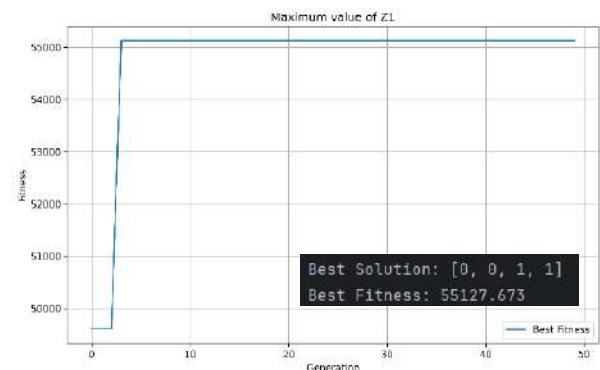


Fig. 1: The best value of Z1 after running GA

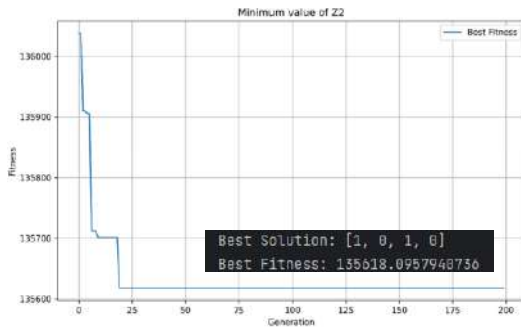


Fig. 2. The best value of Z2 after running GA

After obtaining the best values of Z_1 and Z_2 , the final solution will be decided by minimizing total deviation of both objectives with $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$. The selected suppliers include A_1 and A_3 with the best solution of e is 0.050094

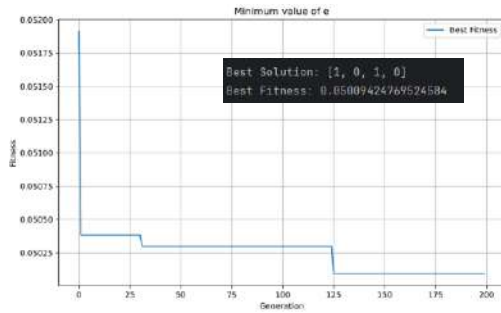


Fig. 3. The best value of e after running GA

IV. ANALYSIS

This section is intended to analyze objective weights' impact on the final decision-making as well as disruptions' impact on the value of total cost Z_2 . In order to gain the general conclusions, an additional situation, slow-moving items, is employed. This case is created for items that have expensive value and a low turnover

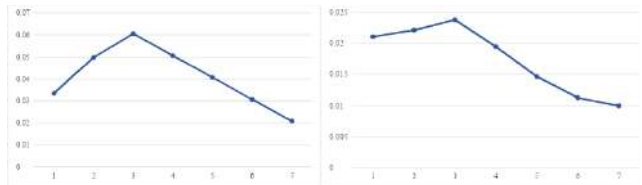


Fig. 4. Objective weight's impact on the deviation e for fast-moving items and slow-moving items

For the analysis on the impact of objective weight, weight α will be adjusted varies between 0.2 and 0.8 for each pair of α_1 and α_2 . As can be seen from the two figures for both scenarios, the more the weight α are unbalanced, the more the trade-off between two objectives will decrease and vice versa. Therefore, achieving high productivity without sacrificing other objectives is difficult. Furthermore, when Z_2 is prioritized, the optimal trade-off achieved, meaning that the deviations from other scenarios are lower. In other words, compared to maximizing total purchasing value (Z_1), minimizing total costs may be more necessary.

V. CONCLUSION

In this study, an uncertain group decision-making framework and risk mitigation under uncertainty and disruptions are used to handle the supplier selection challenge.

The problem is analyzed in depth by solving the potential risk in the financial aspects of all supplier candidates and then combining the scoring supplier process with a simulation of disruptions and inventory management to gain the final decision-making on the selected suppliers.

The multi-objective model is solved through a two-phase approach: supplier shortlisting and evaluation, followed by simulation-optimization with Analytic Model Enhancement (AME). This refinement incorporates disruption information into lead time, improving supplier selection and inventory decisions while reducing disruption risks. Furthermore, in the second phase, a meta-heuristic — Genetic Algorithm (GA) was used to combine the simulation model from the SimPy environment to search in the solution space and find the final optimal solution with the most suitable suppliers.

Although this study considers the intuitionistic fuzzy during the assessment of decision-makers, it does not consider the sustainable criteria (an essential element in many firms nowadays) and other sub-criteria to choose the most flexible supplier. Additionally, other risk considerations in global sourcing should be taken into account even while taking delivery delays or imperfect quality into account. Furthermore, in this study, all suppliers were assumed to be faced with disruptions, but in real life, plants can be flexible to solve the unpredicted dis-ruptions such as backup suppliers with higher unit costs in a certain period, ex-panding suppliers' capacity, or vehicle capacity.

REFERENCES

- [1] M. Govindan, K. Gengatharen, and N. J. Kannan, "Sustainable healthcare supply chain management: A review and research agenda," *Journal of Cleaner Production*, vol. 219, pp. 204–215, 2019.
- [2] D. Ivanov, "Viability of intertwined supply networks: Extending the supply chain resilience angles towards survivability. A position paper motivated by COVID-19 outbreak," *International Journal of Production Research*, vol. 58, no. 10, pp. 2904–2915, 2020.
- [3] D. Ivanov and A. Dolgui, "A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0," *Transportation Research Part E: Logistics and Transportation Review*, vol. 136, p. 101933, 2020.
- [4] A. Hamdan and A. Cheaitou, "Supplier selection and order allocation with green criteria: An MCDM and multi-objective optimization approach," *Computers & Operations Research*, vol. 81, pp. 282–290, 2017.
- [5] N. Kaur and R. Singh, "Multi-criteria decision making to evaluate healthcare supplier selection under uncertainty," *European Journal of Operational Research*, vol. 293, no. 1, pp. 17–27, 2021.
- [6] K. Kilic and N. Yalcin, "Multi-objective decision-making for healthcare supplier evaluation," *Applied Soft Computing*, vol. 95, p. 106524, 2020.
- [7] K. Atanassov, *Intuitionistic Fuzzy Sets: Theory and Applications*, Heidelberg: Springer, 1999.
- [8] N. Sawik, "On the robustness of optimal healthcare supply networks under supply disruptions," *Omega*, vol. 43, no. 1, pp. 83–95, 2014.
- [9] Saputro, T. E., Figueira, G., & Almada-Lobo, B. (2023). Hybrid MCDM and simulation-optimization for strategic supplier selection. *Expert Systems with Applications*, 219, 119624.
- [10] Chakraborty, S., Raut, R. D., Rofin, T. M., & Chakraborty, S. (2024). On solving a healthcare supplier selection problem using MCDM methods in intuitionistic fuzzy environment. *Opsearch*, 1-29.
- [11] Saputro, T. E., Figueira, G., & Almada-Lobo, B. (2019). Integration of supplier selection and inventory management under supply disruptions. *IFAC-PapersOnLine*, 52(13), 2827-2832.
- [12] Saputro, T. E., Figueira, G., & Almada-Lobo, B. (2021). Integrating supplier selection with inventory management under supply disruptions. *International Journal of Production Research*, 59(11), 3304-3322.

Scalable Federated Learning for Video Anomaly Detection in Distributed Environments

Nhat-Minh Dang

Faculty of Information Technology
Hung Yen University of Technology and Education
Hungyen, Vietnam
nhatminh@utehy.edu.vn

Quoc-Viet Hoang

Faculty of Information Technology
Hung Yen University of Technology and Education
Hungyen, Vietnam
viethqict@utehy.edu.vn

Thi-Thu-Trang Do

Faculty of Information Technology
Hung Yen University of Technology and Education
Hungyen, Vietnam
trangdtt@utehy.edu.vn

Van-Quyet Nguyen*

Faculty of Information Technology
Hung Yen University of Technology and Education
Hungyen, Vietnam
quyetict@utehy.edu.vn

Abstract— Video anomaly detection is crucial for large-scale surveillance safety, but centralized approaches face challenges with data volume, transmission overhead, and training delays. This paper introduces a scalable federated learning approach leveraging Apache Spark to enable efficient model training on large distributed datasets. A lightweight 3D CNN-LSTM model is trained in parallel on multiple clients, and global parameters are aggregated using FedAvg. We conduct experiments on a non-IID indoor activity dataset under different client settings, ranging from 2 to 10, to identify the optimal allocation that maximizes accuracy while minimizing training time. Experimental results show that 6 clients provide the best trade-off, achieving 0.7877 accuracy and 0.7554 F1-score while reducing training time by over 30% compared to 2 clients. When the number of clients exceeds this threshold, the dataset becomes excessively fragmented, diminishing the quality of local training and adversely affecting the accuracy of the aggregated global model. These findings highlight the potential of federated learning, when combined with distributed computing, to enable scalable and time-efficient video anomaly detection in large distributed environments.

Keywords—Federated Learning, Video Anomaly Detection, Smart Healthcare, Big Data Analytics

I. INTRODUCTION

Video Anomaly Detection (VAD) plays a crucial role in various surveillance applications such as public safety and healthcare. In hospital environments, VAD assists in detecting incidents such as patient falls, dangerous behaviors, or unauthorized access, thereby enhancing safety and enabling timely responses from medical staff. However, practical deployment faces significant computational challenges, as it requires processing continuous video streams from hundreds of cameras with high storage demands and near real-time processing requirements.

Several traditional deep learning approaches rely on centralized training, where video data from multiple cameras is uploaded to a central server to train a unified neural network model [1]. Centralized models based on CNNs, RNNs, transformers, or memory-augmented networks have achieved high accuracy on public datasets [2][3]. Other approaches utilize Graph Convolutional Networks (GCNs) by extracting skeletal information from videos and constructing spatiotemporal graphs to detect abnormal motions [4]. However, these methods typically assume centralized data availability, which becomes infeasible in large-scale distributed systems due to high storage, bandwidth costs, and expensive execution time issues.

To overcome these limitations, Federated Learning (FL) has emerged as a promising alternative. FL allows local training on edge devices and only transmits model weights to a central server for aggregation [5]. FL has been successfully applied in speech recognition, medical imaging, and video anomaly detection. Some studies have proposed unsupervised FL frameworks enabling cameras to collaborate without sharing raw data [6], or integrated FL with semantic knowledge transfer to achieve performance close to that of centralized models [7]. Nevertheless, these works have largely been limited to small-scale setups and lack comprehensive evaluation in large distributed systems.

Several challenges remain when scaling FL. First, scalability: a hospital network may comprise hundreds of continuously operating cameras, making video transmission to a central server impractical. Even FL is subject to performance degradation as the number of devices increases. Second, there is a need for rapid model updates to adapt to new data. Third, data scarcity and non-IID distributions across regions make it difficult to build a generalized model. Therefore, there is a need for an anomaly detection system that is scalable, capable of fast training, and can effectively leverage distributed and heterogeneous data.

This study investigates federated learning for video anomaly detection in distributed, non-IID environments. The system is implemented on Apache Spark cluster as a simulator to evaluate the impact of varying client allocations on model accuracy, execution efficiency, and scalability. FedAvg is employed for its simplicity, ease of deployment, and minimal hyperparameter requirements, reducing experimental complexity. This choice enables a focused analysis of key factors influencing system performance, such as client allocation, computational cost, and scalability, and provides a foundation for future evaluations of more advanced federated learning algorithms in complex real-world scenarios.

The main contributions of this paper are as follows:

- Implementing the FedAvg algorithm in a distributed environment using Apache Spark, aiming toward real-world deployment in hospital surveillance systems.
- Designing and optimizing the internal layers of a 3D CNN-LSTM model for local training on resource-constrained devices, specifically tailored for video data in abnormal behavior detection.
- Analyzing the impact of the number of clients on model performance, accuracy, and processing time,

and proposing optimal configurations to balance computational cost and model efficiency in heterogeneous federated learning settings.

II. RELATED WORK

Most traditional VAD methods rely on centralized training, where all surveillance video data from edge devices is transmitted to a central server for processing and model training. However, in large-scale surveillance systems, such as hospitals with hundreds of continuously operating cameras, this approach becomes impractical due to high transmission costs, significant storage requirements, and latency in processing [8]. Such systems demand fast, scalable, and distributed training capabilities to support near real-time inference.

FL has emerged as an effective solution for machine learning in distributed environments. FL enables edge devices (e.g., cameras or sensor nodes) to train models locally on their private data and only transmit model parameters to a central server for aggregation [9]. This approach not only reduces the bandwidth required for data transmission but also leverages parallelism in the training process, thereby accelerating convergence time.

Recent studies have demonstrated the feasibility of applying FL to VAD tasks. For instance, Doshi et al. [10] proposed FLVAD, a Transformer based architecture integrated with FL to detect anomalies without centralized data access. Al-Lahham et al. [11] introduced CLAP, a fully unsupervised FL framework that utilizes local feedback to train models collaboratively across multiple cameras. In the healthcare domain, Haripriya et al. [12] showed that combining FL with transfer learning can maintain high accuracy while optimizing computational costs and reducing training time.

Nevertheless, scaling FL to real-world distributed systems presents several challenges:

- **Scalability and training efficiency:** As the number of clients increases, model aggregation time and communication delays grow, leading to slower training processes [13].
- **Limited computational resources:** Edge devices such as surveillance cameras often lack the capability to handle complex models, necessitating lightweight architectures and efficient local training [14].
- **Non-IID data distributions:** Variability in data across devices hampers the convergence of a robust global model and requires adaptive aggregation and coordination strategies [9][13].

Therefore, integrating Federated Learning with a distributed computing platform such as Apache Spark is a promising direction toward building scalable and efficient VAD systems. Apache Spark provides robust resource management and task distribution across multiple nodes, enabling large-scale FL deployment while maintaining flexibility and high performance.

III. METHODOLOGY

This section presents the proposed methodology of the study. We first describe the overall architecture of the federated learning system. We then outline the design of the local learning model. Finally, we introduce the

implementation of the federated training process using Apache Spark.

A. Overall System Architecture

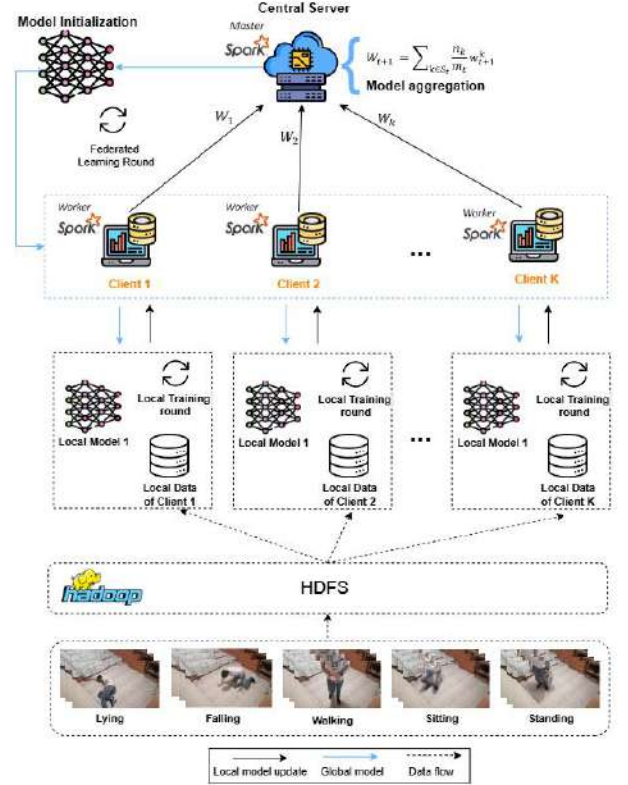


Fig. 1. System Overview

The system is designed to simulate a large-scale distributed deep learning architecture, in which hospital surveillance video data is processed locally at computing nodes following the principles of Federated Learning (FL). The overall system architecture is illustrated in **Fig. 1**.

The first component of the system is the Hadoop Distributed File System (HDFS), which stores preprocessed video segments in the form of short clips. Each clip represents a specific activity, such as falling, lying, walking, sitting, or standing. The data in HDFS is partitioned and evenly distributed across Spark executors to enable parallel processing and ensure load balancing.

Acting as federated clients, the Spark executors perform deep learning model training on their respective data partitions. Each executor may represent a group of cameras, a surveillance zone, or a functional unit within the hospital. The deep learning model deployed at each executor is a lightweight 3D CNN-LSTM architecture, which efficiently captures spatiotemporal features from sequences of video frames.

The Spark driver functions as the central aggregator, responsible for collecting model parameters from the clients using the Federated Averaging (FedAvg) algorithm. After each training round, the global model is updated and broadcast back to the executors for the next federated round.

The proposed architecture enables efficient parallel processing by leveraging multi-core computation at each executor, while preserving data privacy by eliminating the

need to transmit raw video data to a central server. The integration of Apache Spark simplifies scheduling, resource management, and system scalability, supporting both horizontal scaling (increasing the number of executors) and vertical scaling (increasing the training data volume). This makes the system well-suited for large-scale distributed deep learning applications on video data.

B. Design of 3D CNN–LSTM Model Architecture for Video Anomaly Detection

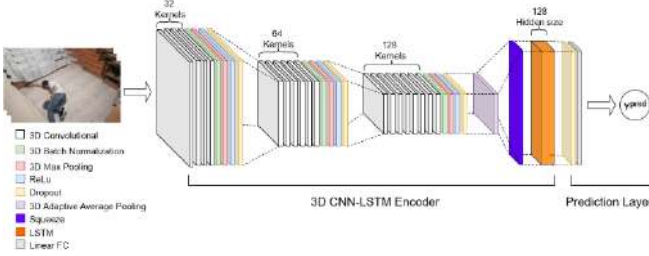


Fig. 2. Proposed 3D CNN-LSTM Model Architecture for Video Anomaly Detection

We propose a model that combines a three-dimensional convolutional neural network (3D CNN) with a Long Short-Term Memory (LSTM) network to jointly capture spatial and temporal features from video sequences for abnormal behavior detection. The model consists of two main components: a 3D CNN–LSTM encoder and a prediction layer. The input is a sequence of video frames with shape (B, T, C, H, W) , where B is the batch size, T is the temporal length, C is the number of channels, and $H \times W$ is the frame resolution. The input passes through three 3D convolutional blocks with increasing kernel sizes of 32, 64, and 128. Each block includes a 3D convolutional layer with a kernel size of $(3 \times 3 \times 3)$ and padding = 1, which helps preserve the spatial-temporal dimensions. This is followed by a ReLU activation, batch normalization, 3D max pooling (with temporal dimension preserved), and dropout to prevent overfitting.

After feature extraction, the model applies an adaptive average pooling layer to reduce the spatial dimensions to (1×1) while fully preserving the entire temporal dimension T . The resulting feature map is then reshaped to a temporal sequence of shape $(B, T, 128)$ and passed through an LSTM layer with a hidden size of 128. Only the output at the final time step is used as a compact representation of the entire video, which is then passed through a dropout layer and a fully connected linear layer for classification. This architecture is designed to be efficient for local training on resource-constrained devices while maintaining the ability to learn rich spatiotemporal representations from video data.

This design was selected because anomaly detection in video requires capturing both spatial appearance and temporal motion patterns. The 3D CNN–LSTM combination provides a balanced trade-off between expressive capacity and computational feasibility, making it suitable for our controlled simulation of federated learning.

C. Federated Learning Implementation using Spark

To implement the FL process training in a distributed environment, we utilize Apache Spark as a parallel processing platform to distribute data and coordinate the training process across multiple computing nodes. In the Spark architecture,

the driver serves as the central coordinator, responsible for task distribution, training management, and aggregation of model parameters after each federated round. The executors function as clients, with each executor performing local model training on its assigned data partition.

The dataset is divided into partitions corresponding to the number of simulated clients. Each executor trains the model on its local data and sends the updated parameters to the driver. The driver aggregates these parameters using the FedAvg algorithm to produce a global model, which is then broadcast back to the executors for the next training round. This process is repeated until convergence or a predefined number of rounds is reached, effectively and flexibly simulating federated learning on a Spark cluster.

IV. EXPERIMENTS

In this section, we first provides an overview of the video dataset used for the anomaly detection task. We then describes the experimental setup, training parameters, and evaluation metrics used to assess the model’s performance in the distributed environment.

A. Dataset

In this study, we use the Indoor Action Dataset released by DaniDeniz et al. [15]. The dataset comprises short video clips capturing daily indoor activities in environments such as kitchens, bedrooms, and living rooms, with a total of 1,228 videos recorded from five different subjects. Although the original dataset contains multiple action labels, we selected only five commonly observed behaviors in patients (see Table I) for model training, aligning with the context of behavior monitoring in hospital settings.

TABLE I. VIDEO DATASET SPLIT

Action Name	Train	Validation	Test	Sum
Falling down	34	11	16	61
Lying on the floor	53	12	36	101
Sitting down	66	22	29	117
Standing up	95	30	41	166
Walking	158	34	90	282
Total	406	109	212	727

The videos are preprocessed into sequences of frames with a spatial resolution of 64×64 pixels, ensuring temporal continuity across frames. To address class imbalance and enhance generalization capability, we apply mild data augmentation techniques such as horizontal flipping, color jittering, and geometric transformations. These augmentations increase the diversity of training samples while preserving the behavioral motion characteristics.

B. Experimental Settings

1) Experimental Environment: The experiments were conducted on a Spark cluster consisting of one coordinator node (master) and two compute nodes (workers), each equipped with 6 CPU cores and 8GB of RAM, and running the Ubuntu 20.04 operating system. Apache Spark was configured to manage distributed processing, and the input video data was partitioned and stored using the HDFS, distributed to the corresponding clients (executors acting as clients in the federated learning framework). We run experiments on the datasets with sequence lengths of 20, 15, and 10 frames, respectively. The number of participating

clients was varied (2, 4, 6, 8, and 10), with each client mapped to a dedicated Spark executor to guarantee isolated computation resources. This design enabled a controlled evaluation of how increasing the client scale affects both model performance and system efficiency.

2) Training Configuration: Each experiment consisted of 50 federated rounds. In each round, all participating clients trained their local models for 2 epochs, using a batch size of 4 and a learning rate of $1e-3$, optimized with the Adam optimizer. To address class imbalance in the dataset, we employed the Focal Loss function with a focusing parameter $\gamma = 1.0$, and incorporated class weights computed based on label frequency (described in detail in Class Weighting Strategy), enabling the model to better focus on underrepresented classes during training.

3) Class Weighting Strategy: To address the class imbalance problem in the training dataset, we compute class weights based on the frequency of labels in the dataset. Specifically, the raw weight w_i for class i is calculated as:

$$w_i = \frac{N}{C \times n_i} \quad (1)$$

where N is the total number of samples, C is the total number of classes, and n_i is the number of samples belonging to class i . This formula aims to assign higher weights to minority classes and lower weights to majority classes.

To prevent instability caused by excessively large weights during training, these raw weights are softened using the natural logarithm function, resulting in the final weight w'_i calculated as:

$$w'_i = \log(1 + w_i) \quad (2)$$

The softened weights w'_i are then converted to tensors and used in the Focal Loss, helping the model focus better on underrepresented classes while maintaining stability during training.

4) Evaluation Metrics: We use Accuracy, Precision, Recall, and F1-Score to measure the effectiveness of abnormal behavior detection. Simultaneously, time-related metrics such as total training time, average training time per round, model update time, and validation time are recorded to assess the efficiency and scalability of the distributed system.

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results to evaluate the model performance and the execution performance in a distributed federated learning environment.

A. Model Performance Evaluation

Table 2 provides important insights into the influence of client quantity and input frame size on model performance in federated learning systems with non-IID data distribution. The results show that increasing the number of clients from 2 to 6 leads to improved performance, as reflected in rising Accuracy and F1-score values. For example, with 10-frame input sequences, Accuracy increases from 0.7170 (2 clients) to 0.7217 (4 clients), peaking at 0.7877 (6 clients), while F1-score improves from 0.6591 to 0.6599 and then increases to 0.7554. However, when the number of clients increases to 8 and 10, performance begins to decline, with Accuracy dropping to 0.7689 and 0.6085, and F1-score decreasing to 0.7285 and 0.5135, respectively. This pattern suggests that a

moderate number of clients can enhance model generalization by improving data diversity and leveraging parallel training. Conversely, an excessive number of clients may fragment the local datasets too severely, reduce the quality of local model updates, and increase communication overhead, all of which negatively affect the global model. These findings highlight the importance of carefully selecting the number of clients to balance computational efficiency and model accuracy in federated learning systems, particularly in real-world applications such as hospital networks where data is inherently distributed.

TABLE II. COMPARISON OF FL PERFORMANCE WITH DIFFERENT CLIENTS AND DATA FRAME TYPES

#Client	Type Frame	Precision	Recall	F1-score	Accuracy
2	10	0.6884	0.6777	0.6591	0.7170
	15	0.7228	0.7779	0.7376	0.7594
	20	0.7170	0.6767	0.6817	0.7170
4	10	0.7035	0.6863	0.6599	0.7217
	15	0.7135	0.7458	0.7224	0.7500
	20	0.7315	0.7791	0.7482	0.7783
6	10	0.7396	0.7789	0.7554	0.7877
	15	0.7324	0.7497	0.7385	0.7830
	20	0.7780	0.7372	0.7171	0.7830
8	10	0.7290	0.7526	0.7285	0.7689
	15	0.6762	0.6984	0.6742	0.7217
	20	0.7295	0.7372	0.7197	0.7406
10	10	0.5249	0.5383	0.5135	0.6085
	15	0.6974	0.6765	0.6576	0.7264
	20	0.5249	0.5383	0.5135	0.6085

However, when the number of clients exceeds the optimal level (above 6 clients), performance degrades because the data becomes overly fragmented. Each client holds too small a local dataset, weakening the quality of local training and subsequently diminishing the effectiveness of global aggregation via FedAvg. In distributed and heterogeneous data settings, choosing an appropriate number of clients is critical to balancing the benefits of leveraging diverse data and maintaining the stability and efficiency of the federated learning process. The results with short sequence lengths (10 frames) further highlight the importance of preserving sufficient spatiotemporal information for the model to learn and predict accurately in such distributed environments.

B. Execution Performance Evaluation

To evaluate the scalability and time efficiency of the Federated Learning system in a distributed environment, we analyzed the training time, validation time, model update time, and average training time per round across different clients and frame types.

Fig. 3 illustrates the total training time of the FedAvg system under varying client counts and input frame lengths. Increasing the number of clients from 2 to 4 significantly reduces total training time, for example, from 3540.27s to 2436.43s in the 10-frame configuration, due to improved parallelism and more efficient workload distribution. This trend is also observed for 15-frame (5364.33s to 3892.08s) and 20-frame (6790.20s to 4474.50s) inputs.

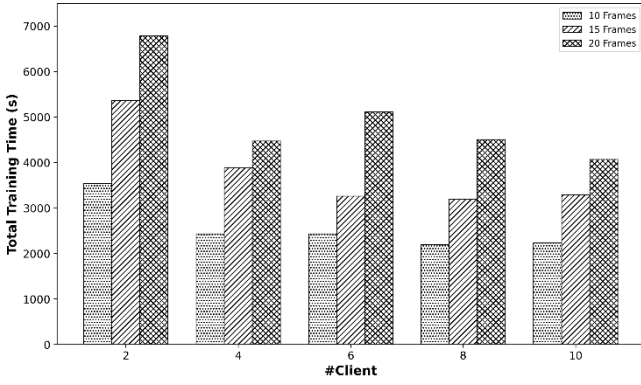


Fig. 3. Comparison of Total Training Time across Different Clients and Frame Types

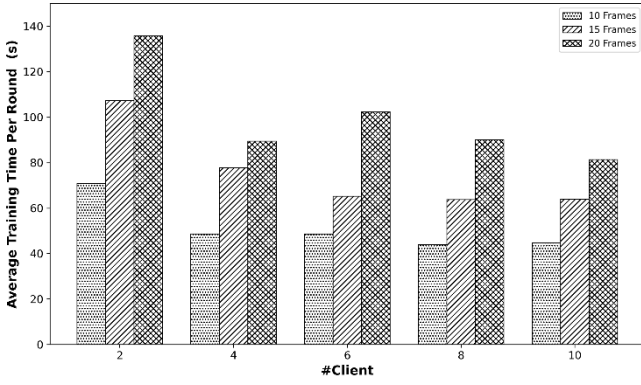


Fig. 4. Comparison of Average Training Time per Round across Different Clients and Frame Types

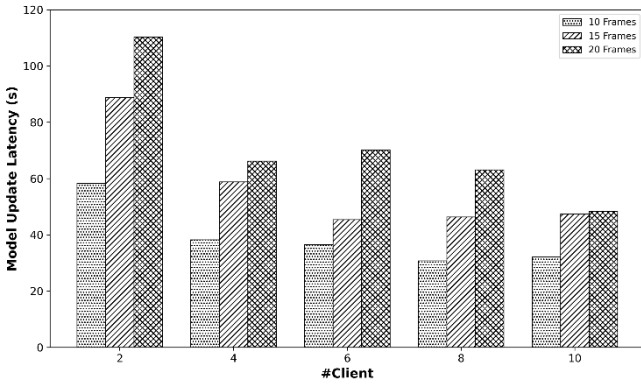


Fig. 5. Comparison of Model Update Time across Different Clients and Frame Types

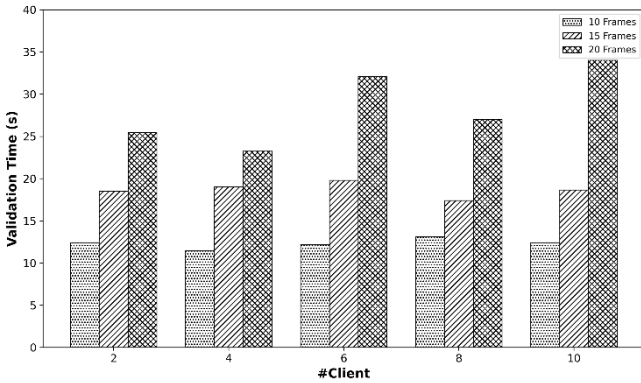


Fig. 6. Comparison of Validation Time across Different Clients and Frame Types

However, when the client count exceeds the optimal threshold of six, the total training time tends to stabilize or slightly increase. For instance, with 20 frames, training time increases from 4474.50s (4 clients) to 5117.84s (6 clients), and fluctuates at 4506.29s (8 clients) and 4069.68s (10 clients). Simultaneously, model accuracy declines, illustrating the trade-off between computational efficiency and model quality due to increased communication and synchronization overhead in large-scale systems.

Fig. 4 presents the average training time per federated round. As the number of clients increases from 2 to 10, round time generally decreases. In the 10-frame setup, it drops from 70.80s (2 clients) to 48.73s (4 clients) and 43.85s (8 clients). However, this metric rises significantly with longer input sequences: for 20-frame inputs, average round time reaches 135.80s (2 clients), 89.49s (4 clients), and 102.36s (6 clients). This reflects increased data complexity and computational cost. When combined with the accuracy trends, these findings highlight the need to carefully select input sequence length and client number to maintain a balance between model performance and computational efficiency.

Fig. 5 depicts the model update time between clients and the central aggregator. As the number of clients increases from 2 to 10, the update time consistently decreases in the 10-frame case from 58.36s (2 clients) to 30.76s (8 clients) and 32.21s (10 clients) due to more efficient synchronization. However, longer sequences still cause moderate increases in latency. For 20-frame inputs, update time peaks at 110.29s (2 clients), decreases to 66.24s (4 clients), but then climbs again to 70.23s (6 clients) and stabilizes around 63.09s (8 clients) and 47.36s (10 clients). This suggests that while scaling helps with synchronization, complex data imposes additional communication load that may impact update consistency and timeliness.

Finally, Fig. 6 shows the validation time after each training round. This metric is mainly influenced by input sequence length rather than client count. For instance, at 10 clients, validation time increases from 12.43s (10 frames) to 18.65s (15 frames) and 34.02s (20 frames). Similar patterns are seen across other client settings. These results indicate that although validation time remains relatively stable across client counts, using longer input sequences to improve accuracy leads to higher validation costs. This emphasizes the importance of balanced system design that considers both model accuracy and computational resource constraints.

Collectively, these findings demonstrate a trade-off between training parallelism and overhead. To ensure both system efficiency and model accuracy, it is essential to balance client allocation and sequence configuration according to available computational resources and task complexity. Such tuning becomes particularly critical in real-time or resource-constrained applications such as smart surveillance or hospital monitoring systems.

VI. CONCLUSION

This study investigated the integration of federated learning with Apache Spark to enable scalable video anomaly detection in distributed, non-IID environments. Experimental results demonstrate that proper client allocation significantly improves training efficiency while maintaining competitive model performance. Nevertheless, increasing the number of clients beyond an optimal threshold leads to excessive data

fragmentation, weakening local learning quality and degrading the aggregated model. These findings highlight the importance of balancing computational cost, scalability, and accuracy when designing federated learning systems for real-world surveillance scenarios. In the future, we intend to delve into more sophisticated federated learning strategies and adaptive training paradigms to achieve superior robustness, communication efficiency, and generalization capabilities in highly dynamic and heterogeneous distributed settings. We will also evaluate the approach on diverse datasets to validate scalability and robustness across different environments. Additionally, we plan to investigate personalized federated learning methods, optimize communication efficiency, and integrate differential privacy to ensure data confidentiality in real-world deployments.

REFERENCES

- [1] G. A. Noghre, "Privacy-preserving Real-world Video Anomaly Detection," 2023 IEEE International Conference on Smart Computing (SMARTCOMP), Nashville, TN, USA, 2023, pp. 235-254, doi: 10.1109/SMARTCOMP58114.2023.00067.
- [2] C. Huang, C. Liu, J. Wen, L. Wu, Y. Xu, Q. Jiang, and Y. Wang, "Weakly supervised video anomaly detection via self-guided temporal discriminative transformer," IEEE Transactions on Cybernetics, vol. 54, no. 5, pp. 3197–3210, 2022.
- [3] Muniyua, John & Wambugu, Geoffrey & Njenga, Stephen. (2021). A Survey of Deep Learning Solutions for Anomaly Detection in Surveillance Videos. International Journal of Computer and Information Technology(2279-0764). 10. 10.24203/ijcit.v10i5.166
- [4] Xu, Y.; Huang, B.; Zhou, C.; Wang, H.; Li, X. Video Anomaly Detection with Hyperbolic Graph Embedding and Masked Normalizing Flows. Electronics 2024, 13, 5013. <https://doi.org/10.3390/electronics13245013>
- [5] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2), 1-210.
- [6] A. Al-lahham, M. Z. Zaheer, N. Tastan and K. Nandakumar, "Collaborative Learning of Anomalies with Privacy (CLAP) for Unsupervised Video Anomaly Detection: A New Baseline," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 12416-12425, doi: 10.1109/CVPR52733.2024.01180. keywords: {Training;Privacy;Data privacy;Protocols;Federated learning;Surveillance;Collaboration},
- [7] F. Qi, R. Pan, H. Zhang, and C. Xu, "FedVAD: Enhancing Federated Video Anomaly Detection with GPT-Driven Semantic Distillation," in Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol. 14389, Cham: Springer, 2024, pp. 234–251. doi: 10.1007/978-3-031-73668-1_14.
- [8] Liu, Jing & Liu, Yang & Lin, Jieyu & Li, Jielin & Cao, Liang & Sun, Peng & Hu, Bo & Song, Liang & Boukerche, Azzedine & Leung, Victor. (2025). Networking Systems for Video Anomaly Detection: A Tutorial and Survey. ACM Computing Surveys. 57. 10.1145/3729222.
- [9] Lu, Zili & Pan, Heng & Dai, Yueyue & Si, Xueming & Zhang, Yan. (2024). Federated Learning With Non-IID Data: A Survey. IEEE Internet of Things Journal. PP. 1-1. 10.1109/JIOT.2024.3376548.
- [10] K. Doshi and Y. Yilmaz, "Privacy-Preserving Video Understanding via Transformer-based Federated Learning," 2023 IEEE Conference on Dependable and Secure Computing (DSC), Tampa, FL, USA, 2023, pp. 1-8, doi: 10.1109/DSC61021.2023.10354099.
- [11] A. Al-lahham, M. Z. Zaheer, N. Tastan and K. Nandakumar, "Collaborative Learning of Anomalies with Privacy (CLAP) for Unsupervised Video Anomaly Detection: A New Baseline," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 12416-12425, doi: 10.1109/CVPR52733.2024.01180.
- [12] Haripriya, R., Khare, N. & Pandey, M. Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings. Sci Rep 15, 12482 (2025). <https://doi.org/10.1038/s41598-025-97565-4>
- [13] J. Wang, S. Guo, X. Xie and H. Qi, "Protect Privacy from Gradient Leakage Attack in Federated Learning," IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, 2022, pp. 580-589, doi: 10.1109/INFOCOM48880.2022.9796841.
- [14] Razaque Mughal, Fahad & He, Jingsha & Das, Bhagwan & Dharejo, Fayaz & Zhu, Nafei & Bhatia, Surbhi & Alzahrani, Saeed. (2024). Adaptive federated learning for resource-constrained IoT devices through edge intelligence and multi-edge clustering. Scientific Reports. 14. 10.1038/s41598-024-78239-z.
- [15] D. Deniz, E. Ros, E. M. Ortigosa, F. Barranco. "Optimized edge-cloud system for activity monitoring using knowledge distillation" in Electronics, 13 (23), 2024.

FedSAE-HS: Hybrid Anomaly Scoring for Semi-Supervised Federated Intrusion Detection

1st Linh Chu Thi
Le Quy Don Technical University
Ha Noi, Viet Nam
chuthilinh189@gmail.com

3rd Trung Ngo Van
Joint Stock Commercial Bank for Foreign Trade of Viet nam
Ha Noi, Viet Nam
ngovantrung.mta@gmail.com

2nd Van Thi Bich Pham
Le Quy Don Technical University
Ha Noi, Viet Nam
vanptb@lqdtu.edu.vn

4th Ly Vu*
Le Quy Don Technical University
Ha Noi, Viet Nam
vu.ly@lqdtu.edu.vn

Abstract—Federated learning (FL) has emerged as a promising paradigm for collaborative network intrusion/anomaly detection in distributed systems, particularly under privacy constraints and data decentralization. Recent semi-supervised FL approaches with autoencoders (AEs) have shown strong potential in this setting. Nevertheless, many of these methods rely on only a single scoring signal, such as reconstruction error (RE) or latent representation z , which often results in unstable detection. To address this, we propose an enhanced federated anomaly detection framework named as Federated Learning with Hybrid Scoring based Shrink AutoEncoder (FedSAE-HS) that applies both RE and latent vector z for calculating an anomaly score. The hybrid scoring strategy allows the model to be sensitive to subtle anomalies, improving detection stability and generalization. We evaluate FedSAE-HS on two benchmark IDS datasets, N-BaIoT and WSN_DS, against four existing federated anomaly detection frameworks. Experimental results show that FedSAE-HS achieves superior performance in AUC, Precision, Recall, F1-score, and False Positive Rate, while converging more stably, confirming the effectiveness of the proposed hybrid scoring strategy for semi-supervised federated anomaly detection.

Index Terms—Federated Learning, Intrusion Detection System, Autoencoder

I. INTRODUCTION

The rapid adoption of Internet of Things (IoT) technologies has led to increasingly diverse and distributed networks that are frequently targeted by cyberattacks due to limited resources and weak security configurations. Intrusion detection systems (IDS) play a critical role in analysing network traffic to detect unauthorised activities [1]. Traditional IDS are based on centralized training and require data collection from all end devices, which presents challenges in privacy, security risks, scalability, and communication [1], [2].

Federated learning (FL) addresses these issues by enabling collaborative model training while keeping data local [2], [3]. A wide range of semi-supervised FL-based IDS methods have emerged, particularly those using AutoEncoder (AE) architectures [4]–[8]. In such models, clients learn benign patterns and jointly detect anomalies without explicit attack labels. These techniques often detect anomalies based on a

single informative signal derived from normal traffic, most commonly reconstruction error (RE) or latent vector z . For example, FedDetect [6] and FedDAE [9] use only RE, while FedSAE relies solely on z [7]. Similarly, FL-MAAE employs RE in a memory-augmented AE under FL settings [10], and the FL-VAE ConvGRU model assesses anomalies purely through latent representations [11]. However, systems that rely solely on RE for anomaly scoring can be highly sensitive to noise and outliers, leading to misclassification of rare but benign patterns [12], [13]. Conversely, models that depend only on latent vector z may fail to capture subtle anomalies that cause minimal shifts in the embedding space [14]. Such limitations can ultimately undermine the overall reliability of anomaly detection.

In this paper, we propose the Federated Learning with Hybrid Scoring based Shrink AutoEncoder (FedSAE-HS) method, an improved version over the original FedSAE framework. Although FedSAE and FedSAE-HS share a similar training loss, the key difference lies in how anomalies are scored during the testing phase. FedSAE-HS integrates both types of signal to produce a hybrid score, leading to improved stability and more discriminative performance, especially when either signal alone may be insufficient.

The main contributions of this paper are:

- We present FedSAE-HS, a federated anomaly detection framework that integrates both RE and the magnitude of the latent vector z in the anomaly scoring phase to enhance detection robustness.
- We provide empirical evidence about the performance of the proposed framework and the four existing federated anomaly detection frameworks in two benchmark IDS datasets N-BaIoT and WSN_DS.
- We empirically demonstrate the effectiveness of the hybrid anomaly scoring method in the FL paradigm by achieving higher ROC-AUC and superior average Precision, Recall, F1-score, and False Positive Rate (FPR), while also exhibiting faster convergence during training.

The rest of the paper is organized as follows. Section II provides research background and related work. Section III presents the proposed framework. The experiment settings are described in Section IV. Section V presents experiment results and discussions. Finally, the conclusion is drawn in Section VI.

II. BACKGROUND AND RELATED WORK

This section provides the foundational concepts, key techniques, and related work in FL and AE-based anomaly detection, with an emphasis on anomaly scoring strategies in semi-supervised federated settings.

A. Federated Learning for Intrusion Detection

FL is a decentralized training paradigm that enables multiple clients to collaboratively learn a shared global model without sharing their raw data [2], [3]. Each client performs local updates based on private data, and only model weights are exchanged with a central server. This setup is particularly beneficial for IDS in IoT environments, where data is inherently distributed, privacy-sensitive, and non-uniform across clients [6].

In a standard FL workflow, the server initializes the global model weight w^0 and sends it to a selected subset of clients. Each client i updates the local model by minimizing its local loss $\mathcal{L}_i$, and returns the updated weights w_i^t to the server. The server aggregates the updates by averaging the local model weights [15] as follows:

$$w^{t+1} = \sum_{i=1}^N \frac{n_i}{n} w_i^t, \quad (1)$$

where t denotes the communication round index, n_i is the number of samples in the client i , and $n = \sum_{i=1}^N n_i$. This process is repeated over multiple communication rounds.

Recent studies have applied FL to IDS tasks under various supervised learning levels. For example, FedFAME [16] and FedUAS [17] introduced strategies for label-scarce scenarios. METALS [18] applied semi-supervised FL with active learning for lightweight IoT systems, while HeteroAdaFedSemi [19] focused on handling heterogeneous computation budgets across clients. In particular, semi-supervised FL approaches that utilize only benign, i.e., normal, data for training have gained attraction due to only using only normal data which is easy to collect in the real world. These methods require no attack labels, making them especially suitable for privacy-sensitive IDS scenarios in which anomalies are rare or hard to label explicitly.

B. AutoEncoder-based Intrusion Detection System

AE models [9], [20], [21] are widely used neural network architectures for anomaly detection in IDS tasks [13], [22], [23]. An AE consists of two components, an encoder $f_\theta(\cdot)$, which maps the input data to a compressed latent vector z , and a decoder $g_\phi(\cdot)$, which reconstructs the original input from that representation [20]. In IDS, autoencoders (AEs) try to learn the behavior of normal data by reducing the RE that measures the difference between the input and output of the AE models. In

many AE-based approaches, intrusions/anomalies are detected when the RE exceeds a predefined or statistically-derived threshold.

Given an input sample $x \in \mathbb{R}^d$, the encoder transforms it into a latent vector as follows:

$$z = f_\theta(x), \quad (2)$$

where $z \in \mathbb{R}^k$ with $k < d$.

The output of the AE model $\hat{x}$ then reconstructs the input via the following equation:

$$\hat{x} = g_\phi(z) = g_\phi(f_\theta(x)). \quad (3)$$

The training objective is to minimize the RE, typically the Mean Squared Error (MSE) as in the following equation:

$$\mathcal{L}_{\text{rec}}(x, \hat{x}) = \|x - \hat{x}\|_2^2. \quad (4)$$

When trained on normal data, the AE aims to reconstruct such inputs with low error. The intrusions/anomalies, which are outside the learned distribution, yield higher RE [13], [24]. Based on this, the anomaly scores can be computed as follows:

$$\text{Score}_{RE}(x) = \|x - \hat{x}\|_2^2. \quad (5)$$

Alternatively, anomaly detection can also be performed based on the magnitude of the latent vector z itself:

$$\text{Score}_z(x) = \|f_\theta(x)\|^2. \quad (6)$$

Although AE-based IDS methods have demonstrated strong performance, their reliance on centralized data aggregation limits their applicability in realistic and privacy-sensitive environments. Semi-supervised FL approaches based on AEs have been introduced to mitigate data sharing concerns, such as FedDetect [6], FedSAE [7], and FedMSE [8].

Anomaly scoring is a critical component of semi-supervised intrusion detection. Existing methods generally fall into two categories. The first is RE-based scoring, which computes the anomaly score as the squared difference between the input x and its reconstructed version $\hat{x}$, i.e., $\text{Score}_{RE}(x) = \|x - \hat{x}\|_2^2$. This approach, used in methods like FedDetect [6] effectively detects input-level deviations but is highly sensitive to noise and outliers. The second is z -based scoring, which measures the deviation of a sample's latent representation from a learned reference such as a centroid vector $\bar{z}$; for instance, FedSAE [7] calculates $\text{Score}_z(x) = \|f_\theta(x) - \bar{z}\|^2$. While z -based scoring tends to be more stable and generalizes better, it may fail to capture subtle deviations essential for anomaly detection. In our paper, we propose a hybrid anomaly scoring strategy for IDS that combines both RE-based and z -based scoring to leverage their complementary strengths.

III. PROPOSED FEDERATED LEARNING METHOD

In this section, we propose FedSAE-HS, a federated IDS framework designed for semi-supervised learning. FedSAE-HS improves the FedSAE framework by integrating both the RE and latent vector z into the anomaly scoring mechanism. This section outlines the core design of FedSAE-HS, including its architecture, loss formulation, and federated training procedure.

A. System Overview

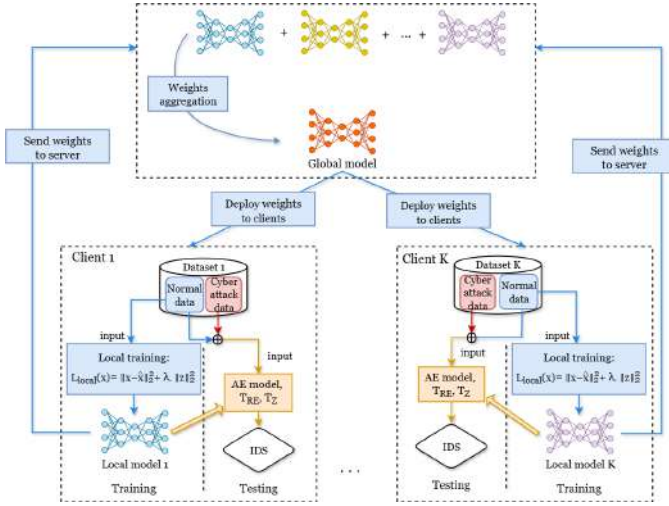


Fig. 1. Overall architecture of the FedSAE-HS framework.

Fig. 1 illustrates the overall architecture of the proposed FedSAE-HS framework for semi-supervised federated anomaly detection. The system consists of a central server and multiple clients, each holding its own dataset containing both normal data. During local training, each client trains an AE model using a joint loss that combines RE and latent representation regularization as presented in (8).

The locally trained weights are sent to the central server, which aggregates them to update the global model. The updated weights are then deployed back to all clients for the next training round. For testing, each client applies the trained AE model and uses two thresholds, one for RE (T_{RE}) and one for z -based scoring (T_z), to detect intrusions. This collaborative training process allows the detection of unseen attacks without sharing data, thereby preserving privacy while improving detection performance.

B. Local Autoencoder Architecture

Each client employs a standard AE composed of an encoder $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^k$ and a decoder $g_\phi : \mathbb{R}^k \rightarrow \mathbb{R}^d$, where d is the input feature dimension and $k < d$ is the size of the latent space. For an input $x \in \mathbb{R}^d$, the forward pass is:

$$z = f_\theta(x), \quad \hat{x} = g_\phi(z), \quad (7)$$

where z is the latent vector and $\hat{x}$ is the reconstructed input.

The local training objective consists of two components, i.e., the RE between the input x and its reconstruction $\hat{x}$, and a latent shrinkage term that penalizes the magnitude of the vector z . The overall loss is as follows:

$$\mathcal{L}_{\text{local}}(x) = \|x - \hat{x}\|_2^2 + \lambda \cdot \|z\|_2^2, \quad (8)$$

where λ is a regularization coefficient that encourages the encoder to produce compact vectors z close to the origin without collapsing the representation.

The federated training proceeds over multiple communication rounds as follows:

- 1) Initialization (first round only): The central server initializes the global model weights w^0 .
- 2) At the beginning of each round t , the server broadcasts the current global model weights w^t to clients.
- 3) Each client updates its local model weight by minimizing the loss $\mathcal{L}_{\text{local}}$ on its private normal data.
- 4) The clients send their updated local model weights w_i^{t+1} back to the server.
- 5) The server aggregates weights using the averaging operator [15] as the following equation:

$$w^{t+1} = \sum_{i=1}^N \frac{n_i}{n} w_i^t, \quad (9)$$

where n_i is the sample size on client i , and $n = \sum_i n_i$.

For each client i , we perform a validation pass on its own benign data to compute the RE threshold ($Score_{RE}$) and the z threshold ($Score_z$). Specifically, we calculate the scalar scores for each validation sample as:

$$Score_{RE}(x) = \|x - \hat{x}\|_2^2, \quad Score_z(x) = \|z\|_2^2,$$

We then compute the Gaussian mean and standard deviation of these thresholds in local training to obtain $(\mu_{RE,i}, \sigma_{RE,i})$ and $(\mu_{z,i}, \sigma_{z,i})$, where μ denotes the mean value and σ denotes the standard deviation of the corresponding score distribution over the validation samples of the client i .

Using a scaling factor γ^1 , the per-signal thresholds are defined as:

$$T_{RE,i} = \mu_{RE,i} + \gamma \cdot \sigma_{RE,i}, \quad T_{z,i} = \mu_{z,i} + \gamma \cdot \sigma_{z,i}. \quad (10)$$

During testing phase in the client i , for each testing input data x , we calculate $z = f_\theta(x)$ and $\hat{x} = g_\phi(z)$, evaluate $Score_{RE}(x)$ and $Score_z(x)$, and calculate the normalized anomaly score:

$$Score(x) = \frac{Score_{RE}(x)}{T_{RE,i}} + \frac{Score_z(x)}{T_{z,i}}. \quad (11)$$

We classify x as anomalous if $Score(x) > \tau$, with a default cut-off $\tau = 2.0$.

The two signals in Eq. (11) capture different but complementary aspects of the data. The reconstruction error reflects deviations in the input space and is effective for detecting large distortions, but it can also be sensitive to noise or outliers. On the other hand, the latent norm represents structural variations in the embedding space, which provides robustness to noisy inputs but may fail to expose subtle anomalies that barely alter the reconstruction.

By normalizing and combining both signals into a single hybrid score, our method leverages their complementary strengths, mitigating the weaknesses of each individual measure and yielding a more stable and generalizable anomaly detector.

¹In the experiments, we tune the parameter γ by running multiple trials with values $\gamma = 1, 2, 3, 4, 5$, and select the optimal value for each experiment (e.g., around 4 for FedSAE-HS).

IV. EXPERIMENTAL SETUP

This section describes the datasets, training configurations, evaluation metrics, and experimental scenarios used in our study.

A. Datasets

To evaluate the performance of FedSAE-HS, we use two public intrusion detection datasets: N-BaIoT [25] and WSN_DS [26]. These datasets represent different types of IoT environments and provide various attack scenarios. Table I shows the total number of samples in training, validation and test sets, using a split ratio of 70%-10%-20% with a fixed random seed for reproducibility.

The N-BaIoT dataset [25] contains data collected from IoT devices infected with Mirai and Bashlite malware. In this study, we used a balanced subset of 190,463 samples extracted from selected devices, including both benign and malicious traffic. Each data point is represented by 115 statistical features computed over various time windows. The dataset includes multiple types of botnet attacks, such as UDP, Scan, ACK, and SYN flooding.

The WSN_DS dataset [26] originates from a simulated wireless sensor network scenario including various attack types such as Blackhole, Grayhole, Flooding, and Scheduling. In our experiments, we use the public version containing 19 numerical attributes. We further preprocess the dataset by binarizing the categorical labels (Normal = 0, attack = 1) and removing the identifier and label columns, resulting in 17 numerical features used for training and testing.

TABLE I
SUMMARY STATISTICS OF THE EVALUATED DATASETS

Dataset	Train	Val	Test	Features
N-BaIoT	133,324	19,046	38,093	115
WSN_DS	262,263	37,466	74,932	19

B. Hyper-Parameter Settings

All experiments were conducted using 20 clients. The number of local training epochs was set to 1 per communication round. The batch size was fixed at 128 for both training and validation. The learning rate was chosen based on empirical grid search: 0.01 for WSN_DS and 0.001 for N-BaIoT, which aligns with the scale and complexity of the datasets.

The coefficient λ used in the proposed FedSAE [7] and FedSAE-HS method is set to 1. This value was empirically selected to balance the contributions of the RE and the latent vector z in the loss function. All experimental models were trained for 500 global communication rounds.

The aggregation strategy was set to a simple averaging [15] across all baselines. We used the same input dimension as the number of features after preprocessing, i.e., 17 for the WSN_DS and 115 for the N-BaIoT.

We evaluated performance using multiple metrics. In addition to ROC-AUC, which provides a measure widely used in anomaly detection, we also report Precision, Recall, F1-score,

and FPR [27] to better reflect classification quality and false alarm behavior in intrusion detection.

C. Experimental Scenarios

We conducted two experiments to evaluate the effectiveness of the proposed FedSAE-HS method in federated anomaly detection settings.

First, we compared the performance of FedSAE-HS with those of four baseline methods: FedDetect [6], FedDAE [9], FedSAE [7], and FedMSE [8]. For each dataset, we report per-client and average ROC-AUC to capture both local and global detection performance. In addition, we present the average Precision, Recall, F1-score, and FPR to provide a broader evaluation of classification quality.

Second, we analyzed the convergence behavior of FedSAE-HS by visualizing the training dynamics over communication rounds. Specifically, we track the reconstruction loss and the latent loss on a representative client. This provides insight into how each component of the objective evolves during training and contributes to overall optimization stability.

V. PERFORMANCE EVALUATION

A. Comparison with Semi-supervised learning methods

TABLE II
AUC (%) COMPARISON ON THE N-BaIoT DATASET

Client	FedDetect	FedDAE	FedSAE	FedMSE	FedSAE-HS
Client 1	95.62	87.12	97.37	97.38	98.38
Client 2	93.51	85.52	92.85	92.86	94.15
Client 3	83.24	80.38	99.05	99.06	99.19
Client 4	94.48	88.02	91.67	91.88	94.36
Client 5	90.56	73.55	90.02	91.03	91.31
Client 6	75.56	74.2	89.75	90.94	91.68
Client 7	94.12	77.91	95.47	95.85	95.49
Client 8	80.49	82.03	94.63	93.27	93.03
Client 9	81.49	78.61	99.03	98.54	99.01
Client 10	86.49	83.31	98.26	98.27	99.52
Client 11	96.92	92.19	96.87	96.70	98.21
Client 12	80.41	82.38	96.46	96.17	96.00
Client 13	72.10	75.37	92.67	92.19	87.60
Client 14	88.15	80.64	93.21	93.57	92.11
Client 15	79.23	88.00	95.01	96.02	96.29
Client 16	89.92	85.68	99.32	99.33	96.21
Client 17	84.81	81.76	97.42	98.01	99.51
Client 18	79.77	78.27	95.21	94.05	95.41
Client 19	78.65	77.02	94.15	95.16	96.23
Client 20	76.77	78.48	99.57	98.61	97.22
Average	85.12	81.52	95.40	95.44	95.53

Table II presents the AUC performance of FedSAE-HS in comparison with four semi-supervised federated baselines, namely FedDetect [6], FedDAE [9], FedSAE [7], and FedMSE [8], on the N-BaIoT dataset. On average, FedSAE-HS achieves the best AUC score of 95.53%, outperforming all other methods. While FedSAE (95.40%) and FedMSE (95.44%) show strong results, FedSAE-HS slightly improves over them, and significantly surpasses FedDetect (85.12%) and FedDAE (81.52%). FedSAE-HS demonstrates more stable performance across clients with the highest AUC score in 11 out of 20 clients. In the rest of the clients, FedSAE-HS maintains a notably higher AUC, suggesting improved

TABLE III
AUC (%) COMPARISON ON THE WSN_DS DATASET

Client	FedDetect	FedDAE	FedSAE	FedMSE	FedSAE-HS
Client 1	91.71	89.65	89.87	90.37	93.08
Client 2	93.40	91.81	89.92	90.42	93.17
Client 3	92.03	92.96	90.31	90.81	93.69
Client 4	93.55	93.23	89.58	90.08	93.32
Client 5	92.12	91.55	90.49	90.99	93.32
Client 6	92.29	94.88	90.43	90.93	94.00
Client 7	93.14	96.08	90.39	90.89	94.28
Client 8	92.56	90.92	89.60	90.10	93.17
Client 9	92.75	91.99	90.44	90.94	94.16
Client 10	91.62	92.03	90.13	90.63	93.18
Client 11	92.51	89.28	89.77	90.27	93.07
Client 12	91.19	88.85	89.40	89.90	92.29
Client 13	92.06	89.18	90.83	91.33	93.90
Client 14	93.02	92.43	90.12	90.62	94.34
Client 15	92.87	92.58	90.09	90.59	93.97
Client 16	92.75	91.39	89.81	90.31	93.47
Client 17	92.71	89.29	89.68	90.18	93.52
Client 18	91.95	89.28	90.08	90.58	93.48
Client 19	91.90	92.53	90.40	90.90	94.44
Client 20	91.88	92.45	90.68	91.18	94.08
Average	92.40	91.62	90.10	90.60	93.59

robustness under client heterogeneity. Compared to FedSAE, the additional use of the RE in FedSAE-HS leads to more discriminative anomaly representations, allowing the model to better identify outliers in different distributions.

Performance comparisons of the five methods on the WSN_DS dataset are shown in Table III. Following a similar trend, FedSAE-HS obtains the highest mean AUC of 93.59%, comparing to 90.10% (FedSAE), 90.60% (FedMSE), 91.62% (FedDAE), and 92.40% (FedDetect). FedSAE-HS also consistently ranks the best performance in most clients (16 of 20). The gains are more pronounced on clients that typically show high variability, indicating that the hybrid scoring method in FedSAE-HS helps mitigate inconsistency caused by local data differences.

TABLE IV
PERFORMANCE COMPARISON (PRECISION, RECALL, F1, AND FPR) ON THE N-BaIoT DATASET

Metric	FedDetect	FedDAE	FedSAE	FedMSE	FedSAE-HS
Precision	98.18	97.51	98.33	98.34	98.47
Recall	80.35	75.64	99.33	99.22	99.93
F1	87.83	84.90	98.81	98.77	99.19
FPR	8.77	11.47	10.05	9.97	9.29

TABLE V
PERFORMANCE COMPARISON (PRECISION, RECALL, F1, AND FPR) ON THE WSN_DS DATASET

Metric	FedDetect	FedDAE	FedSAE	FedMSE	FedSAE-HS
Precision	49.84	52.63	34.31	31.95	45.57
Recall	96.53	93.66	98.84	98.84	98.84
F1	65.64	66.59	50.86	48.20	62.20
FPR	10.20	9.36	19.53	21.79	12.33

Beyond ROC-AUC, we further evaluate Precision, Recall, F1-score, and FPR, which are crucial for imbalanced intrusion detection tasks. Results on N-BaIoT (Table IV) show that

FedSAE-HS achieves the best overall performance across all reported metrics, with the highest Precision (98.47%), Recall (99.93%), and F1-score (99.19%), while also maintaining the lowest FPR (9.29%).

On WSN_DS (Table V), FedSAE-HS again provides competitive results. Although its Precision (45.57%) is slightly lower than FedDAE (52.63%), it offers a more favorable trade-off with high Recall (98.84%) and a lower FPR (12.33%), yielding strong overall detection capability.

Overall, the results show that FedSAE-HS not only achieves the best ROC-AUC across most clients but also consistently improves Precision, Recall, and F1 while reducing FPR. These gains highlight the effectiveness of the hybrid scoring strategy in enhancing anomaly detection accuracy and stability across federated settings.

B. Convergence Ability

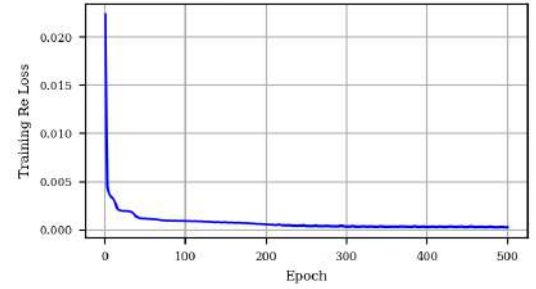


Fig. 2. Reconstruction loss of Client 1 on the N-BaIoT dataset.

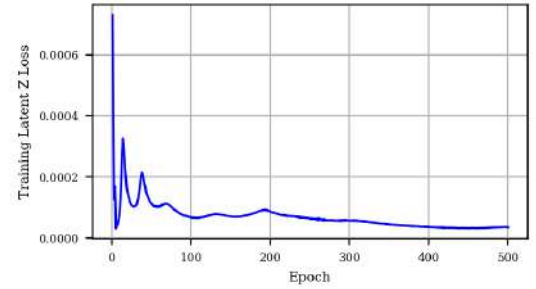


Fig. 3. Latent vector representation of Client 1 on the N-BaIoT dataset.

To examine the convergence characteristics of the proposed FedSAE-HS method, we visualize the training behavior of Client 1 on the N-BaIoT dataset. Specifically, we present two learning curves over 500 communication rounds: the RE, and the z loss.

As shown in Figure 2, the reconstruction loss decreases sharply within the first 50 rounds and then converging smoothly. Meanwhile, the latent loss (Fig. 3) shows moderate fluctuations during the initial stages but becomes increasingly stable as training progresses. These observations indicate that the model successfully optimizes both components of the loss function in parallel.

The consistent reduction and stabilization of both RE and z loss components strengthen the effectiveness of the proposed unified objective. This behavior demonstrates that FedSAE-HS maintains stable optimization dynamics across local clients.

VI. CONCLUSION

This paper introduced FedSAE-HS, a semi-supervised federated anomaly detection framework for IDS that integrates a hybrid scoring strategy combining reconstruction error and latent deviation. Unlike methods relying on a single signal, this dual-cue design enhances sensitivity to diverse anomalies while preserving a simple, label-free AE architecture.

Comprehensive experiments on the N-BaIoT and WSN_DS datasets show that FedSAE-HS consistently surpasses four semi-supervised FL baselines, achieving higher per-client ROC-AUC, superior average Precision, Recall, F1-score, and lower FPR, along with faster convergence and stable training dynamics. These results confirm the practical value of FedSAE-HS in decentralized, privacy-sensitive environments. Future work will explore dynamic weighting of scoring signals, personalization strategies, and large-scale real-time deployments.

REFERENCES

- [1] A. Khraisat, A. Alazab, M. Alazab, A. Obeidat, S. Singh, and T. Jan, "Federated learning for intrusion detection in iot environments: a privacy-preserving strategy," *Discover Internet of Things*, vol. 5, no. 1, p. 72, 2025.
- [2] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.
- [3] J. Zhao, S. Bagchi, S. Avestimehr, K. Chan, S. Chaterji, D. Dimitriadis, J. Li, N. Li, A. Nourian, and H. Roth, "The federation strikes back: A survey of federated learning privacy attacks, defenses, applications, and policy landscape," *ACM Computing Surveys*, vol. 57, no. 9, pp. 230:1–230:37, 2025.
- [4] M. J. Idrissi, H. Alami, A. El Mahdaoui, A. El Mekki, S. Oualil, Z. Yartaoui, and I. Berrada, "Fed-anids: Federated learning for anomaly-based network intrusion detection systems," *Expert Systems with Applications*, vol. 234, p. 121000, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417423015026>
- [5] P. T. Duy, D. T. T. Hien, T. D. Luong, N. H. Quyen, and V.-H. Pham, "Fed-evolver: An automated evolving approach for federated intrusion detection system using adversarial autoencoder in sdn-enabled networks," *Internet of Things*, vol. 28, p. 101397, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S254266052400338X>
- [6] T. Zhang, C. He, T. Ma, L. Gao, M. Ma, and S. Avestimehr, "Federated learning for internet of things," in *Proceedings of the 19th ACM conference on embedded networked sensor systems*. Association for Computing Machinery, 2021, pp. 413–419.
- [7] T. A. Vu, T. P. Tran, L. Vu, and Q. U. Nguyen, "Shrink autoencoder for federated learning-based iot anomaly detection," in *2022 9th NAFOS-TED Conference on Information and Computer Science (NICS)*, 2022, pp. 383–388.
- [8] R. Beuran *et al.*, "Fedmse: Semi-supervised federated learning approach for iot network intrusion detection," *Computers & Security*, vol. 151, p. 104337, 2025.
- [9] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.
- [10] L. Xing, K. Wang, H. Wu, H. Ma, and X. Zhang, "Fl-mae: An intrusion detection method for the internet of vehicles based on federated learning and memory-augmented autoencoder," *Electronics*, vol. 12, no. 10, 2023. [Online]. Available: <https://www.mdpi.com/2079-9292/12/10/2284>
- [11] K. Zhang, Y. Jiang, L. Seversky, C. Xu, D. Liu, and H. Song, "Federated variational learning for anomaly detection in multivariate time series," in *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, 2021, pp. 1–9.
- [12] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International conference on learning representations*, 2018.
- [13] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. Association for Computing Machinery, 2017, pp. 665–674.
- [14] L. Vu, V. L. Cao, Q. U. Nguyen, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, "Learning latent representation for iot anomaly detection," *IEEE Transactions on Cybernetics*, vol. 52, no. 5, pp. 3769–3782, 2022.
- [15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.
- [16] S. Malaviya, M. Shukla, P. Korat, and S. Lodha, "Fedfame: a data augmentation free framework based on model contrastive learning for federated semi-supervised learning," in *Proceedings of the 38th ACM/SIGAPP symposium on applied computing*, 2023, pp. 1114–1121.
- [17] Y. Li and Y. Li, "Semi-supervised federated learning for collaborative security threat detection in control system for distributed power generation," *Engineering Applications of Artificial Intelligence*, vol. 148, p. 110374, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0952197625003744>
- [18] O. Aouedi, G. Jajoo, and K. Piamrat, "Metals: semi-supervised federated active learning for intrusion detection systems," in *2024 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2024, pp. 1–6.
- [19] Y. Xu, L. Wang, H. Xu, J. Liu, Z. Wang, and L. Huang, "Enhancing federated learning with server-side unlabeled data by adaptive client and data selection," *IEEE Transactions on Mobile Computing*, vol. 23, no. 4, pp. 2813–2831, 2023.
- [20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: <http://www.deeplearningbook.org>
- [21] V. L. Cao, M. Nicolau, and J. McDermott, "Learning neural representations for network anomaly detection," *IEEE Transactions on Cybernetics*, vol. 49, no. 8, pp. 3074–3087, 2019.
- [22] Z. Cheng, S. Wang, P. Zhang, S. Wang, X. Liu, and E. Zhu, "Improved autoencoder for unsupervised anomaly detection," *International Journal of Intelligent Systems*, vol. 36, no. 12, pp. 7103–7125, 2021.
- [23] M. Said Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, "Network anomaly detection using lstm based autoencoder," in *Proceedings of the 16th ACM symposium on QoS and security for wireless and mobile networks*, 2020, pp. 37–45.
- [24] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *ACM Computing Surveys*, 2019. [Online]. Available: <https://arxiv.org/abs/1901.03407>
- [25] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-baiot: Network-based detection of iot botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018.
- [26] I. Almomani, B. Al-Kasasbeh, and M. AL-Akhras, "Wsn-ds: A dataset for intrusion detection systems in wireless sensor networks," *Journal of Sensors*, vol. 2016, pp. 1–16, 2016.
- [27] Google Developers, "Classification: Accuracy, recall, precision, and related metrics," <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall>, 2025, access Date: 20-4-2025.

Building Reliable Covert Timing Channels via Error Correction Codes

Manh Hung Ngo*, Van Tong*, Duc Tran*, Giang Nguyen Linh* and Tuyen Ngoc Le[†]

*Hanoi University of Science and Technology, Hanoi, Vietnam

Email: hung.nm226083@sis.hust.edu.vn, {vantv, ductq, giangnl}@soict.hust.edu.vn

[†]Department of Electronic Engineering, Ming Chi University of Technology, New Taipei, 24301 Taiwan, ROC

Center for Reliability Engineering, Ming Chi University of Technology, New Taipei, 24301 Taiwan, ROC

Email: tuyennl@mail.mcut.edu.tw

Abstract—Improving the reliability of covert timing channels (CTCs) influenced by varying network factors such as latency and packet loss is a critical challenge. To address this difficulty, we empirically assess the effectiveness of applying error correction code (ECC) techniques as potential solutions. The results indicate that ECCs greatly increase data recovery; however, the performance between schemes differs significantly. Our research reveals the Reed-Solomon code as the most successful way to handle erroneous data in CTCs, while other ECC techniques, such as LDPC in the employed configuration, do not demonstrate an effective fit for these types of channels. We also quantify the reliability-time trade-off for processing larger data, establishing a practical limit on the size of fully recoverable messages. This study gives in-depth quantitative insights and practical design criteria, contributing to the building of more reliable hidden network channels.

Index Terms—Network covert channel, Covert Timing Channels, Forward Error Correction, Reed-Solomon codes

I. INTRODUCTION

A network covert channel is a hidden communication pathway in which messages are secretly transmitted to the recipient by using systems or protocols in unexpected manners within the network environment [1]. Network covert channels can be either covert storage channels (CSCs) or covert timing channels (CTCs). The former incorporates information by altering packet fields or metadata, including unused header bits and TCP sequence numbers. The latter encodes data by employing the temporal characteristics of network traffic, such as packet arrangements, packet transmission rates, and inter-packet delays [2].

Although CSCs can offer faster transmission and higher throughput, these advantages come with the cost of increased probability of detection. CSCs can be easily identified using header normalization, protocol conformance validation, or modern deep packet inspection. CTCs, on the other hand, have a lower bitrate but are much more challenging to detect. Lou et al. [3] introduced TCPScript, a covert communication encoding method via TCP data bursts, allowing greater reliability through clustering methods and exploiting rich TCP protocol attributes. TCPScript demonstrates higher data rates and decoding precision compared to other tools, such as IPTIME and JitterBug. Furthermore, El-Atawy and Al-Shaer [4] examined the technique of packet reordering to establish a covert channel. In

their paper, they improve the reliability of this channel by using specific permutations, with a distribution designed to replicate real traffic and augment stealthiness.

Recently, CTCs that rely on the manipulation of inter-packet delays (IPDs) have gained popularity. Archibald et al. [5] designed a model-based CTC in which IPDs of packet streams generated by Skype are modulated in accordance with a statistical model of the application. The encoded symbols are particularly modulated using multiple IPDs (m-IPDs), which can be allocated to different partitions of delay distribution. The main goal of this approach is to avoid buffer underruns and overruns. These CTCs have been demonstrated to be largely undetected, even when subjected to Kullback-Leibler divergence and Kolmogorov-Smirnov tests.

Zhang et al. [6] observed that IPDs might be affected by network packet loss. These issues may cause decoding errors, ultimately leading to the loss of confidential data. The authors suggested adopting Forward Error Correction (FEC) to minimize data loss by transmitting a specific quantity of redundant information. In their work, the Reed-Solomon (RS) code was utilized, which is especially advantageous for burst-error correction and efficient for channels with memory capabilities. Zhang et al. [7] employed the Gray Code with variable-length encoding to improve robustness to network delays. The authors proposed utilizing interleaving to rearrange speech and video packets to address the issue, as the IPDs of Voice over Long-Term Evolution (VoLTE) traffic display notable regularity, making the modifications easily detectable.

In both [6] and [7], the RS and Gray code were utilized to enhance the robustness of User Datagram Protocol (UDP)-based channels. In practice, UDP may experience unreliability due to packet loss and reordering. The Transmission Control Protocol (TCP) is a better alternative to the User Datagram Protocol (UDP), as it is favored for reliable data transmission by adversaries. The prevalence of TCP in network traffic further complicates the detection of TCP-based covert channels. Nevertheless, TCP presents new issues for CTCs as its reliability characteristics and congestion mechanisms can reduce the decoding accuracy. In CTCs, because data is encrypted before being transmitted, a single erroneous bit during decoding may lead to a complete failure in message recovery. This intrinsic

unreliability results in a significant gap in this research, which aims to enhance the dependability of CTCs.

This research aims to strengthen the reliability of TCP-based CTCs against disruptive network phenomena, such as latency and packet loss, through the introduction of Forward Error Correction (FEC). Although the principles of FEC are well-established and have been investigated for improving covert channels [8], [9], there appears to be a limited systematic examination and comparison of this technique for optimizing TCP-based CTCs under real-world network environments. To this end, this paper provides the following contributions:

- We offer a comprehensive evaluation of the implications of various delay settings on the ability to recover confidential messages transmitted via TCP-based CTCs.
- We examine the effectiveness of three error correction methodologies: Reed-Solomon (RS), Bose-Chaudhuri-Hocquenghem (BCH), and modern iterative codes such as Low-Density Parity-Check (LDPC) in enhancing the reliability of CTCs.

The remainder of this paper is structured as follows. Section II describes the fundamentals of the explored error-correction code (ECC) schemes. In Section III, we outline the methodology used in our research. Afterwards, Section IV presents our experimental setup and analyzes our comparative performance results. Finally, Section V concludes the paper and discusses potential avenues for future research.

II. FORWARD ERROR CORRECTION

This section presents the foundations of FEC and different Error Correction Codes (ECC), including RS, BCH, and LDPC.

A. FEC Principles

FEC is a family of techniques that enhance data reliability by adding redundant information within transmitted data, enabling error correction without retransmission [10]. Error Correction Codes (ECCs) are a subset of FEC. While FEC includes various methods for handling data loss, this paper specifically looks at ECC families because they have clear correction abilities and are easier to decode in CTCs that rely on inter-packet delay (IPD).

Given k information symbols, FEC aims at encoding them into n codeword symbols. The error-correcting capability t is defined as

$$t = \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor \quad (1)$$

where $d_{\min}$ is the code's minimum distance. The floor function $\lfloor x \rfloor$ ensures an integer bound.

In addition to t , the code rate $R = k/n$ is used to measure efficiency. A lower R indicates increased resilience to network disruptions such as packet loss or latency, albeit at the cost of increased overhead. Hence, to balance overhead costs and dependability in CTCs, it is crucial to use suitable FEC systems with optimal coding rates.

B. Reed-Solomon Codes

Reed-Solomon (RS) codes use m bits for each symbol and function in the finite field $\mathbb{F}_{2^m}$. An $RS(n, k)$ code can correct $t = \lfloor n - k/2 \rfloor$ symbol errors and has an information length of k ($k < n$) and a block length of $n = 2^m - 1$ [11]. In RS, the generator polynomial $g(x)$ is defined as

$$g(x) = \prod_{j=1}^{2t} (x - \alpha^j) \quad (2)$$

where α is a primitive element of $\mathbb{F}_{2^m}$. According to [12], upon receiving a polynomial $r(x) = c(x) + e(x)$, where $e(x)$ denotes the error polynomial, the syndrome polynomial is calculated as

$$S(x) = \sum_{j=0}^{2t-1} S_j x^j \text{ where } S_j = r(\alpha^{j+1}) \quad (3)$$

RS codes achieve the singleton bound with minimum distance $d_{\min} = n - k + 1$. The symbol-level error correction of RS codes makes them especially proficient in mitigating burst errors in timing channels.

C. BCH Codes

A binary Bose–Chaudhuri–Hocquenghem (BCH) [13] code operates on $\mathbb{F}_2$. It has a block length of $n = 2^m - 1$, an information length of $k \geq n - mt$, and the ability to correct t bit errors with a minimum distance $d_{\min} \geq 2t + 1$ for positive integers $m \geq 3$ and $t < 2^{m-1}$. The generator polynomial is given as

$$g(x) = \text{LCM}[M_1(x), M_3(x), \dots, M_{2t-1}(x)] \quad (4)$$

where $M_i(x)$ is the minimal polynomial of α^i over $\mathbb{F}_2$ and LCM is the least common multiple. Given a received vector r , the syndrome can be computed as

$$S_j = r(\alpha^j) = \sum_{i=0}^{n-1} r_i (\alpha^j)^i \quad (5)$$

The bit-level error correction of BCH codes renders them suited for covert timing channels influenced by random bit error, where it appears isolated rather than clustered.

D. LDPC Codes

Low-density parity-check (LDPC) codes [14] have attained popularity in coding theory. The codes are fully parallelizable and can operate at rates approaching Shannon's capacity [15]. Assume that $\mathbf{H}$ is a parity-check matrix of size $m \times n$. An LDPC code is defined by a set of codewords $\mathbf{c}$ that satisfy the equation $\mathbf{H}\mathbf{c}^T = 0$. This code is deemed linear since any linear combination of codewords results in another valid codeword. The term low-density refers to the relatively small number of 1s present in the matrix $\mathbf{H}$.

LDPC decoding relies on Belief Propagation (BP) and the Tanner graph of the code [16]. The Tanner graph is a bipartite graph that consists of m check nodes and n variable nodes. The BP algorithm operates by circulating computations throughout the Tanner graph. It is important to note that the decoding

performance is impeded by small cycles in the Tanner graph. Hence, it is crucial to ensure that the girth of the graph is not excessively small [17].

Although LDPC codes are known to approach the Shannon limit, this is often accomplished via soft-decision decoding algorithms such as Belief Propagation, which require probabilistic channel information (Log-Likelihood Ratios). For the scope of this initial investigation, we chose a more straightforward and less sophisticated approach to establishing a performance baseline. As a result, we adopted a hard-decision decoding (Hard-LDPC) approach. This involved demodulating the received Inter-Arrival Times into a binary stream, which was then processed using an iterative Weighted Bit-Flipping (WBF) method.

III. METHODOLOGY

This paper presents an extension of the covert channel educational analysis protocol (CCEAP) [18]. Specifically, we utilize CCEAP to create TCP-based CTCs and validate their reliability under various network conditions, both with and without ECCs.

A. ECC Configurations

To evaluate the channel's reliability, we selected three distinct Error Correction Codes (ECCs), each chosen to test against the error characteristics of a covert timing channel. A key factor in this context is the trade-off between reliability and overhead. A lower code rate offers more protection but reduces the throughput of the covert timing channel. As shown in Table I, we aimed for a coding rate of around 0.85 across all schemes. RS and LDPC were configured accordingly, while BCH required a slightly lower rate due to construction constraints. This rate was chosen as a realistic compromise, since it allows for extensive error correction while maintaining a reasonable data throughput.

1) *Reed-Solomon (RS)*: Covert timing channels over the internet are highly prone to burst errors. Network phenomena such as sudden latency spikes or TCP congestion control can distort a series of consecutive inter-packet delays, leading to a cluster of correlated errors.

The RS(255, 223) arrangement was chosen due to its compatibility with $\mathbb{GF}(2^8)$. The codeword length of $n = 255$ is the natural limit for an 8-bit symbol code, whereas $k = 223$ gives 32 bytes of redundancy. This architecture was chosen to provide a strong defense against predicted burst failures while still retaining high data throughput.

2) *BCH*: To test the alternative hypothesis that channel errors are random, we used the bit-oriented BCH coding.

We selected BCH(255, 191) to maintain a consistent codeword length of $n = 255$ with the RS code, enabling a fair and direct comparison of their performance on the same block size. The data length of $k = 191$ represents a standard parameter set for this configuration that adheres to the code's construction rules while providing a significant data payload.

3) *Hard-LDPC*: This modern iterative code was included to serve as a high-performance baseline against the classic algebraic codes. The configuration was chosen to create a modern benchmark with a standard block size. A codeword length of $n = 1024$ is standard in digital communications, and the data length of $k = 834$ was intentionally chosen to ensure that the percentage of data to redundancy was equivalent to the other codes, ensuring a fair comparison of overhead.

B. System Architecture

Fig. 1 depicts our system architecture, which comprises three primary conceptual components: a Sender entity, the Network Channel, and a Receiver entity.

We assume that an adversary possesses the capability to acquire sensitive information, including user passwords and data. To bypass the protection mechanisms, he can utilize the prevalent AES algorithm to encrypt the data and transfer the encrypted information through a CTC. The Sender entity is responsible for transmitting the data by modulating the packet delay, employing IPD modulation, and augmenting with ECC for robustness. As illustrated in Fig. 1 the data undergoes preprocessing through various ECC techniques, specifically RS, BCH, and LDPC. The specific parameters used by these methods are presented in Table I. In general, ECC introduces redundancy by generating bits or symbols from these parameters and, hence, producing an encoded data block.

In IPD modulation, a binary 1 is represented by a short delay d_1 , while a binary 0 is represented by a long delay d_0 . The Sender entity then transmits a sequence of packets following the CCEAP protocol over the established TCP connection. After each call to the send() function, the application introduces a pause using the usleep() function, which corresponds to the target IPD (d_0 or d_1).

The Receiver entity performs the reverse operations. Upon receiving data, it parses the CCEAP packets and computes inter-arrival times (IATs) as the time difference between successive packets. Each measured IAT is checked against a decision threshold, typically set at the midpoint between d_0 and d_1 . An IAT shorter than the threshold is interpreted as binary 1, while IAT longer than the threshold is interpreted as binary 0. This processing step may lead to an erroneous bitstream. This erroneous bitstream arises from fluctuations in network conditions, resulting in variations in end-to-end packet delay. Given that the data is encrypted with AES, a single bit error may lead to a complete failure in data recovery. To overcome this challenge, the Receiver entity performs ECC decoding and uses known parameters to detect and attempt to correct errors within the bitstream, up to the correction capacity t .

TABLE I: The parameters and error correction capability of different ECC techniques

Error Correction Code	n	k	Rate	EC Capabilities
RS	255	223	0.874	16 byte-errors
BCH	255	191	0.749	8 bit-errors
Hard-LDPC	1024	834	0.814	Iterative

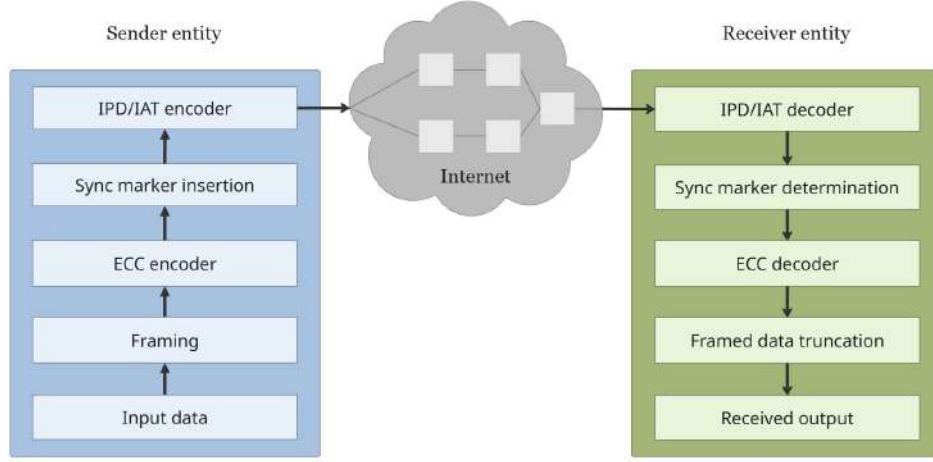


Fig. 1: End-to-end system architecture for IPD-based CTCs with ECC

C. Evaluation Metrics

The performance of CTCs is measured using the **Packet Success Rate (PSR)** metric. PSR measures the ability to successfully decode 500-byte packets following each transmission over the simulated network channel. This statistic is defined as the ratio of successful decodings (i.e., the decoded bitstream contains no errors or errors are rectified by the packet-level ECC, if present) compared to the total number of 500-byte packet transmissions. The PSR is hence calculated as

$$\text{PSR} = \left(\frac{S}{T} \right) \times 100\% \quad (6)$$

where S represents the total number of successful packet decodings, and T denotes the total number of packet transmissions.

Message Recovery Rate (MRR) evaluates the accuracy of a data file or message (with sizes 100B, 1KB, 4KB) that is recovered after a single transmission over a network channel. MRR is defined as the percentage of successfully decoded bits (compared to the original data) at the receiver following the conclusion of transmission, demodulation, and file-level ECC decoding. MRR is as follows

$$\text{MRR} = \left(\frac{C}{N} \right) \times 100\% \quad (7)$$

where C is the number of correctly decoded bits, and N is the total number of bits in the original data.

IV. EXPERIMENTS

To accurately assess the network conditions in which our CTC operates, we utilized the **iperf** tool and **ping** command to measure network parameters. The experimental configuration consisted of a client and a server connected via the TCP protocol. **iperf** and **ping** were configured to generate network traffic and gather critical parameters, including *average bit rate*, *latency*, and *packet loss*. Each test involved transmitting 357

MB of data for 60 seconds, ensuring a long enough measuring duration to collect reliable results.

TABLE II: Measured TCP network path characteristics

Characteristics	Measured value
Data transfer size	357 MB
Average bitrate	49.8 Mbps
Average latency	29.2 ms
Average packet loss	0.24%

To ensure accuracy, we run the **iperf** and **ping** 10 times, and calculate the mean values of the parameters. The measurement findings are summarized in Table II, which offers an overview of the network characteristics during the experiment. Based on this finding, we conduct further tests to evaluate the performance of CTCs under the stated network parameters.

A. Experimental Setup

We conducted two distinct experiments to assess and enhance the reliability of the TCP-based covert channel utilizing ECCs. The initial experiment aimed to identify the ideal delay pairs for effective CTCs via the Internet. Initially, a 500-byte packet was delivered over the LAN as a baseline and subsequently over the Internet with three distinct delay pairs: 20/50 ms (low delay), 50/100 ms (medium delay), and 100/200 ms (high delay). Note that the delays were measured in **milliseconds**. These combinations were selected to represent constant to poor internet conditions. We evaluated four configurations: one without ECC and three that used various types of ECC codes. The procedure involves transmitting packets via LANs and the Internet for various configurations and latencies, followed by decoding and analysis. The collected data includes success rate, defined as the percentage of accurately recovered messages utilized to identify the most dependable delay pairs across varying network conditions.

The second experiment aimed to identify the largest packet size that could guarantee a 100% recovery rate based on the

best delay pair found in the first experiment. Messages were delivered in varied sizes and transmitted over the Internet, using the delay pair determined in the previous experiment. The data sizes represented each specific file: 100B for a text file carrying a command delivered over the Internet, and 1KB and 4KB for the stolen password files. The implementation procedure includes transmitting the packet, decoding the received message, and evaluating the recovery accuracy. The measurement was the recovery rate for each packet size, defined as the proportion of messages retrieved completely without mistakes. This experiment eventually established the maximum packet size that could consistently achieve a 100% recovery rate, thereby setting a practical limit on the capacity of the IPD covert channel under the conditions examined.

B. Results

This section presents the quantitative results from the two experiments described in Section IV-A, which aim to evaluate the performance and determine the limitations of the CTCs when applying error-correcting codes under simulated network conditions, as summarized in Table II.

The initial experiment assesses the efficacy of ECC schemes for 500-byte packets across various delay conditions, with findings detailed in Table III. This experiment evaluates the decoding success rate of 500-byte packets delivered over a LAN, serving as a baseline reference, and over a simulated Internet channel with three delay pairs. The four configurations analyzed consist of the No ECC scenario and three error correction coding methodologies discussed in Section IV-A.

Table III indicates a substantial difference between optimal channel parameters and actual internet circumstances. The LAN channel attains a 100% Packet Success Rate, whereas the non-ECC scenario rapidly fails, peaking at 17% even under the highest delay pair (100/200 ms). This underlines the severity of faults in real-world channels and the necessity for error correction techniques. The implementation of ECC coding techniques demonstrates the potential to significantly enhance PSR. For the delay pair of 20/50 ms, none of the three ECC codes have any impact, resulting in 0% PSR. For the delay pair of 50/100 ms, RS attains a 10% PSR, whereas both BCH and Hard-LDPC maintain a PSR of 0%. The differences in error recovery among ECC codes become most noticeable at the 100/200 ms delay pair, with RS displaying superior performance at 87% PSR, substantially surpassing Hard-LDPC's 60% PSR and BCH's 47% under the same conditions. All three ECC codes considerably improve PSR compared to non-ECC (17%) for the delay pair of 100/200 ms, suggesting their efficacy as correction solutions in error-prone network environments.

The results from the first experiment show that the 100/200 ms delay pair is where ECC codes demonstrate the most resilience and RS is the most effective scheme in this test set for 500-byte packets (87% PSR). The 100/200 ms delay pair was, thus, selected for the setup of the second experiment to test the limit on the packet size that can be reliably conveyed.

The goal of the second experiment is to determine the largest message size that may achieve a 100% Message Recovery

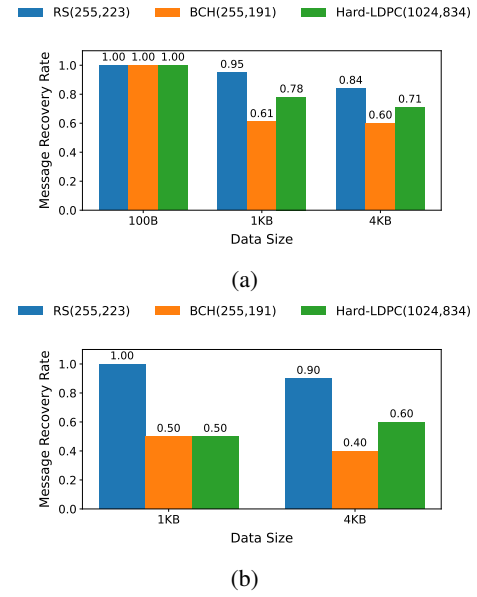


Fig. 2: The recovery rate of different ECC codes under different packet sizes: (a) when data is sent without fragmentation and (b) when data is divided into smaller packets.

Rate (MRR) using the most efficient configuration identified in Experiment 1. This experiment evaluates the recovery rates for messages of 100B, 1KB, and 4KB transmitted over the covert channel using different ECC schemes. We investigate two distinct transmission strategies: sending the message as a single, continuous bitstream (non-fragmented) and dividing the message into a sequence of smaller packets (fragmented).

Fig. 2a illustrates the MRR when messages are transmitted as a single, uninterrupted stream. The data reveals a clear trend: as the message size increases, the probability of uncorrectable errors grows, leading to a decline in recovery rates across all codes. For a small 100B message, all three ECCs achieve perfect recovery (100% MRR). However, for the 1KB message, the performance of RS(255,223) degrades slightly to 95%, while BCH(255,191) and Hard-LDPC(1024,834) experience more substantial drops to 61% and 78%, respectively. This trend continues with the 4KB message, where RS maintains superior performance at 84% MRR, compared to 71% for Hard-LDPC and 60% for BCH. These results underscore that for continuous transmissions over a noisy channel, longer data streams are more susceptible to accumulating errors that exceed the correction capability of the codes.

In contrast, Fig. 2b shows the outcomes when the message is fragmented into smaller packets before transmission. This strategy was evaluated for the larger 1KB and 4KB messages where the non-fragmented approach showed degradation. The fragmentation strategy substantially enhances the performance of the RS code. Most notably, for a 1KB message, this approach allows RS to achieve a full 100% recovery rate, a significant improvement from the 95% seen in the non-fragmented case. Its MRR for the 4KB file also improves from 84% to 90%.

TABLE III: Packet success rate for 500-Byte packet decoding across varying delay pairs, measured in milliseconds

Delay pairs	LAN	Internet			
		No ECC	RS (255,223)	BCH (255,191)	Hard-LDPC (1024,834)
20/50 ms	100%	0%	0%	0%	0%
50/100 ms	100%	0%	10%	0%	0%
100/200 ms	100%	17%	87%	47%	60%

Achieving greater reliability, particularly when dealing with large amounts of data, involves certain trade-offs. The time it takes for packets to traverse a non-ideal network connection could increase the probability of errors or packet loss. To mitigate this, dividing data into smaller packets is proposed to ensure reliability, as Fig. 2b indicates. However, if the packets split into too many smaller units, the overall time spent transmitting and processing each packet can grow considerably, making this method impractical for transferring large amounts of data. As a result, CTCs, in this context, are more suitable for carrying short messages or files, often commands, or sensitive data files of restricted size (e.g., password files).

The trade-off between reliability and transmission time is illustrated by the processing time statistics. Table IV presents the time required to transmit and decode entire 4 KB messages for each coding scheme with and without fragmentation.

TABLE IV: Transmission and decoding time for 4 KB messages with and without fragmentation strategies

Coding scheme	Non-fragmented (4 KB)	Fragmented (4 KB)
BCH(255,191)	145 min	175 min
RS(255,223)	103 min	150 min
Hard-LDPC(1024,834)	107 min	135 min

As shown in Table IV, fragmentation consistently increases the total time required. For the RS code, achieving a 6 percentage point increase in MRR (from 84% to 90%) for the 4KB message comes at the cost of an additional 47 minutes, making the fragmented transmission approximately 1.46 times longer than the continuous one. Although fragmentation improves the recovery rate for large messages when using RS, it does so at the cost of significantly higher processing time.

V. CONCLUSION AND FUTURE WORK

This paper first established that a TCP-based covert timing channel is inherently unreliable, demonstrating a baseline Packet Success Rate (PSR) of just 17% for uncoded transmissions. To overcome this, we evaluated a fragmentation strategy that treats each file chunk as an individually correctable packet. This approach proved highly effective, transforming the unreliable channel into a robust data exfiltration method. When combined with Reed-Solomon codes, our strategy achieved a 100% recovery rate for files up to 1 KB, establishing a practical limit for reliable covert transfers. In the future, we will focus on implementing LDPC with a soft-decision decoder to provide a more comprehensive comparison.

ACKNOWLEDGEMENT

This work is supported by the Ministry of Science and Technology of Vietnam (MOST) under Grant No. KC.01.04/21-30.

REFERENCES

- [1] S. Al-Eidi, O. Darwish and Y. Chen, "Covert timing channel analysis either as cyber attacks or confidential applications," *Sensors*, vol. 20, no. 8, 2417, Apr. 2020.
- [2] S. Wendzel, S. Zander, B. Fechner, and C. Herdin, "Pattern-based survey and categorization of network covert channel techniques," *ACM Computing Surveys*, vol. 47, pp. 50:1-26, Apr. 2015.
- [3] X. Luo, E. W. W. Chan and R. K. C. Chang, "TCP covert timing channels: Design and detection," in 2008 IEEE International Conference on Dependable Systems and Networks With FTCS and DCC (DSN), Anchorage, AK, USA, 2008, pp. 420-429.
- [4] A. El-Atawy and E. Al-Shaer, "Building covert channels over the packet reordering phenomenon," in IEEE INFOCOM 2009, Rio de Janeiro, Brazil, 2009, pp. 2186-2194.
- [5] R. Archibald and D. Ghosal, "Design and analysis of a model-based covert timing channel for Skype traffic," in 2015 IEEE Conference on Communications and Network Security (CNS), Florence, Italy, 2015, pp. 236-244.
- [6] L. Zhang, G. Liu, J. Zhai, and Y. Dai, "Improving reliability of covert timing channel to packet loss," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, pp. 544-553, 2015.
- [7] X. Zhang, L. Zhu, X. Wang, C. Zhang, H. Zhu, and Y. Tan, "A packet-reordering covert channel over VoLTE voice and video traffics," *Journal of Network and Computer Applications*, vol. 126, pp. 29-38, 2019.
- [8] J. Keller, S. Imhof, and P. Sobe, "Error correction and erasure codes for robust network steganography," *Journal of Systems Architecture*, vol. 153, 103191, 2024.
- [9] A. Marzin, M. Schwartz and M. Segal, "Covert communication by exploiting error-correcting codes," in 2024 20th International Conference on the Design of Reliable Communication Networks (DRCN), Montreal, QC, Canada, 2024, pp. 143-150.
- [10] Z. Yuan and X. Zhao, "Introduction of forward error correction and its application," in 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), Yichang, China, 2012, pp. 3288-3291.
- [11] S. B. Wicker and V. K. Bhargava, "An introduction to Reed-Solomon codes," in *Reed-Solomon Codes and Their Applications*, IEEE, 1994, pp. 1-16.
- [12] J. Vinck, "Coding concepts and Reed-Solomon codes," May 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2205.01044>
- [13] D. Azinovic, K. Tittelbach-Helmrich, and Z. Stamenkovic, "Performance investigation on BCH codec implementations," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2016, pp. 280-285.
- [14] Z. Tu and S. Zhang, "Overview of LDPC codes," in 7th IEEE International Conference on Computer and Information Technology (CIT 2007), Aizu-Wakamatsu, Japan, 2007, pp. 469-474.
- [15] T.J. Richardson, M.A. Shokrollahi, and R.L. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE transactions on information theory*, 47(2), 619-637.
- [16] A. Shokrollahi, "LDPC codes: an introduction," in *Coding and Cryptography*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 85-110.
- [17] D. Declercq, David, M. Fossorier, and E. Biglieri, "Channel Coding: Theory, Algorithms, and Applications," Academic Press Library in Mobile and Wireless Communications. Academic Press, 2014.
- [18] S. Wendzel and W. Mazurczyk, "POSTER: An educational network protocol for covert channel analysis using patterns," in *Proc. ACM Conf. Comput. Commun. Secur.*, New York, NY, USA, 2016, pp. 1739-1741. [Online]. Available: <https://doi.org/10.1145/2976749.2989037>

Privacy-Preserving Personalized Federated Learning for Encrypted Traffic Classification

Hai Anh Tran, Van Tong, Duc Tran and Giang Nguyen*
Hanoi University of Science and Technology, Hanoi, Vietnam
Email: {anhth, vantv, ductq, giangn}@soict.hust.edu.vn

Abstract—Effective encrypted traffic classification is crucial for modern network management, impacting security monitoring, Quality of Service (QoS) assurance, and network optimization. Traditional centralized classification methods pose significant privacy and scalability concerns, particularly in large-scale distributed communication networks. To overcome these challenges, we propose a Personalized Encrypted trAffic classification with diffeRential privacy-enhanced Learning (PEARL) framework designed specifically for encrypted traffic classification across heterogeneous network domains. Our methodology leverages a globally shared feature extractor combined with client-specific classifiers, enabling personalized adaptations to diverse network conditions. To robustly address privacy risks during federated collaboration, we incorporate Differential Privacy (DP), adding controlled noise to client updates before aggregation. Experimental evaluations demonstrate that PEARL maintains high classification accuracy, achieving 93.5% accuracy and 93.1% F-score with DP, closely approaching centralized performance (95.7% accuracy). Furthermore, personalization improves client-specific accuracy by up to 10% compared to global models, while DP mechanisms introduce only moderate computational overhead, ensuring practical deployment in real-world network environments.

Index Terms—Encrypted Traffic Classification, Personalized Federated Learning, Differential Privacy, Network Security and Management, Privacy-Preserving Machine Learning

I. INTRODUCTION

The exponential growth of encrypted network traffic, driven by widespread adoption of protocols such as TLS/SSL and VPN, has significantly enhanced user privacy and data confidentiality. While encryption provides essential security and privacy benefits, it simultaneously complicates essential network management tasks such as traffic monitoring, anomaly detection, and resource allocation. Traditional deep packet inspection techniques, which rely on accessing payload data, become ineffective when dealing with encrypted traffic, thereby reducing network operators' capability to ensure Quality of Service (QoS), enforce security policies, and detect malicious activities or network intrusions [1], [2]. Consequently, advanced machine learning (ML)-based methods that leverage metadata—such as packet size distributions, inter-arrival times, and handshake patterns—have become critical for enabling effective classification and management of encrypted network traffic.

Centralized ML-based traffic classification methods, however, pose significant limitations for deployment within mod-

ern distributed telecommunications infrastructures. Centralized data collection involves transmitting potentially sensitive traffic metadata to a centralized server, raising critical privacy concerns and regulatory compliance issues. Additionally, scalability and network overhead associated with centralized data aggregation become prohibitive as the number of network domains and traffic volumes increase.

Federated Learning (FL) offers an effective decentralized alternative, enabling network domains to collaboratively train a global ML model without sharing raw data [3]. Despite its advantages, FL is still vulnerable to privacy inference attacks, where adversaries—either a curious server or malicious clients—attempt to reconstruct sensitive information from exchanged model updates [4], [5]. Furthermore, standard federated models often struggle with performance degradation due to data heterogeneity across different network domains, which significantly impacts network reliability and overall QoS.

To address these critical challenges, we propose a **Personalized Encrypted trAffic classification with diffeRential privacy-enhanced Learning (PEARL)** framework specifically designed for secure and effective encrypted traffic classification across distributed communication networks. The proposed solution features a globally shared feature extractor that captures common network-traffic characteristics, combined with client-specific classifiers to enable personalized adaptation to each network domain's unique traffic patterns. Moreover, to robustly mitigate privacy risks, we incorporate Differential Privacy (DP), adding carefully calibrated noise to client model updates prior to aggregation. This ensures strong privacy guarantees by preventing adversaries from inferring sensitive client data, while simultaneously maintaining high classification accuracy.

PEARL significantly benefits network communication infrastructures by enhancing operators' ability to securely manage network resources, reliably detect anomalies and intrusions, and maintain QoS standards across diverse network conditions. Specifically, the contributions of this paper include:

- Formulating encrypted traffic classification explicitly as a privacy-aware federated learning challenge, emphasizing practical implications for network management, security, and QoS.
- Proposing a Personalized Federated Learning (PFL) approach integrated with DP to address data heterogeneity and privacy threats effectively within distributed communication networks.

*Corresponding author.

- Evaluating the proposed framework comprehensively on real-world encrypted traffic datasets, demonstrating robust privacy protection and superior classification performance under realistic heterogeneous network scenarios.

The remainder of this paper is organized as follows. Section II reviews relevant literature on encrypted traffic classification, federated learning, and privacy-preserving techniques. Section III details the proposed framework, emphasizing personalization and differential privacy mechanisms. Section IV presents comprehensive experimental evaluations and analysis. Finally, Section V concludes the paper with practical implications and future research directions.

II. RELATED WORK

A. Encrypted Traffic Classification

Encrypted traffic classification has emerged as a critical research area due to the increasing deployment of encryption protocols like TLS/SSL and VPN, which obscure packet payload information and complicate traditional packet-inspection methods. Early efforts primarily relied on statistical analysis of network metadata, such as packet lengths, timing intervals, and flow durations, to classify encrypted traffic [6], [7]. With the advancement of ML and deep learning (DL) techniques, recent works have adopted supervised and unsupervised approaches to further improve classification accuracy and robustness. Anderson et al. [1] demonstrated that ML-based classification leveraging flow metadata can accurately identify encrypted malware traffic, while Shafiq et al. [2] performed an extensive systematic study evaluating various DL models for encrypted traffic classification tasks.

B. Federated Learning for Network Traffic Analysis

FL has attracted considerable attention in networking research due to its inherent privacy preservation and decentralized training paradigm [3], [8]. In the context of network traffic analysis, FL enables distributed clients (e.g., network gateways or IDS sensors) to collaboratively build a global model without sharing sensitive raw data. For instance, Nguyen et al. [9] proposed DIoT, a federated anomaly detection method for IoT networks, highlighting FL's capability to handle privacy-sensitive network data. Similarly, Zhao et al. [10] explored FL specifically for mobile network traffic, emphasizing FL's advantages in scalability and privacy-preserving analytics in large-scale telecommunications environments. Despite these benefits, standard federated approaches often suffer from performance degradation due to non-IID (heterogeneous) data distributions across different network domains [11]. This limitation motivates further research into personalized federated learning.

C. Personalized Federated Learning

To address heterogeneity challenges, Personalized Federated Learning (PFL) techniques have been introduced, allowing models to adapt to individual clients' unique data characteristics while leveraging shared global knowledge [12]. Recent approaches, such as personalized federated averaging [13], multi-task federated learning [14], and personalization layers

within neural networks [15], have shown promising results in scenarios characterized by significant data diversity. Notably, Arivazhagan et al. [15] introduced a multi-head FL architecture, enabling clients to personalize models effectively while still benefiting from shared representations. However, most PFL studies to date have focused on generic datasets or image/text classification tasks, leaving PFL approaches for encrypted traffic analysis relatively unexplored.

D. Privacy-Preserving Mechanisms in Federated Learning

Although FL inherently avoids centralized raw data collection, recent research highlights vulnerabilities to privacy inference attacks, where adversaries reconstruct sensitive client data from exchanged model parameters or gradients [4], [5]. DP, a mathematically rigorous privacy framework, has been widely adopted to mitigate such attacks [4], [16]. In federated learning contexts, DP is typically implemented by adding carefully calibrated noise to model updates, effectively obscuring individual client contributions while retaining global utility [4], [17]. Geyer et al. [4] showed DP's effectiveness in protecting client-level privacy within FL. Despite these advances, integrating differential privacy into PFL, particularly for encrypted traffic classification within distributed communication networks, remains understudied.

E. Discussions

While existing works have individually addressed encrypted traffic classification, FL, personalized modeling, and differential privacy, the integration of these techniques to tackle practical challenges in distributed network management and security is currently limited. Our work uniquely combines PFL and DP specifically for encrypted traffic classification. This integration addresses crucial gaps in prior literature by effectively handling data heterogeneity, enhancing privacy protection, and ensuring high classification accuracy suitable for real-world network communication scenarios.

III. METHODOLOGY

In this section, we outline the proposed approach, PEARL, for encrypted traffic classification using PFL with a multi-head model architecture. The PEARL's primary focus is on mitigating privacy attacks, where an adversary attempts to infer sensitive client data from model updates. In FL, an honest-but-curious server or malicious clients may attempt to reconstruct local data from shared gradients or weight updates. To counter these privacy threats, we incorporate DP into the aggregation process, ensuring that client updates remain private while still contributing to global learning.

A. Problem Formulation

Encrypted traffic classification is crucial for modern cybersecurity, as traditional deep packet inspection techniques are ineffective on encrypted data. To address this, ML methods leverage observable metadata (e.g., packet sizes, timing, cipher handshakes) to infer traffic types or malicious activity.

Formally, given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^M$, where $\mathbf{x}_i$ represents encrypted flow features and y_i denotes class labels (e.g., malicious or benign), the classification task involves learning a model $f(\mathbf{x}; \theta)$ that minimizes the cross-entropy loss:

$$\min_{\theta} \frac{1}{M} \sum_{i=1}^M \ell(f(\mathbf{x}_i; \theta), y_i). \quad (1)$$

Since training data is often distributed across organizations, centralized learning raises privacy and legal concerns. Hence, we adopt FL to collaboratively train classification models without sharing raw data.

B. Personalized Federated Learning Setup

To address the distributed nature of encrypted traffic data, we adopt a PFL approach with a multi-head model architecture. In FL, multiple clients (e.g., network gateways, organizational IDS deployments, or user devices) collaboratively train a global model under the coordination of a central server without sharing their raw data. Each client i maintains its local dataset $\mathcal{D}_i$ of encrypted traffic flows and only exchanges model updates with the server, ensuring data confidentiality.

Multi-Head Model Architecture: We design the federated model as a neural network with a shared base and personalized heads for each client. Let $\mathbf{w}$ denote the parameters of the shared base (feature extractor) and $\mathbf{v}_i$ denote the parameters of client i 's head (classifier layer). For an input $\mathbf{x}$, client i 's model prediction is:

$$f_i(\mathbf{x}) = g_i(h(\mathbf{x}; \mathbf{w}); \mathbf{v}_i), \quad (2)$$

where $h(\mathbf{x}; \mathbf{w})$ is the shared feature representation and $g_i(\cdot; \mathbf{v}_i)$ is the client-specific classifier. The shared base $\mathbf{w}$ is learned collaboratively across all clients, while the local head $\mathbf{v}_i$ is trained independently for each client, allowing adaptation to different traffic distributions. The federated training objective is:

$$\min_{\mathbf{w}, \{\mathbf{v}_i\}_{i=1}^N} \sum_{i=1}^N \frac{n_i}{n_{\text{total}}} \mathcal{L}_i(\mathbf{w}, \mathbf{v}_i), \quad (3)$$

where $\mathcal{L}_i(\mathbf{w}, \mathbf{v}_i)$ represents the local classification loss on client i 's encrypted traffic data. This structure ensures a balance between shared knowledge transfer and personalized decision-making per client.

C. Federated Communication Process with Differential Privacy

To mitigate privacy attacks, we apply DP at the client level before sending updates to the server. DP ensures that the contribution of any single client cannot be distinguished from the overall model updates, making it difficult for an adversary to infer specific training data.

As shown in Algorithm 1, the training process of PEARL consists of multiple communication rounds. Each communication round is described as follows:

Algorithm 1 The proposed PEARL algorithm with Differential Privacy

- 1: **Initialize:** Global backbone parameters $\mathbf{w}^{(0)}$
 - 2: **for** each communication round $t = 0, 1, 2, \dots, T - 1$ **do**
 - 3: **Server:** Broadcast $\mathbf{w}^{(t)}$ to all clients
 - 4: **for** each client i **in parallel do**
 - 5: Receive global backbone $\mathbf{w}^{(t)}$
 - 6: Perform local training on $\mathcal{D}_i$:

$$\begin{aligned} \mathbf{w}_i^{(t+1)} &= \mathbf{w}^{(t)} - \eta \nabla_{\mathbf{w}} \mathcal{L}_i(\mathbf{w}^{(t)}, \mathbf{v}_i^{(t)}) \\ \mathbf{v}_i^{(t+1)} &= \mathbf{v}_i^{(t)} - \eta \nabla_{\mathbf{v}_i} \mathcal{L}_i(\mathbf{w}^{(t)}, \mathbf{v}_i^{(t)}) \end{aligned}$$
 - 7: Add Gaussian noise for Differential Privacy:

$$\tilde{\mathbf{w}}_i^{(t+1)} = \mathbf{w}_i^{(t+1)} + \mathcal{N}(0, \sigma^2 I)$$
 - 8: Send $\tilde{\mathbf{w}}_i^{(t+1)}$ to the server
 - 9: **end for**
 - 10: **Server:** Aggregate noisy updates:

$$\mathbf{w}^{(t+1)} = \sum_{i=1}^N \frac{n_i}{n_{\text{total}}} \tilde{\mathbf{w}}_i^{(t+1)}$$
 - 11: **end for**
 - 12: **Return:** Final global backbone parameters $\mathbf{w}^{(T)}$
-

- 1) **Global Model Broadcast:** The server initializes the global backbone $\mathbf{w}^{(0)}$ and shares it with all clients.
- 2) **Local Training:** Each client trains its local model using its own data:

$$\mathbf{w}_i^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla_{\mathbf{w}} \mathcal{L}_i(\mathbf{w}^{(t)}, \mathbf{v}_i^{(t)}), \quad (4)$$

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} - \eta \nabla_{\mathbf{v}_i} \mathcal{L}_i(\mathbf{w}^{(t)}, \mathbf{v}_i^{(t)}). \quad (5)$$

- 3) **Differentially Private Update Submission:** Clients add noise to their model updates before sending them to the server to ensure privacy. Specifically, each client perturbs its update with Gaussian noise:

$$\tilde{\mathbf{w}}_i^{(t+1)} = \mathbf{w}_i^{(t+1)} + \mathcal{N}(0, \sigma^2 I), \quad (6)$$

where $\mathcal{N}(0, \sigma^2 I)$ represents Gaussian noise with variance σ^2 . This ensures that individual client updates remain private while still allowing useful information for model aggregation.

- 4) **Private Model Aggregation:** The server aggregates noisy model updates using:

$$\mathbf{w}^{(t+1)} = \sum_{i=1}^N \frac{n_i}{n_{\text{total}}} \tilde{\mathbf{w}}_i^{(t+1)}. \quad (7)$$

- 5) **Model Distribution:** The updated global backbone is shared with clients for the next round.

By incorporating differential privacy, we ensure that no single client's contribution can be uniquely identified, providing robust protection against privacy attacks while maintaining effective encrypted traffic classification.

IV. EXPERIMENTS

A. Experimental Setup

We evaluate our approach using publicly available datasets: ISCX VPN-NonVPN, CICIDS2017, and USTC-TFC2016. We simulate a FL scenario using the Flower framework with 10 clients, each representing a distinct network domain. Each client receives between 2,000 and 3,500 samples with varying class distributions to emulate realistic non-i.i.d. scenarios.

The model consists of a shared backbone for feature extraction and personalized classification heads per client. We train using the Adam optimizer ($\eta = 0.001$), batch size of 64, and $E = 5$ local epochs per round, over $T = 100$ communication rounds. To preserve client privacy, Gaussian noise (σ^2 ranging from 0.01 to 1.0) is added to model updates for DP.

B. Experimental results

1) *Performance Evaluation*: To evaluate the effectiveness of our proposed PEARL model with DP, we compared its performance against two baseline approaches: a standard FL model without personalization and a fully centralized learning model. Table I summarizes key performance metrics, including accuracy, precision, recall, and F-score. Our personalized federated approach achieves performance closely matching the centralized model, demonstrating the effectiveness of personalized modeling even under privacy constraints. Furthermore, our approach significantly outperforms standard federated learning without personalization, underscoring the importance of personalization in federated scenarios with heterogeneous network traffic.

TABLE I: Performance Comparison Among Different Approaches

Method	Accuracy	Precision	Recall	F-score
PEARL with DP	0.935	0.932	0.929	0.931
Standard FL	0.881	0.874	0.862	0.865
Centralized Learning	0.957	0.953	0.950	0.952

Additionally, Figure 1 shows a detailed confusion matrix obtained by our PEARL model with DP, providing deeper insight into class-level performance. The confusion matrix highlights strong predictive accuracy across multiple encrypted traffic classes, such as VPN_Email, VPN_Streaming, VPN_P2P, and Tor, with most diagonal values significantly higher than off-diagonal ones, indicating accurate classification across the majority of classes. However, minor misclassifications are observed mainly among classes sharing similar network behavior patterns. For example, some VPN_Email traffic is occasionally confused with Skype or SSH, likely due to similarities in their flow characteristics and metadata profiles.

Interestingly, the class VPN_VoIP exhibits somewhat lower recall, reflecting higher misclassification rates with classes such as Skype and VPN_Streaming. Such misclassifications are plausible due to their shared reliance on real-time packet transmission and similar traffic patterns, making these classes inherently more challenging to distinguish based solely on encrypted metadata.

Overall, these results demonstrate that PEARL effectively addresses data heterogeneity and privacy constraints, achieving robust classification performance across diverse encrypted network traffic classes. The minor class confusions highlight avenues for further improvement, potentially by incorporating additional features or fine-tuning the personalized components of the model to enhance performance on challenging classes.

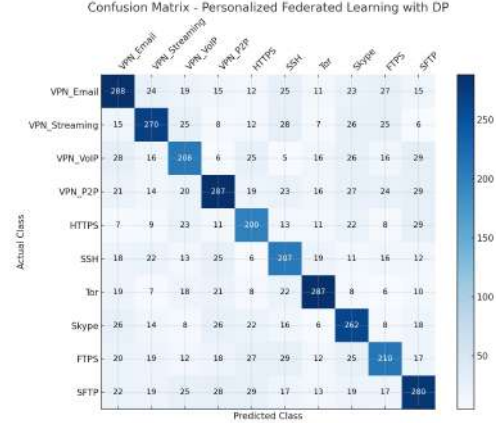


Fig. 1: Confusion Matrix of PEARL with DP

2) *Privacy Analysis*: To quantify the impact of DP noise variance (σ^2) on model performance, we systematically varied the noise parameter and analyzed its influence on classification accuracy. Figure 2 illustrates the observed trade-off between privacy (controlled via DP noise variance) and the corresponding accuracy of our PEARL model. As expected, increasing the DP noise variance strengthens privacy protection by obscuring individual clients' contributions, thereby reducing susceptibility to inference attacks. However, this privacy enhancement inevitably introduces additional randomness, affecting the precision of model parameter estimation and leading to a decline in accuracy.

Specifically, the analysis shows a clear, gradual decline in accuracy as the noise variance increases. When the noise variance (σ^2) is minimal (e.g., around 0.01 to 0.1), accuracy remains high (approximately between 94% and 92%), indicating robust performance while still ensuring moderate privacy protection. As the noise variance grows further (e.g., from 0.3 to 0.6), we observe a more pronounced drop in accuracy, which stabilizes around 88% to 90%. Beyond this region (noise variance greater than 0.6), accuracy continues to decrease, reaching approximately 82% at the highest evaluated noise variance of 1.0.

This behavior highlights a critical operational trade-off: while higher noise variance significantly improves privacy protection, the corresponding accuracy degradation can negatively impact practical network management and security monitoring tasks. Therefore, in real-world network deployments, selecting an appropriate DP noise parameter should carefully balance these competing requirements. For instance, network administrators may prefer modest noise levels (σ^2 between 0.1 and 0.4) that provide substantial privacy protection without excessively

compromising accuracy, thus maintaining effective network security and QoS.

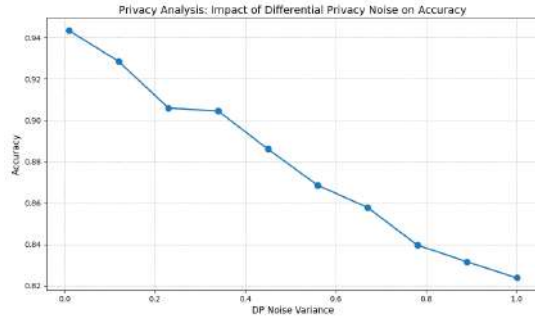


Fig. 2: Privacy analysis illustrating the trade-off between accuracy and DP noise variance.

3) *Personalization Effectiveness*: To assess the impact of personalization in federated encrypted traffic classification, we compared the accuracy of our PEARL models against a global federated model without personalization. Figure 3 provides a detailed comparison across 10 distinct federated clients, each representing different network domains with unique data distributions and traffic characteristics.

The results clearly demonstrate that personalized models consistently outperform the global federated model across all clients, emphasizing the critical role of personalization when addressing data heterogeneity in distributed network environments. Specifically, accuracy achieved by the personalized models ranges between approximately 85% to over 94%, compared to global federated models, which achieve notably lower accuracy between approximately 76% and 86%.

Particularly notable is the performance improvement on clients with distinct or more specialized traffic patterns (e.g., Client_5 and Client_10), where personalized models exhibit accuracy improvements exceeding 10% compared to global models. These significant performance gains highlight the personalized model's capability to adapt effectively to local data distributions and unique encrypted traffic characteristics, thereby substantially improving the classification quality at individual network domains.

Overall, the observed accuracy improvements provided by personalization not only enhance individual client performance but also have substantial implications for overall network management, security enforcement, and QoS assurance. The results reinforce the importance and practical value of adopting personalized approaches within federated learning frameworks, particularly in diverse and dynamic telecommunications environments.

4) *Ablation Studies*: We performed extensive ablation studies to explicitly evaluate and isolate the contributions of personalization and DP mechanisms to the overall classification performance. Table II presents accuracy and F-score metrics for four distinct model configurations: PEARL with DP, PEARL without DP, Standard FL with DP, and Standard FL without DP.

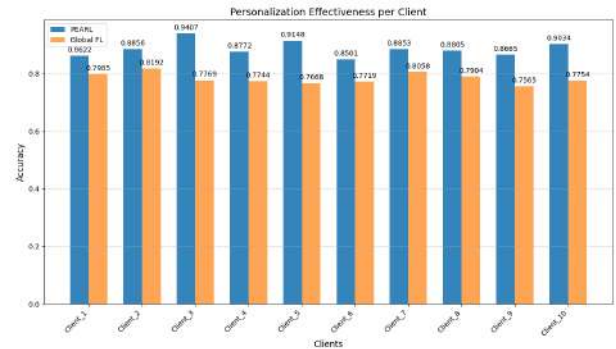


Fig. 3: Comparison of personalized and global model accuracy per client, highlighting the effectiveness of personalization in addressing data heterogeneity.

Results demonstrate that personalization consistently provides significant performance improvements compared to standard federated learning. Specifically, PEARL without DP achieves the highest accuracy (94.1%) and F-score (93.8%), underscoring the value of tailoring model components to individual network domains' unique data distributions. Standard FL, both with and without DP, shows noticeably lower performance (accuracy around 87-88%), reflecting the challenges posed by data heterogeneity in federated scenarios when personalization is not employed.

Integrating DP introduces a moderate reduction in accuracy across both personalized and standard federated learning methods. For example, incorporating DP into the personalized FL framework slightly decreases accuracy from 94.1% to 93.5%, a modest drop in exchange for significantly enhanced privacy protection. This slight performance compromise highlights an acceptable and practical privacy-accuracy trade-off, demonstrating that our PEARL model with DP maintains robust performance while providing critical privacy assurances required in real-world telecommunications networks.

Overall, this ablation analysis confirms the critical importance of personalization to manage data heterogeneity effectively and reveals the moderate yet manageable impact of DP on model performance. These insights validate our combined PFL approach with DP as a highly practical and secure solution, capable of reliably addressing privacy and personalization requirements in modern encrypted network traffic classification scenarios.

TABLE II: Ablation Study on Personalization and DP Mechanisms

Method	Accuracy	F-score
PEARL with DP	0.935	0.931
PEARL without DP	0.941	0.938
Standard FL with DP	0.872	0.865
Standard FL without DP	0.881	0.875

5) *Computational Efficiency*: We evaluated the computational overhead introduced by personalization and DP by measuring the average training time per communication round

across different model configurations. Figure 4 illustrates the comparative results for four setups: PFL with and without DP, and Standard FL with and without DP.

Interestingly, while personalization introduces additional computational complexity at the client level due to maintaining distinct classification heads, the overall training time for PFL with DP remains relatively low, averaging approximately 164.4 seconds per round. In contrast, PFL without DP incurs a higher overhead of about 245.3 seconds per round, likely attributed to additional client-side optimization steps in the absence of DP-induced noise stabilization.

Standard FL models, particularly those with DP, require similar training times (approximately 248.2 seconds per round), demonstrating that the shared global model's optimization overhead remains significant in non-personalized setups. The lowest training time (146.7 seconds per round) is observed in Standard FL without DP, due to the absence of both personalization and privacy-preserving mechanisms.

These findings highlight a notable trade-off between privacy, personalization, and computational efficiency. Despite moderate overheads, the training time for PEARL with DP remains practical for real-world deployment, especially when considering the privacy guarantees and performance benefits achieved. Furthermore, the relatively lower training time compared to other configurations suggests that integrating DP with personalization introduces stabilization effects, reducing convergence time in federated settings.

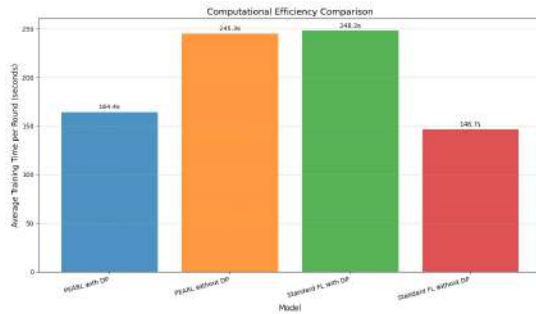


Fig. 4: Average training time per communication round across different federated learning configurations.

V. CONCLUSION

In this paper, we proposed PEARL, a Personalized Encrypted Traffic classification with Differential privacy-enhanced Learning framework enhanced with DP for encrypted traffic classification. The proposal effectively addresses the challenges posed by privacy attacks inherent in FL, ensuring data confidentiality and personalized performance across heterogeneous network environments. Experimental results confirmed that integrating DP with personalized federated models significantly improves privacy protection without compromising classification accuracy. Future work will explore further optimizations in differential privacy mechanisms and extend evaluations to real-time traffic scenarios.

ACKNOWLEDGMENT

This work is supported by the Ministry of Science and Technology of Vietnam (MOST) under Grant No. KC.01.04/21-30.

REFERENCES

- [1] B. Anderson, S. Paul, and D. McGrew, "Identifying encrypted malware traffic with contextual flow data," *Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec)*, p. 35–46, 2018.
- [2] M. Z. Shafiq, A. Khakpour, Z. Anwar, and A. X. Liu, "Encrypted traffic classification using deep learning: A systematic study," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, p. 2276–2290, 2020.
- [3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, p. 1273–1282.
- [4] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," in *Proceedings of the Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning on the Phone and other Consumer Devices*, 2017.
- [5] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2019, p. 691–706.
- [6] R. Alshammari and A. N. Zincir-Heywood, "Can encrypted traffic be identified without port numbers, ip addresses and payload inspection?" *Computer Networks*, vol. 55, no. 6, pp. 1326–1350, 2011.
- [7] W. Wang, M. Zhu, J. Wang, and X. Zeng, "Malware traffic classification using convolutional neural network for representation learning," *International Conference on Information Networking (ICOIN)*, pp. 712–717, 2017.
- [8] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [9] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, "Diot: A federated self-learning anomaly detection system for iot," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 756–767.
- [10] Y. Zhao, J. Zhao, L. Jiang, R. Tan, D. Niyato, Z. Li, L. Lyu, and Y. Liu, "Mobile edge computing, blockchain and reputation-based crowdsourcing iot federated learning: A secure, decentralized and privacy-preserving system," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2388–2401, 2021.
- [11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [12] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *NeurIPS*, pp. 3557–3568, 2020.
- [13] Y. Jiang, J. Konecny, K. Rush, S. Kannan *et al.*, "Improving federated learning personalization via model agnostic meta learning," *arXiv preprint arXiv:1909.12488*, 2019.
- [14] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4427–4437.
- [15] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, S. Choudhary *et al.*, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.
- [16] M. Abadi, A. Chu, I. Goodfellow, B. McMahan *et al.*, "Deep learning with differential privacy," in *Proceedings of CCS*, 2016, pp. 308–318.
- [17] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Ldp-fed: Federated learning with local differential privacy," *Proceedings of the ACM International Conference on Artificial Intelligence in Information and Communication (ICAIC)*, pp. 61–68, 2020.

Predicting Portfolio Price Crashes in the Vietnamese Stock Market Using Large Language Models and Deep Learning Models with Temporal Relational Reasoning

Minh-Huong Nguyen¹, Thanh-Tam Do Thi², Hong-Viet Tran^{1*}, Van-Tan Bui³, Phuong-Thai Nguyen¹

¹*University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam*

²*Faculty of Economic Information Systems and E-Commerce, Thuongmai University, Hanoi, Vietnam*

³*Faculty of Information Technology, University of Economics - Technology for Industries, Hanoi, Vietnam*

nguyenminhhuong1222@gmail.com, tam.dtt@tmu.edu.vn, thviet@vnu.edu.vn, bvtan@uneti.edu.vn, thainp@vnu.edu.vn

* *The Corresponding Author*

Abstract—Predicting stock portfolio crashes in Vietnam is challenging due to dynamic financial news. Large language models (LLMs) offer promise via zero-shot reasoning. This study introduces Temporal Personalized Relational Reasoning (TPRR), an enhanced framework building on Temporal Relational Reasoning (TRR), to model human-like cognitive processes for crash detection. TPRR integrates the Temporal Personalized PageRank (TPP) algorithm for improved graph weighting through temporal decay and personalized prioritization. We also propose a hybrid architecture combining deep learning for entity extraction and sentiment analysis with LLMs for relationship expansion, optimizing efficiency and robustness. Experiments on Vietnamese financial news (Jan 1–Apr 7, 2025) show that TPRR with deep learning achieves an AUROC of 0.781, outperforming baseline TRR (0.738). This work underscores TPRR’s potential for complex financial forecasting tasks in Vietnam.

Index Terms—temporal relational reasoning, portfolio crash prediction, large language models, deep learning, personalized pagerank

I. INTRODUCTION

In Vietnamese stock investment, diversified portfolios are used to mitigate risks across sectors such as technology, banking, and real estate. Nevertheless, rare events—like the 2008 financial crisis, COVID-19 in 2020, or the 2025 tariff crisis—can cause portfolio crashes, defined as a sudden drop in portfolio return of 2% or more [1]. Detecting such crashes remains challenging due to data scarcity and imbalance; in our 45-day evaluation (Feb 5–Apr 8, 2025), only three crash days occurred.

Large Language Models (LLMs) [2] show promise for this task thanks to their zero-shot reasoning over Vietnamese financial news. Detecting crashes requires systems that: (1) dynamically process continuous news, (2) analyze relational networks between events and stocks, and (3) account for temporal dynamics. Existing frameworks like Graph-of-Thoughts (GoT) [3] and Think-on-Graph (ToG) [4] do not fully satisfy these needs.

We propose the Temporal Relational Reasoning (TRR) framework, which emulates cognitive processes through: (1)

inferring entities and relations from news, (2) chaining information with temporal dependencies, (3) filtering critical “impact chains” using Temporal Personalized PageRank (TPP), and (4) reasoning over the temporal-relational graph to predict crashes. To optimize performance, TRR is combined with a hybrid architecture: a pre-trained NER model and a sentiment classifier extract entities and initial impacts, while an LLM (Gemini 2.0 Flash) identifies complex relationships, iteratively expanding the graph efficiently.

We further extend TRR to Temporal Personalized Relational Reasoning (TPRR) by integrating TPP into the Graph-based Temporal Relational Reasoning (GTRR) module, enhancing filtering of critical impacts in temporal graphs. Using Vietnamese financial news from Jan 1–Apr 7, 2025, we evaluate this framework on ten representative stocks (FPT, SSI, VCB, VHM, HPG, GAS, MSN, MWG, GVR, VCG) across diverse sectors.

II. RELATED WORKS

A. Stock Price Prediction

Early research represented stock data as Subject-Relation-Object triples [5]–[7], later shifting to Knowledge Graph embeddings [8]–[10] from static databases like Freebase [11] or Wikidata [12]. These methods largely ignore dynamic news-driven changes. Recent works [13] leverage LLMs to infer relations from news combined with Graph Neural Networks (GNNs). In contrast, our framework focuses on zero-shot reasoning to detect portfolio crashes from emerging events.

B. Reasoning Frameworks Integrated with LLMs

LLMs show strong zero-shot reasoning [14]. Frameworks like Tree-of-Thoughts (ToT) [15] and GoT [3] emulate cognitive problem-solving. Integrating LLMs with KGs, as in StructGPT [16] or ToG [4], improves reliability, but mainly extracts chains from static KGs. Our approach extracts subgraphs from dynamic impact graphs to predict portfolio crashes.

C. Temporal Relational Reasoning (TRR)

TRR [1] simulates human cognition by processing temporal-relational information in financial news, using a decaying memory bank, PageRank filtering, and reasoning over the Temporal-Relational Graph. Our TPRR extension integrates TPP to improve temporal graph weighting, enhancing LLM reasoning. TRR operates in four stages:

- **Brainstorming:** Generate graphs of cascading impacts.
- **Memory:** Retrieve historical impact chains.
- **Attention:** Extract critical impact chains.
- **Reasoning:** Analyze subgraphs to predict crashes.

III. RESEARCH METHOD

The main problem of the study is modeled as follows: given a set J of news articles $X = \{x_1, x_2, \dots, x_J\}$, to provide a binary prediction regarding whether a portfolio P will crash on the subsequent day, i.e., $\hat{y} \in \{0, 1\}$. The portfolio P comprises N stocks, $P = \{s_1, s_2, \dots, s_N\}$, where s_i denotes a specific stock and $i \in \{1, 2, \dots, N\}$. Applying the four stages of TRR [1] to the research problem:

Brainstorming Phase: This phase aims to capture the comprehensive impact of each news event on the selected portfolio. It first identifies all potential propagation chains leading to the stocks within the portfolio. This process is modeled as a directed graph $G = (Z, A)$. The set of vertices Z comprises the initial news articles x_j with $j \in \{1, \dots, J\}$, intermediate affected entities e_h with $h \in \{1, \dots, H\}$, and culminates at the stocks in the portfolio s_n with $n \in \{1, \dots, N\}$. The set of edges A denotes the directed positive/negative impact between vertices, each associated with a timestamp representing when the relationship was formed.

To generate new vertices for the graph, we iteratively prompt Large Language Models (LLMs), specifically models from the Google Gemini family, in two main steps:

- Initial Entity Extraction:** From each news article (x_j), the LLM identifies directly affected entities (stocks, sectors, companies, etc.) and their impact polarity (positive/negative), linking up to 5 entities per article.
- Iterative Relationship Expansion (Chain Effect):** Entities identified in the previous step (the "frontier") are continually fed into the LLM to discover new entities influenced by them, expanding indirect relationships and chain effects, with up to 4 entities per call.

Each individual impact chain is denoted as C , where $C \in G$. Each chain C originates from a news article x_j , traverses through affected entities, and concludes at a stock s_n . This process continues until reaching a portfolio stock or the iteration limit. Duplicate entities are merged into a single vertex to maintain graph consistency.

Memory Phase: Human memory of news events tends to decay over time; thus, past news is presumed to have less impact on the current market. This fading is modeled by an exponential decay applied to the edge weights:

$$w(u, v) = \exp(-\lambda \cdot t_{u,v}) \quad (1)$$

where $t_{u,v}$ is the time difference (in days) between the occurrence of impact (u, v) and the prediction day, and λ is the decay rate constant. Following [1], $\lambda = 1$ is used by default but can be tuned.

The memory module operates daily in two steps:

- For each affected entity generated by the LLM, the module searches for prior mentions and incorporates them into the daily graph G , creating a temporal context graph:

$$G_{\text{temporal}} = \bigcup_{\forall e_h \in Z} (G \cup M_{e_h}) \quad (2)$$

- At the end of each day, the memory module is updated with the daily impact chains, i.e., $M_{e_h} = \bigcup_{\forall e_h \in Z} (M_{e_h} \cup C_{e_h})$, storing temporal context for future phases.

This mechanism enables TRR to combine current and past information in a unified temporal graph while modeling memory decay consistent with previous research [1].

Attention Phase: The temporal context graph G_{temporal} is often too large for direct LLM processing due to token limits. While other studies aggregate or prune paths [17], we preserve an informative relational graph (see Figure 1) to provide a comprehensive market overview.

Inspired by the PageRank algorithm [7], we assign importance scores to entities while incorporating temporal decay:

$$PR(e_h) = \sum_{b \in B_{e_h}} \frac{PR(b)}{L_b} \cdot w(b, e_h) \quad (3)$$

where B_{e_h} is the set of parent nodes of entity e_h , L_b is the out-degree of node b , and $w(b, e_h)$ is the decayed edge weight from Eq. 1.

Entities are ranked by $PR(e_h)$, and the top- q highest-scoring entities are retained (default $q = 6$, as in [1]). Chains C containing these entities form the reduced subgraph G_{TRR} , representing the most critical daily information.

Reasoning Phase: Finally, to determine whether a portfolio crash will occur, we reason based on the temporal-relational subgraph G_{TRR} . The subgraph is represented as relational tuples:

$$(t, z_s, a, z_o)$$

where t is the timestamp, z_s and z_o are subject and object entities, and a is the direction of impact.

Building on prior works [9], [18], we employ an LLM to reason over G_{TRR} represented as relational tuples [19]. The prediction is generated as:

$$\hat{y} \sim p_{\text{reason}}(\hat{y} \mid P, G_{TRR}) \quad (4)$$

where $\hat{y}$ is a binary outcome (0: no crash, 1: crash), simulating how investors evaluate both direct and indirect news impacts on their portfolios.

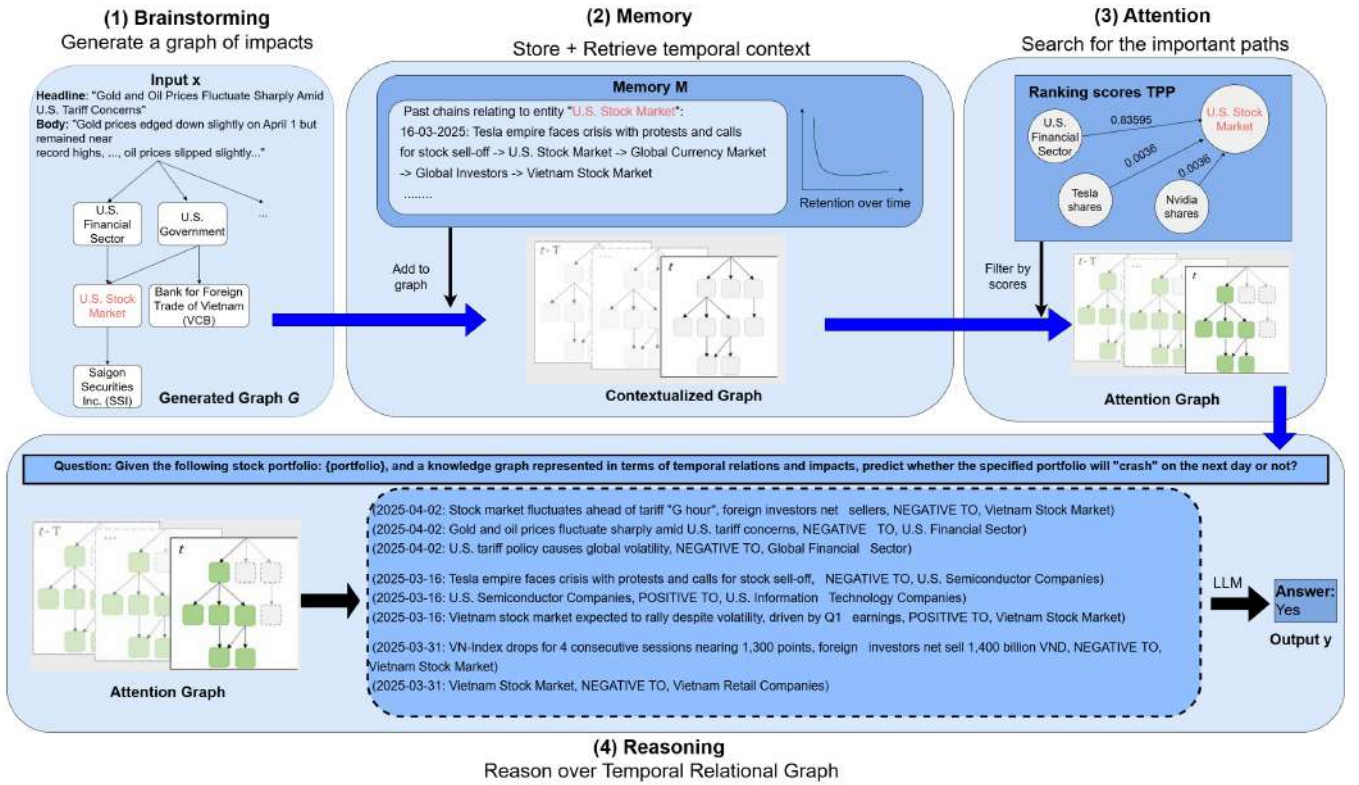


Fig. 1. Components of the improved TPRR framework using the Temporal Personalized PageRank (TPP) algorithm in step (3), simulating human cognitive processes including (1) brainstorming, (2) memory, (3) attention, and (4) reasoning.

IV. PROPOSED METHOD

In this study, we propose an effective hybrid approach to construct a knowledge graph from financial news data, combining the strengths of both specialized Deep Learning (DL) models and Large Language Models (LLMs). This method aims to optimize processing performance and operational costs while ensuring the quality of entity and relationship extraction. In addition, in the attention phase of the TRR framework [1], we apply the Temporal Personalized PageRank algorithm (TPP)—an improved variant of the classic PageRank algorithm [20].¹

A. A Hybrid Knowledge Graph Building

The proposed architecture operates in two main stages: **Stage 1: Entity Extraction and Initial Impact Identification.** To achieve high efficiency and reduce the load on LLMs, we prioritize the use of specialized DL models for foundational tasks. Specifically, a pre-trained Named Entity Recognition (NER) model (NlPHUST/ner-vietnamese-electra-base) is employed to identify and extract relevant entities from news article titles and descriptions. Subsequently, a sentiment classification model (wonrax/phobert-base-vietnamese-sentiment) is applied to determine the initial impact (positive, negative, or neutral) of the article content on the extracted

entities. It is important to note that sentiment analysis here only provides a coarse polarity score, whereas relationship expansion in Stage 2 captures richer semantic and causal links. An LLM fallback mechanism is also implemented: in cases where DL models fail to extract entities (empty output) or assign a confident impact, the system defaults to using an LLM (Gemini 2.0 Flash). This ensures robustness and completeness.

Stage 2: Relationship Expansion. To capture complex relationships and deeper contextual understanding between entities, we rely entirely on an LLM. The LLM (Gemini 2.0 Flash) is used to analyze the entities extracted in Stage 1 and identify diverse relationships between them (e.g., "Company A positively impacts Industry B"). This step goes beyond polarity, enabling the construction of causal and sectoral impact chains. Newly discovered target entities are added to the "frontier," and the relationship extraction process is iterated a specified number of times, allowing for deeper exploration of connections within the information network. This hybrid architectural approach offers several significant benefits:

- **High Efficiency:** By leveraging the fast processing speed of DL models for simpler tasks, we significantly minimize the number and frequency of LLM calls, leading to faster data processing. Processing 20 news articles per day is completed in approximately 15 minutes, compared to 30+ minutes with LLM-only setups.
- **Cost Optimization:** Reducing reliance on LLMs for

¹<https://github.com/nguyenminhhuong0208/Temporal-Personalized-Relation-Reasoning.git>

basic tasks substantially lowers API costs associated with utilizing large language models.

- **Accuracy and Robustness:** The combination of specialized models and an intelligent fallback mechanism ensures high accuracy for each task and resilience against atypical data cases or model failures.

B. Temporal Personalized Relational Reasoning (TPRR)

The Temporal Personalized Relational Reasoning (TPRR) framework integrates the Temporal Personalized PageRank (TPP) algorithm, an extension of the classic PageRank by Brin and Page [20], by incorporating personalization and temporal factors to enhance its capability in analyzing financial networks. Unlike traditional PageRank, where scores are uniformly distributed among nodes, TPP utilizes a personalization vector $p(n)$ to prioritize nodes directly related to the portfolio and those with direct influence on them. This personalization vector $p(n)$ is defined as follows:

- $p(n) = \text{WEIGHT\_DIRECT\_INFLUENCER}$ (default = 5) for nodes that are direct influencers (neighbors of stocks and sectors in the portfolio).
- $p(n) = \text{WEIGHT\_PORTFOLIO\_ITEM}$ (default = 0.5) for nodes belonging to PORTFOLIO_STOCKS or PORTFOLIO_SECTOR.
- $p(n) = \text{WEIGHT\_OTHER\_NODES}$ (default = 0.1) for all remaining nodes.

Subsequently, the personalization vector $p(n)$ is normalized so that its values sum to 1 before being used in the TPP algorithm. Temporal weights $w(u, v)$ are also applied to reflect the decay of information value over time. Two components are considered:

- **Node-to-node decay (R_{decay}):** Based on the time difference between source and destination nodes, calculated as $e^{-\lambda_{\text{decay}} \cdot \Delta t_{\text{nodes}}}$.
- **Edge age decay:** Based on the time since the edge was formed, calculated as $e^{-\lambda_{\text{decay}} \cdot \Delta t_{\text{edge}}}$.

The final edge weight is obtained by multiplying both decay factors. By default $\lambda_{\text{decay}} = 1$, but it is a tunable hyperparameter, consistent with prior work [1].

The TPP score update is formally expressed as:

$$PR(n) = (1 - \alpha) \cdot p(n) + \alpha \sum_{(u,v) \in E} \frac{PR(u) \cdot w(u, v)}{d_u} \quad (5)$$

where $\alpha = 0.85$ is the damping factor, d_u is the out-degree of node u , $w(u, v)$ is the temporal weight, and $p(n)$ is the personalization prior. Here, both λ_{decay} and q (the number of top entities retained) are tunable hyperparameters, following the TRR framework, and can be adjusted to balance between recall and precision of critical entities.

V. EXPERIMENT

Experimental Setup: From the FireAnt platform, a forum focused on finance and securities in Vietnam, the research team collected 24,000 articles over the period from January 1, 2025, to April 7, 2025. The gemini-2.0-flash lite model was

Algorithm TPP($G, \alpha, p, \lambda_{\text{decay}}, t_{\text{current}}, \text{max\_iters}$)

Require: G is directed graph with nodes and edge/node timestamps, α is damping factor, p is personalization vector, λ_{decay} is decay coefficient, t_{current} is current timestamp, max_iters : max iterations

Ensure: PR : PageRank score for each node

```

1: procedure TPP( $G, \alpha, p, \lambda_{\text{decay}}, t_{\text{current}}, \text{max\_iters}$ )
2:    $\text{sum\_p} \leftarrow \sum_{n \in G.\text{nodes}()} p[n]$ 
3:   for each node  $n$  in  $G.\text{nodes}()$  do
4:      $p[n] \leftarrow p[n] / \text{sum\_p}$ 
5:   end for
6:   for each node  $n$  in  $G.\text{nodes}()$  do
7:      $PR[n] \leftarrow p[n]$ 
8:   end for
9:   for each edge  $(u, v)$  in  $G.\text{edges}()$  do
10:     $w(u, v) \leftarrow$  calculated as per Equation (2)
11:     $G[u][v][\text{"temporal\_weight"}] \leftarrow w(u, v)$ 
12:   end for
13:   for  $i$  from 1 to  $\text{max\_iters}$  do
14:      $\text{new\_PR} \leftarrow \{\}$ 
15:     for each node  $n$  in  $G.\text{nodes}()$  do
16:        $\text{sum\_incoming} \leftarrow 0$ 
17:       for each predecessor  $u$  of  $n$  in  $G$  do
18:          $d_u \leftarrow \text{out\_degree}(u)$ 
19:          $\text{sum\_incoming} \leftarrow \text{sum\_incoming} + (PR[u] \cdot$ 
20:            $G[u][n][\text{"temporal\_weight"}] / d_u)$ 
21:       end for
22:        $\text{new\_PR}[n] \leftarrow (1 - \alpha) \cdot p[n] + \alpha \cdot \text{sum\_incoming}$ 
23:     end for
24:      $PR \leftarrow \text{new\_PR}$ 
25:     if convergence criterion met or  $i = \text{max\_iters}$  then
26:       BREAK
27:     end if
28:   end for
29:   return  $PR$ 
30: end procedure

```

employed to summarize the articles and select 20 significant and prominent articles each day.² Additionally, both the gemini 2.0-flash-lite and gemini-2.5-flash models were utilized for most tasks involving brainstorming, memory, attention, and reasoning, with a temperature parameter of 0.1 to ensure stable and accurate responses. Predictions were conducted over 45 trading days (February 5, 2025, to April 8, 2025) within a total collection period of 62 days. To evaluate daily price changes for the 10 stock codes (FPT, SSI, VCB, VHM, HPG, GAS, MSN, MWG, GVR, VCG), we computed the equal-weighted portfolio return (each stock treated equally). A day is labeled as a “crash” if the equal-weighted average decline exceeds 2% compared to the previous day’s close. Within this evaluation window, three crash days were observed: 2025-04-03, 2025-04-04, and 2025-04-08

Implementation: We first collected all raw posts/articles, then used Gemini (gemini 2.0-flash-lite) to produce a short

²Dataset: https://huggingface.co/datasets/nguyenminhhuong0208/FireAnt_financial_news

summary for each article and a daily ranking. From the daily pool (24,000 total across the collection period) we used an LLM prompt to select 20 representative summaries per day (the prompt asks the model to produce 20 concise salient news items for that day). Where the DL NER or sentiment modules failed (empty extraction or low-confidence sentiment), an LLM fallback was used to ensure entity + impact coverage. Average observed LLM latency in our environment for a single prompt (with our templates and token sizes) was 15 seconds. It is important to note that the DL-based sentiment model only provides polarity signals, while the LLM stage is responsible for capturing richer relationships and causal structures. Thus, the hybrid pipeline leverages DL for efficiency but still relies on LLMs for semantic depth

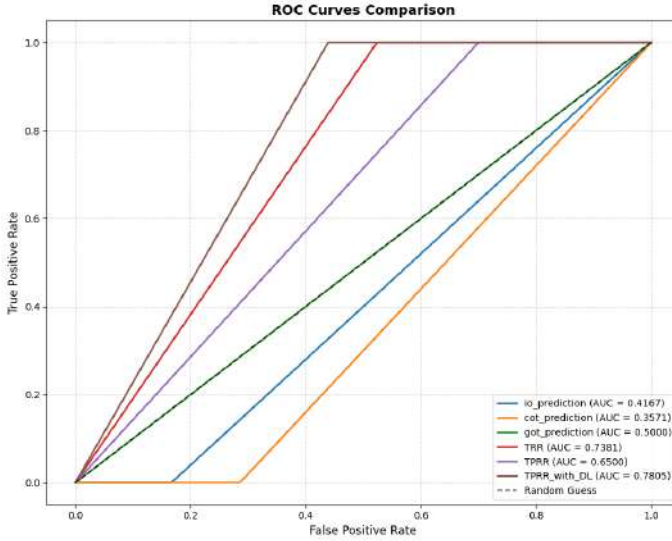


Fig. 2. ROC Curves of the Tested Models. Hybrid TPRR achieves the highest AUROC.

Result: Table I summarizes the performance of six methods evaluated in the experiments: Input-Output (IO) prompting, Chain-of-Thoughts (CoT), Graph-of-Thoughts (GoT), Temporal Relational Reasoning (TRR), Temporal Personalized Relational Reasoning (TPRR), and TPRR with Deep Learning (TPRR with DL). TRR is used as the baseline. Note that the three simple prompting baselines (IO, CoT, GoT) did not return reliable binary crash labels in our setup (hence missing recall/F1 entries); they were primarily included to show that structured graph-based reasoning outperforms naive prompting for this task.

TABLE I
EXPERIMENTAL RESULTS ACROSS MODELS.

Model	AUC	Recall	F1-score
IO	0.417	—	—
CoT	0.357	—	—
GoT	0.500	—	—
TRR	0.738	1.000	0.214
TPRR	0.650	1.000	0.177
TPRR with DL	0.781	1.000	0.250

TABLE II
CONFUSION MATRIX COMPARISON

Metric	TRR	TPRR	TPRR with DL
TP	3	3	3
FP	22	28	18
TN	20	12	23
FN	0	0	0

From Table II, the TPRR with DL model shows superior performance in detecting crash events, achieving True Positives (TP) = 3 and False Negatives (FN) = 0, indicating perfect recall ($\text{Recall} = 1.0$). However, it recorded False Positives (FP) = 18, resulting in a low precision ($\text{Precision} = \frac{3}{3+18} \approx 0.1429$), while True Negatives (TN) = 23 yield an overall accuracy of $\frac{3+23}{44} \approx 0.5909$. Interestingly, the TPRR model without Deep Learning (DL) preprocessing underperforms the baseline TRR in AUROC, mainly because its additional hyperparameters—personalization weights, λ_{decay} , and entity cut-off q —were left at default values ($\lambda_{\text{decay}} = 1$, $q = 6$) without tuning. As discussed in [1], these parameters strongly affect the precision–recall balance. In contrast, when combined with DL preprocessing, the hybrid TPRR with DL achieves the highest AUROC (0.781), outperforming TRR by 4.3%.

VI. DISCUSSION AND FUTURE WORK

While the Temporal Personalized Relational Reasoning (TPRR) framework demonstrates potential in predicting stock portfolio crashes, several limitations remain. Firstly, although recall is perfect (1.000), the F1-score is relatively low (0.250), reflecting high false positive rates. Future work should therefore focus on improving precision—by refining attention mechanisms, filtering chains more selectively, or constraining LLM reasoning prompts. Secondly, the dataset covers only Vietnamese financial news from January to April 2025, a short period with limited crash events (3 in 45 trading days), raising concerns about generalizability. Expanding both the temporal span and data source diversity will improve robustness. Thirdly, although the hybrid design reduces LLM reliance, the relationship expansion stage still depends on stochastic LLM outputs. Even with a low temperature (0.1), variability can cause non-deterministic graph construction; thus, constrained decoding, rule-based post-processing, or ensemble prompting should be explored. Finally, hyperparameter optimization was not performed due to time constraints. Parameters such as the temporal decay coefficient λ , the number of top-ranked entities q , personalization weights $p(n)$, and damping factor α were set to heuristic defaults, likely explaining why TPRR underperformed TRR in isolation. In future work, we will systematically tune and analyze the sensitivity of these hyperparameters to identify optimal settings, improving precision and clarifying how personalization and temporal weighting interact in financial reasoning.

Key Contributions: Despite these limitations, our work makes three main contributions:

- Extends Temporal Relational Reasoning (TRR) into Temporal Personalized Relational Reasoning (TPRR), introducing personalization and two-component temporal decay for PageRank-based filtering of financial news graphs.
- Designs a hybrid pipeline combining DL-based NER and sentiment models with LLM-based relation expansion (with fallback), improving efficiency and robustness while keeping inference costs practical.
- Provides an empirical evaluation on Vietnamese financial news for portfolio crash detection and releases code, dataset, and prompt templates for reproducibility.

VII. CONCLUSION

This study proposed the Temporal Personalized Relational Reasoning (TPRR) framework, enhancing Temporal Relational Reasoning (TRR) for predicting stock portfolio crashes in the Vietnamese stock market. By integrating the Temporal Personalized PageRank (TPP) algorithm, TPRR incorporates both temporal decay and personalized prioritization to identify critical impact chains. A hybrid architecture improves efficiency by combining deep learning models for entity extraction and sentiment analysis with large language models for relationship expansion. Experiments on Vietnamese financial news (data collected Jan 1–Apr 7, 2025; predictions made Feb 5–Apr 8, 2025) show that TPRR with deep learning achieved an AUROC of 0.781, improving by 4.3% over TRR (0.738), with perfect recall (1.000) and higher F1-score (0.250). These results highlight the promise of combining personalization, temporal reasoning, and hybrid pipelines in financial forecasting. Future directions include expanding datasets to longer time spans and more diverse sources, systematically tuning λ_{decay} and q , and applying TPRR to broader macroeconomic forecasting tasks. This will improve precision, scalability, and the real-world applicability of crash prediction systems.

REFERENCES

- [1] K. J. L. Koa, Y. Ma, R. Ng, H. Zheng, and T.-S. Chua, "Temporal relational reasoning of large language models for detecting stock portfolio crashes," 2024. [Online]. Available: <https://arxiv.org/abs/2410.17266>
- [2] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2307.06435>
- [3] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, and T. Hoefler, "Graph of thoughts: Solving elaborate problems with large language models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, p. 17682–17690, Mar. 2024. [Online]. Available: <http://dx.doi.org/10.1609/aaai.v38i16.29720>
- [4] S. Ma, C. Xu, X. Jiang, M. Li, H. Qu, C. Yang, J. Mao, and J. Guo, "Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation," 2025. [Online]. Available: <https://arxiv.org/abs/2407.10805>
- [5] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Using structured events to predict stock price movement: An empirical investigation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1415–1425. [Online]. Available: <https://aclanthology.org/D14-1148/>
- [6] X. Ding, Y. Zhang, T. Liu, and J. Duan, "deep learning for event-driven stock prediction," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, 2015, p. 2327–2333.
- [7] Y. Zhang and S. Clark, "Syntactic processing using the generalized perceptron and beam search," *Computational Linguistics*, vol. 37, no. 1, pp. 105–151, Mar. 2011. [Online]. Available: <https://aclanthology.org/J11-1005/>
- [8] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Knowledge-driven event embedding for stock prediction," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Y. Matsumoto and R. Prasad, Eds. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2133–2142. [Online]. Available: <https://aclanthology.org/C16-1201/>
- [9] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T.-S. Chua, "Temporal relational ranking for stock prediction," *ACM Transactions on Information Systems*, vol. 37, no. 2, pp. 1–30, mar 2019. [Online]. Available: <http://dx.doi.org/10.1145/3309547>
- [10] R. Sawhney, S. Agarwal, A. Wadhwa, and R. R. Shah, "Deep attentive learning for stock movement prediction from social media text and company correlations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 8415–8426. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.676/>
- [11] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," ser. SIGMOD '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1247–1250. [Online]. Available: <https://doi.org/10.1145/1376616.1376746>
- [12] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, p. 78–85, Sep. 2014. [Online]. Available: <https://doi.org/10.1145/2629489>
- [13] Z. Chen, L. Zheng, C. Lu, J. Yuan, and D. Zhu, "Chatgpt informed graph neural network for stock movement prediction," *SSRN Electronic Journal*, 2023. [Online]. Available: <http://dx.doi.org/10.2139/ssrn.4464002>
- [14] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," 2023. [Online]. Available: <https://arxiv.org/abs/2205.11916>
- [15] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023. [Online]. Available: <https://arxiv.org/abs/2305.10601>
- [16] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J.-R. Wen, "StructGPT: A general framework for large language model to reason over structured data," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9237–9251. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.574/>
- [17] Z. Liao, T. Liu, Y. He, and L. Lin, "Effective temporal graph learning via personalized pagerank," *Entropy*, vol. 26, no. 7, 2024. [Online]. Available: <https://www.mdpi.com/1099-4300/26/7/588>
- [18] L. Lin, P. Yuan, R.-H. Li, C. Zhu, H. Qin, H. Jin, and T. Jia, "Query-centered temporal community search via time-constrained personalized pagerank," 2023. [Online]. Available: <https://arxiv.org/abs/2302.08740>
- [19] Z. Jiang, B. Liu, M. Peng, W. Xu, Y. Xiao, Z. Shan, and M. Peng, "Towards explainable temporal reasoning in large language models: A structure-aware generative framework," *arXiv preprint arXiv:2505.15245*, 2025.
- [20] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Technical Report 1999-66, november 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: <http://ilpubs.stanford.edu:8090/422/>

A New Two-Stage Fusion-Based Approach to Cyber Threat Detection Using Graph Neural Networks

Thu Lam Bui

Academy of Cryptography Techniques
Hanoi, Vietnam
0000-0002-9812-4970

Thi Thoa Vi*

Academy of Cryptography Techniques
Hanoi, Vietnam
0009-0003-6938-3548

Cong Thanh Nguyen

Academy of Cryptography Techniques
Hanoi, Vietnam
0009-0005-4214-8091

Quang Long Pham

Academy of Cryptography Techniques
Hanoi, Vietnam
0009-0002-7960-284X

Thanh Tra Ung

International University Vietnam
National University Ho Chi Minh City
Ho Chi Minh, Vietnam
0009-0001-6724-3600

Duy Trung Pham

Academy of Cryptography Techniques
Hanoi, Vietnam
0009-0009-0986-709

Abstract—Traditional intrusion detection systems (IDS) face significant challenges in identifying sophisticated cyber attacks due to their inability to fully capture the structural and relational information embedded in network systems. This paper proposes a novel two-stage fusion approach leveraging Graph Neural Networks (GNNs) that combine neighbor sampling-based SAGEConv and Graph Convolutional Networks (GCN) in an attack graph framework to improve detection accuracy. The attack graph models the system status, including static attributes such as components, configurations, and existing vulnerabilities. Our FD-GNN model integrates two parallel branches to utilize both large-scale graph processing and local structural representation learning. Experimental results on the CICUNSW-NB15 dataset demonstrate that FD-GNN achieves a validation accuracy of 98.6 % and F1-score of 93.4 %, outperforming baseline models. The model exhibits high generalization ability and robust multi-class classification performance across various attack types.

Index Terms—GNN, Intrusion Detection System, SAGEConv, GCN, Attack Graph, Feature Fusion

I. INTRODUCTION

In the current digital era, cyberattacks have become increasingly sophisticated and diverse, posing significant challenges to the cybersecurity of organizations and critical infrastructures [1]. Intrusion Detection Systems (IDS) play a pivotal role in protecting information systems against both external and internal threats [2]. However, traditional IDS techniques—namely, signature-based and anomaly-based methods—often struggle with processing flat data representations and fail to capture the complex structures and relationships inherent in modern network environments [1], [2].

Graph Neural Networks (GNNs) have emerged as a promising solution to these limitations [3]. GNNs are capable of learning informative representations from graph-structured data, enabling models to recognize structural patterns of cyberattacks rather than relying solely on conventional features [1], [3]. Among GNN architectures, SAGEConv (a variant of Graph Sample and Aggregation) utilizes neighbor sampling to efficiently process large-scale graphs and has demonstrated excellent scalability for such environments [4], [5]. In contrast,

Graph Convolutional Networks (GCN) leverage message passing mechanisms, excelling at capturing local graph structures and the nuanced interactions between nodes [6].

To harness the benefits of both approaches, this research proposes a new architecture featuring a two-stage fusion of SAGEConv and GCN for cyber intrusion detection, Fusion Dual-Graph Neural Network (FD-GNN). The proposed method employs an attack graph to represent the system's operational state, encompassing static attributes such as components, configurations, and known vulnerabilities [7], [8]. The FD-GNN model allows the system to leverage SAGEConv's strength in neighbor sampling for large, complex topologies while simultaneously utilizing GCN's effectiveness in local structure learning and message passing. This synergistic design enables accurate and scalable intrusion detection, addressing critical challenges faced by conventional and single-model IDS solutions.

The rest of the paper is organized as follows. First, we review the related works in Section II. Section III, we presents the proposed GNN model for attack detection. Section IV presents the experimental results and analysis. We conclude the paper with conclusions drawn from the experimental results and future works in Section V.

II. RELATED WORK

Traditional intrusion detection systems can be categorized into two main groups: signature-based detection and anomaly-based detection [2]. The signature-based method uses a database of known attack patterns to match against network traffic, whereas the anomaly-based method identifies abnormal activities by comparing them with a baseline of normal behavior [9]. However, these methods face limitations in detecting novel (zero-day) attacks and cannot capture the complex structural relationships present in network data [1], [10].

With the advancement of machine learning, numerous studies have applied algorithms such as Support Vector Machine (SVM), Random Forest, and Neural Networks to the intrusion

detection problem [9]. These methods have significantly improved detection performance compared to traditional approaches, especially in handling high-dimensional data and identifying complex attack patterns [11], [12]. Nonetheless, most of these methods still process data in a flat format, neglecting the crucial structural information of network topologies [1].

Recently, Graph Neural Networks (GNNs) have been widely adopted in cybersecurity due to their capability to process graph-structured data [1], [10]. Lo et al. proposed E-GraphSAGE, the first successful application of GNN for IoT intrusion detection using flow data [4]. This study demonstrated that integrating edge features and topological information can considerably enhance detection performance.

Other researches such as GNN-IDS by Chen et al. incorporated attack graphs with real-time measurements to represent both static and dynamic attributes of computer networks [10]. Termos et al. introduced Centrality-based E-GraphSAGE, extending the E-GraphSAGE model by integrating centrality measures to improve detection performance in IoT environments [13].

Although notable progress has been made, the current literature still exhibits several limitations:

- Lack of prediction explainability: State-of-the-art deep-learning intrusion-detection models still struggle to provide interpretable outputs. In addition, most studies emphasize only one attribute type and do not effectively integrate static attributes from attack graphs with dynamic attributes extracted from network flows [14].
- Inefficient information integration: Existing methods often utilize a single type of GNN architecture, thus failing to exploit the complementary advantages offered by diverse architectures [3].
- Lack of multi-class classification capability: Many research studies focuses on binary classification, paying little attention to multi-class classification for differentiating various attack types [11], [12]. In fact, detecting a specific type of attack as early as possible helps increase response capabilities and minimize losses for businesses.
- Not optimized for real-time processing: Several methods have high computational complexity, rendering them unsuitable for real-time applications [15].

The proposed method addresses the above-mentioned limitations by combining SAGEConv and Graph Convolutional Networks (GCNs) through a dedicated feature-fusion mechanism. By jointly leveraging attack-graph information and dynamic flow data, the model not only achieves a marked improvement in intrusion-detection performance but also yields interpretable predictions. Its effectiveness is empirically validated on the CIC-UNSW-NB15 dataset, a modern benchmark for network-intrusion research.

III. PROPOSED METHOD

We propose a multi-class intrusion detection model based on a spatio-temporal graph constructed from the attack graph and real-time system parameters. The execution flow is illustrated

in Figure 1. We first introduce the key components for constructing a spatio-temporal graph from the attack graph and real-time data. Next, we introduce the proposed FD-GNN model, featuring parallel processing branches and a feature fusion layer.

A. Stage 1: Input Graph Construction with Spatio-Temporal Graph Fusion

1) *AttackGraph*: The information-gathering phase records the current system posture from multiple sources, encompassing network assets, system configurations, and known vulnerabilities. These data are harvested through vulnerability-scanning tools and monitoring platforms, thereby providing a comprehensive snapshot of the system state.

- The attack graph is then constructed using three principal node types [8], [16] that delineate the routes an adversary might leverage to compromise the system:
- Information nodes – encode network details such as assets, configurations, and vulnerabilities. They describe the static state of the environment—e.g., IP addresses, service ports, software versions, and documented CVEs.
- Action nodes – represent interaction rules and attack techniques that can be executed. Typical examples include vulnerability exploitation, privilege escalation, and lateral movement within the network.
- Privilege nodes – denote access rights and the eventual outcome of asset compromise, ranging from ordinary user privileges to administrator authority or access to sensitive data.

Relationships among these nodes follow a coherent principle: a cyber-attack commonly starts from a information node, traverses an action node, and culminates in a new privilege node with higher access rights. For instance, a low-privilege user might execute malicious code to obtain administrative control of a server.

2) *Experimental System Modeling*: We employ the MulVAL framework to model an experimental system that simulates a realistic operational environment consisting of the following components:

- Firewall Cluster: A pair of firewalls protecting the system
- DMZ Zone: Hosting a public web server
- Server Zone: Containing an internal file server
- LAN Zone: Internal network for users

The system is configured with three primary security vulnerabilities:

- CVE-2021-41773: A path traversal vulnerability in Apache HTTP Server, allowing attackers to access files outside of the configured directories [17]
- CVE-2022-24405: A command injection vulnerability in OX App Suite, enabling remote command execution [18]
- CVE-2018-15473: A user enumeration vulnerability in OpenSSH versions 7.7 and earlier [19]

The attack graph generated from this system contains **nine** selected privileged nodes, each representing a distinct compromised state. These states are mapped to **nine** attack labels in the CICUNSW-NB15 dataset.

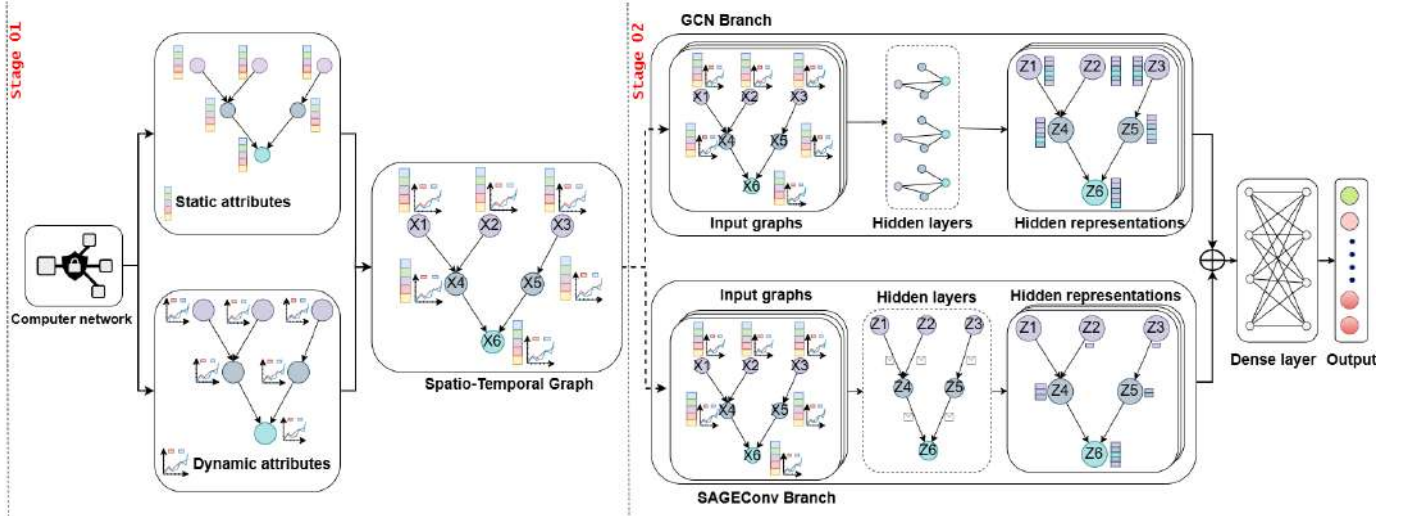


Fig. 1: Two stage of Proposed Method

3) *Input Graph Construction*: The attack graph is represented through edges and nodes, with each node characterized by a textual statement (S). Dataset entries consist of system snapshots captured at specific time points, which are subsequently transformed into graph structures. The procedure for generating the spatio-temporal graph is executed as follows:

- Static features are extracted for each node from the attack graph to form matrix $S \in \mathbb{R}^{N \times F_s}$, maintaining a fixed topology defined by edge_index.
- Real-time measurements for sample b are acquired and masked to preserve values exclusively at action nodes, with all other nodes assigned zeros, resulting in $R^{(b)} \in \mathbb{R}^{N \times F_t}$.
- Feature concatenation is performed along the dimension axis for each node: $x^{(b)} = [S; R^{(b)}]$. This process is repeated across all samples to yield batch $X \in \mathbb{R}^{B \times N \times (F_s + F_t)}$.
- Labeling scheme: Benign instances are assigned zero vectors across action nodes; malicious instances employ one-hot encoding based on the targeted action node (with measurements selectively injected solely into the target node).

B. Stage 2: Dual-Branch GNN Architecture with Feature Fusion Layer

The FD-GNN model is designed with a dual-branch parallel architecture to leverage the strengths of both SAGEConv and GCN components:

- **Branch 1 – SAGEConv with Neighbor Sampling**
 - Utilizes SAGEConv with a neighbor sampling mechanism to efficiently process large-scale graphs.
 - Supports inductive learning, enabling the model to handle unseen nodes during inference and deployment.
 - Employs sampling strategies to mitigate the neighbor explosion issue commonly encountered in Graph Neural Networks.

- **Branch 2 – GCN with Message Passing**

- Incorporates the Graph Convolutional Network to capture local structural information within the graph.
- Implements efficient message passing to aggregate information from neighboring nodes.
- Exploits the spectral properties of the graph to enhance representation and learning capacity.

This dual-branch design allows the FD-GNN model to achieve both scalability and expressive power by integrating inductive learning with rich structural representation capabilities.

1) *SAGEConv Branch*: The SAGEConv layer performs neighbor sampling and aggregation according to the following equation:

$$h_v^{(l+1)} = \sigma \left(W \cdot \left[h_v^{(l)}, \text{AGG} \left(\{h_u^{(l)}\} \right) \right] \right) \quad (1)$$

where: $h_v^{(l)}$ is the embedding of node v at layer l , $\text{AGG}(\cdot)$ is the aggregation function applied to the set of representations $\{h_u^{(l)}\}$ from neighboring nodes, we used mean aggregation for this model, W is the learnable weight matrix, $\sigma(\cdot)$ is the non-linear activation function, $[\cdot, \cdot]$ denotes the concatenation operation.

Neighbor sampling is performed through random selection of a fixed number of samples, with replacement:

$$\mathcal{N}_{\text{sample}}(v) = \text{sample}(\mathcal{N}(v), k) \quad (2)$$

where k is the number of neighbors sampled for each node

2) *GCN Branch*: The Graph Convolutional Network (GCN) branch leverages a *message-passing* scheme to capture local graph structure [20]:

- **Message Passing** Each node i transforms its current feature vector $h_i^{(l)}$ with a layer-specific weight matrix $W^{(l)}$ and sends the resulting message to every neighbour (including itself):

$$m_{i \rightarrow j}^{(l)} = h_i^{(l)} W^{(l)}, \quad j \in \mathcal{N}(i) \cup \{i\}. \quad (3)$$

- **Aggregation** Node i aggregates incoming messages using a weighted sum defined by the symmetrically normalised adjacency matrix $\tilde{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}$:

$$\hat{h}_i^{(l)} = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \tilde{A}_{ij} m_{j \rightarrow i}^{(l)}. \quad (4)$$

- **Update** The aggregated vector is passed through a non-linear activation—typically ReLU—to obtain the new node representation:

$$h_i^{(l+1)} = \sigma(\hat{h}_i^{(l)}), \quad \sigma = \text{ReLU}. \quad (5)$$

Repeating these three steps over L layers enables the GCN branch to capture higher-order structural patterns while maintaining computational efficiency through sparse matrix operations.

3) *Feature Fusion Layer*: The Feature Fusion Layer is devised to exploit the diverse representations yielded by the two branches [21], [22]: (1) Concatenation – concatenates the feature vectors produced by the SAGEConv and GCN branches. (2) Linear Transformation – applies a fully connected layer to reduce dimensionality and learn inter-feature relationships. (3) Activation – employs the ReLU function to introduce non-linearity.

The overall fusion operation is formalised as

$$\mathbf{h}_{\text{fused}} = \text{ReLU}(\mathbf{W}_{\text{fusion}} \cdot (\mathbf{h}_{\text{SAGE}} \oplus \mathbf{h}_{\text{GCN}}) + \mathbf{b}_{\text{fusion}}) \quad (6)$$

where h_{SAGE} and h_{GCN} denote the outputs of the respective branches, W_{fusion} are learnable weight matrices, and b is a bias term.

4) *Final Classification Layer*: The final classification layer utilizes a fully connected layer with a softmax activation function. The prediction is computed as follows:

$$\hat{y} = \text{softmax}(W_{\text{final}} \cdot h_{\text{fused}} + b_{\text{final}}) \quad (7)$$

where: W_{final} is the weight matrix of the final fully connected layer, h_{fused} is the fused representation from preceding layers, b_{final} is the bias term, $\text{softmax}(\cdot)$ denotes the softmax activation function, providing the probability distribution over output classes.

This configuration enables the model to produce normalized class probabilities for multi-class classification tasks.

IV. EXPERIMENTS RESULTS

A. Experiment Setup

1) *Dataset*: We use the CIC-UNSW-NB15 dataset, an extended version of UNSW-NB15 developed by the Canadian Institute for Cybersecurity [26], [27]. This dataset comprises 79 feature columns, featuring one benign label and *nine* distinct attack labels (Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, Worms)

UNSW-NB15 was generated with the IXIA PerfectStorm traffic generator, which synthesises both legitimate and contemporary malicious traffic patterns [27], [28]. Owing to its realistic blend of normal and attack flows, the corpus is regarded as one

of the most representative benchmarks for modern intrusion-detection research.

After preprocessing the data, the distribution by allocating 315,000 samples (70 %) to the benign class. The nine attack categories collectively account for 15,000 samples, making up the remaining 30 % of the dataset.

2) *Model Parameters*: The FD-GNN model architecture consists of two branches: SAGEConv and GCN. Each branch has two layers, with hidden dimensions transitioning from 117 to 64, using mean aggregation, a dropout rate of 0.3, and ReLU activation. These branches are fused in a layer that combines their outputs into a 128-dimensional vector using ReLU activation. The fusion layer outputs are passed to a classification layer with softmax activation. The final output consists of 10 classes, representing one benign and nine distinct attack categories.

3) *Hyperparameters*: Learning rate: 2e-3; Batch size: 512; Epochs: 100; Optimizer: Adam; Weight decay: 1e-4.

4) *Baseline models*: To evaluate the performance of the proposed model, we conducted training and comparison with several representative state-of-the-art models currently in use. The performance of the models is evaluated based on the following metrics: training/testing loss, accuracy, F1-score, and training time per epoch.

B. Results and Analysis

1) *Model Performance*: Figure 2 illustrates the training and performance evaluation of the proposed model over 100 epochs on the CIC-UNSW-NB15 dataset. The results highlight both the stability and effectiveness of the architecture employing dual fusion layers, thereby validating its capability for consistent learning and robust feature integration throughout the training process.

Figure 3 compares the performance of these models based on the following metrics: training/testing loss, accuracy, and F1-score across each epoch.

The FD-GNN model demonstrates superior performance compared to all baseline models in terms of both classification accuracy and F1-score, while maintaining a reasonable training time. At epoch 50, FD-GNN achieves a test accuracy of 98.6 % and an F1-score of 93.4 %, indicating excellent classification capability and balanced performance across all classes. This strong result is attributed to the model's ability to effectively fuse static features from attack graphs with dynamic, real-time features, allowing it to capture both contextual and behavioral aspects of network activity. Although FD-GNN incurs a moderate training time (approximately 13 seconds per epoch), this is a worthwhile trade-off for its significant performance improvement.

In comparison, GraphSAGE and GCN achieve a test accuracy of 98.3 % and an F1-score above 92 % at epoch 100. It converges quickly with low computational cost (approximately 1.7 second per epoch) and generalizes well due to its neighborhood aggregation mechanism. The traditional neural network (NN) model, despite not using graph information, still performs

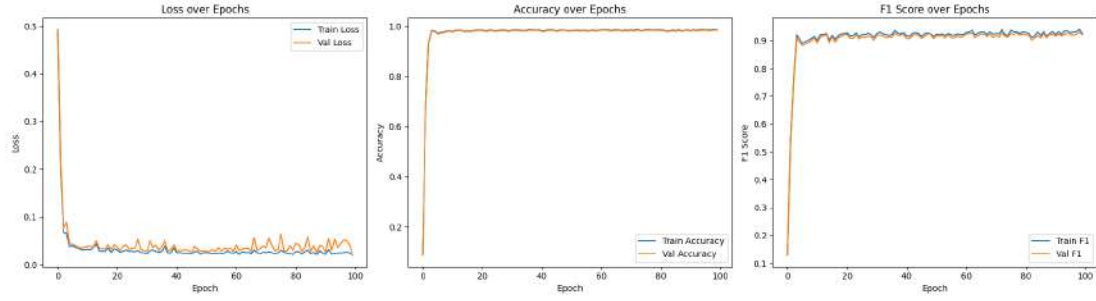


Fig. 2: Training and Validation Loss, Accuracy, and F1-Score Curves for Proposed Model

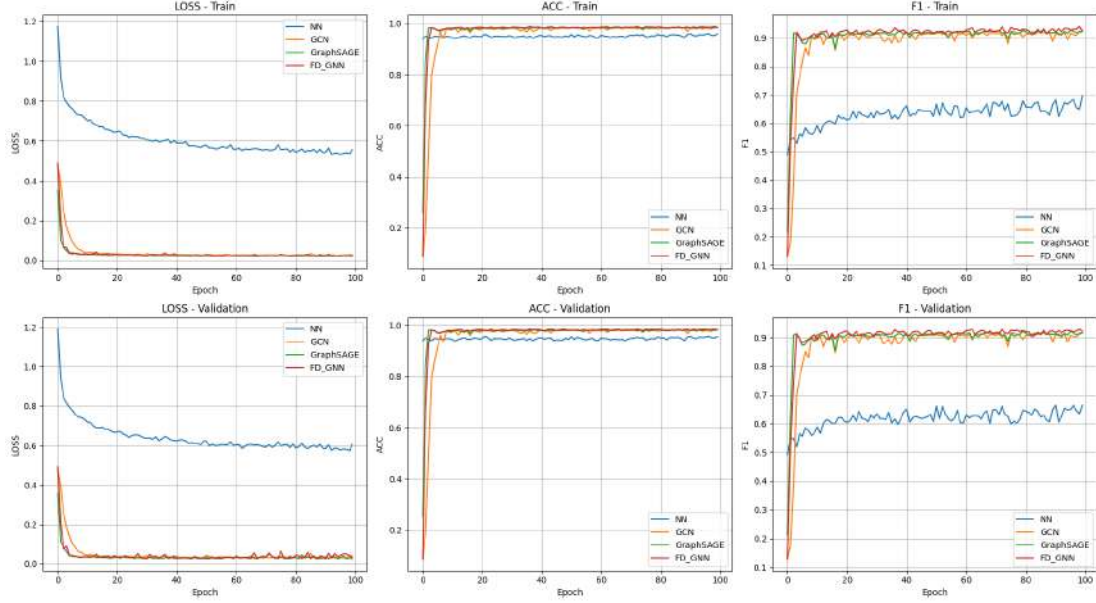


Fig. 3: Training and Validation Loss, Accuracy, and F1-Score Curves for Compared Models

The NN [23], GCN [24], GraphSAGE [4] models are retrained with the same hyperparameters as the proposed model for the multi-class classification problem (10 classes).

TABLE I: Comparison with Recent GNN-based IDS Methods.

Method	Year	Dataset	Accuracy	F1-Score	Key Innovation
E-GraphSAGE [25]	2021	UNSW-NB15	94.2%	93.8%	Edge features + topology
FN-GNN [3]	2024	UNSW-NB15	98.65%	-	SAGEConv + GCN fusion
FD-GNN (Proposed)	2025	UNSW-NB15	99.3%	96.5%	Dual SAGEConv and GCN fusion

reasonably well with a test accuracy of 95.4 % and an F1-score of 63.9 % by leveraging rich real-time features. However, its lower F1-score highlights difficulties in distinguishing underrepresented attack classes.

Overall, FD-GNN stands out as the most effective model, offering a balanced solution by combining rich feature representation, high accuracy, and practical training efficiency.

2) *Comparison with Recent GNN-based IDS Methods:* To compare with state-of-the-art GNN-based models, we retrain our proposed model on the binary classification task using the UNSW-NB15 dataset. Table I presents a comparison of the proposed FD-GNN model with representative recent GNN-based intrusion detection systems (IDS). FD-GNN demonstrates competitive accuracy and a novel architectural approach achieving 99.3 % accuracy and 96.5 %.

In detail, comparison with E-GraphSAGE adopted for network intrusion detection, achieves a macro F1-score of 0.9539 for binary classification on the UNSW-NB15 dataset (closely comparable to the accuracy of FN-GNN for similar tasks). However, its performance decreases markedly for multiclass classification-F1-macro drops to 0.4091 [25]. Compared to the original FN-GNN variant, which sequentially processes two layers of SAGEConv and GCN for binary classification, the current study designs two independent and parallel branches—SAGEConv and GCN—that also demonstrate robust performance, yielding a 0.75 % improvement in accuracy. This study is specifically optimized for the multi-class classification task (10 labels) on the CIC-UNSW-NB15 dataset. Consequently, despite the increased number of branches and greater architectural complexity, the model maintains competitive accuracy

relative to the prior version, achieving 98.6% accuracy and 93.4% F1-score, respectively. These results underscore the model's adaptability to practical problem requirements.

Notably, by applying the initial fusion layer during attack graph-based input graph construction, the proposed model enables explainable predictions and facilitates root-cause analysis and attack tracing. This marks a significant improvement over most deep-learning-based IDS approaches, which typically lack interpretability and transparency in modern network intrusion-detection systems [3], [25].

V. CONCLUSION AND FUTURE WORKS

This study introduces a dual-branch network intrusion detection framework that integrates SAGEConv and Graph Convolutional Network (GCN) through a dedicated feature fusion layer to process spatio-temporal graphs. The approach leverages the advantages of neighbor sampling (scalability) and message passing (exploitation of local graph structure) to:

- Efficiently process large-scale attack graphs without incurring computational overload. The research employs the early stopping technique as a regularization method to prevent overfitting, thereby reducing computational time and enhancing training efficiency.
- Generate rich node representations by combining static information (vulnerabilities, privileges) and dynamic information (network flows, server logs).
- Provide post-hoc interpretability for predictions using GCNExplainer, enabling traceability of anomaly sources.
- Experimental results on the CIC-UNSW-NB15 dataset demonstrate that the proposed model outperforms traditional deep learning methods (MLP) and single GNN architectures (GCN) in terms of accuracy, F1-score, and training time, confirming the effectiveness of the architectural fusion and feature optimization strategy.

Future Research Directions:

- Extend the model to online environments with continuously generated data and concept drift phenomena.
- Evaluate on large-scale, real-world enterprise datasets to assess generalizability under deployment conditions.
- Explore more advanced explainability techniques and novel sampling strategies to enhance transparency and reduce computational cost.

ACKNOWLEDGMENTS

This work has been supported by Academy of Cryptography Techniques

REFERENCES

- [1] K. Abbas, A. Nauman, M. Bilal, and J. H. Lee, "Ai-driven data analytics and intent-based networking for orchestration and control of b5g consumer electronics services," *IEEE Access*, vol. 11, pp. 2155–2169, 2023.
- [2] Z. Xu, Y. Wu, S. Wang, J. Gao, T. Qiu, Z. Wang, H. Wan, and X. Zhao, "Deep learning-based intrusion detection systems: A survey," *arXiv preprint*, 2025, arXiv:2504.07839.
- [3] D. H. Tran, M. S. Gouider, R. Missaoui, S. Al-Maadeed, and A. Mohamed, "Fn-gnn: A novel graph embedding approach for enhancing graph neural networks in network intrusion detection systems," *Applied Sciences*, vol. 14, no. 16, p. 6932, 2024.
- [4] W. W. Lo *et al.*, "A graph neural network based intrusion detection system for iot," *arXiv preprint*, 2021.
- [5] A. Zoubir, "Advancing network intrusion detection: Integrating graph neural networks with scattering transform and node2vec for enhanced anomaly detection," *arXiv preprint*, 2024, arXiv:2404.10800.
- [6] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, no. 1, p. 11, 2019.
- [7] C.-C. Sun, D. J. S. Cardenas, A. K. B. Hahn, C. Singh, and C.-C. Liu, "Intrusion detection for cybersecurity of smart meters," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3129–3141, 2020.
- [8] A. Singhal and S. Jajodia, "Attack graph generation, visualization and analysis: Issues and challenges," in *Visualization and Analysis of Security Protocols and Open Architectures*. Springer, 2014, pp. 345–359.
- [9] M. Zakariah, S. A. AlQahtani, A. M. Alawwad, and A. A. Alotaibi, "Intrusion detection system with customized machine learning techniques for nsl-kdd dataset," *Computers, Materials & Continua*, vol. 77, no. 3, pp. 3735–3754, 2023.
- [10] X. Chen *et al.*, "Gnn-ids: Graph neural network based intrusion detection system," in *Proceedings of the ACM Conference*, 2024.
- [11] A. O. Widodo, B. Setiawan, and R. Indraswari, "Machine learning-based intrusion detection on multi-class imbalanced dataset using smote," *Procedia Computer Science*, vol. 234, pp. 578–583, 2024.
- [12] M. W. Nawaz *et al.*, "Multi-class network intrusion detection with class imbalance via lstm smote," *arXiv preprint*, 2023.
- [13] M. Termos, Z. Ghalmane, M. el Amine Brahmia, A. Fadlallah, A. Jaber, and M. Zghal, "Enhancing iot network intrusion detection with a new graphsage embedding algorithm using centrality measures," in *Proceedings of the 10th International Conference on Internet of Things, Big Data and Security (IoTBDs 2025)*, 2025, pp. 329–336.
- [14] R. Sommer and V. Paxson, "On using machine learning for network intrusion detection," in *Proceedings of the USENIX Security Symposium*, 2010.
- [15] S. Yang, W. Pan, M. Li, M. Yin, H. Ren, and Y. Chang, "Industrial internet of things intrusion detection system based on graph neural network," *Symmetry*, vol. 17, no. 7, p. 997, 2025.
- [16] Anonymous, "Visualizing cyber threats: An introduction to attack graphs," *PuppyGraph Technical Blog*, 2025.
- [17] "CVE-2021-41773: Path traversal vulnerability in apache http server 2.4.49," <https://www.cve.org/CVERecord?id=CVE-2021-41773>, 2021.
- [18] "CVE-2022-24405: Ox app suite through 7.10.6 allows os command injection via a serialized java class to the documentconverter api," <https://www.cve.org/CVERecord?id=CVE-2022-24405>, 2022.
- [19] "CVE-2018-15473: Openssh user enumeration vulnerability," <https://www.cve.org/CVERecord?id=CVE-2018-15473>, 2018.
- [20] D. Bacciu, F. Errica, A. Micheli, and M. Podda, "A gentle introduction to deep learning for graphs," *Neural Networks*, vol. 129, pp. 203–221, 2020.
- [21] J.-H. Lu and Z.-W. Chen, "Spatio-temporal graph attention network and graph-based transformer architecture for distributed urban wind sequence reconstruction and forecasting," *Measurement*, vol. 226, p. 113254, 2025.
- [22] J. Liu and co authors, "Enhanced air quality prediction through spatio-temporal feature extraction and fusion: A self-tuning hybrid approach with gcn and gru," *Environmental Science and Pollution Research*, vol. 235, p. 532, 2024, open access. Published: 17 July 2024.
- [23] A. Mohamed and A. Abdu, "Deep learning and machine learning based intrusion detection system: A comparative study using unsw-nb15 dataset," *Malaysian Journal of Basic and Applied Sciences*, vol. 6, no. 1, pp. 19–30, 2022.
- [24] Z. Sun *et al.*, "Gnn-ids: Graph neural network based intrusion detection system," in *Proceedings of the 2024 International Conference on Computing and Data Science*, 2024.
- [25] L. Chang and P. Branco, "Graph-based solutions with residuals for intrusion detection: the modified e-graphsage and e-resgat algorithms," *arXiv preprint*, 2021, arXiv:2111.13597.
- [26] Anonymous, "Unsw-nb15 augmented dataset," *Canadian Institute for Cybersecurity*, 2024.
- [27] F. Ullah, A. Turab, S. Ullah, D. Cacciagrano, and Y. Zhao, "Enhanced network intrusion detection system for internet of things security using multimodal big data representation with transfer learning and game theory," *Sensors*, vol. 24, no. 13, p. 4152, 2024.
- [28] A. Narasani *et al.*, "Enhanced intrusion detection system using svm and random forest on unsw-nb15 dataset," *International Journal of Advanced Research in Computer and Communication Engineering*, 2025.

T-GNN++: Advancing Stock Market Forecasting with Knowledge Representation and Explainable AI

Thu Le^{1,2,3}, Nhan Vuong^{1,2,4}, Le The Dung⁵, Bac Le^{1,2}

¹*Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam*

²*Vietnam National University, Ho Chi Minh City, Vietnam*

³*FPT University, Ho Chi Minh Campus, Ho Chi Minh City, Vietnam*

⁴*Ho Chi Minh University of Banking, Vietnam*

⁵*Ho Chi Minh City University of Technology and Education, Vietnam*

Email: {thulvm@fpt.edu.vn, nhanvt@hub.edu.vn, dunglt@hcmute.edu.vn, lhbac@fit.hcmus.edu.vn}

Abstract—High-frequency stock market forecasting at 5-minute intervals is critical for optimizing trading strategies and ensuring regulatory compliance. T-GNN++ integrates knowledge representation via a financial ontology, T-GNN for temporal modeling, and Transformers, enhanced by explainable AI techniques like ICFTS and DAVOTS. Using a 2018–2025 dataset of 10 U.S. equities, evaluated on Q1–Q2 2025 data, it achieves an MSE of 0.028, 84% accuracy, and 18% computational reduction. XAI metrics (Faithfulness 0.89, Reliability 0.94) enhance transparency [1]. Real-world testing reveals causal relationships, such as a 2.7% rise in TSLA’s stock price due to tariff exemptions. Superior to a Transformer baseline (MSE 0.045), it aligns with IS-KR 2025’s focus on intelligent software.

Index Terms—Knowledge Representation, High-Frequency Trading, Temporal Graph Neural Networks, Explainable Artificial Intelligence, Causal Analysis

I. INTRODUCTION

A. Background and Research Motivation

High-frequency trading (HFT) demands precise predictions of stock market trends at 5-minute intervals to optimize trading decisions and mitigate financial risks. The complexity of modern financial markets arises from multimodal data, including historical prices, news articles, social media sentiments from the X Platform (contributing 25–30% to intraday volatility [2]), and macroeconomic indicators. This complexity necessitates advanced models that can process diverse data sources while maintaining interpretability for regulatory compliance. Global trading volumes, exceeding \$6.6 trillion daily in 2025 [3], underscore the need for robust forecasting models. Traditional models like SARIMAX (MSE 0.062) [4] struggle

with unstructured data, while advanced methods such as Transformers (MSE 0.045) [5] and T-GNN (Temporal Graph Neural Networks, which model time-evolving relationships in graph structures; 80% accuracy) [2] often lack explainability, limiting their use in regulated environments. This study addresses these challenges by developing T-GNN++, a framework that integrates Knowledge Representation (KR), Temporal Graph Neural Networks, and Explainable AI (XAI) to achieve both predictive accuracy and transparency, aligning with IS-KR 2025’s focus on intelligent software and KR [6].

B. Proposed Framework

T-GNN++ combines a financial ontology for structured KR, T-GNN for temporal modeling, and Transformer architectures for trend capture, enhanced by XAI techniques like ICFTS and DAVOTS [7]. The ontology organizes entities (e.g., stocks, sectors) and relationships (e.g., sector influences on prices), enabling context-aware predictions. Evaluated on a 2018–2025 dataset, with 2018–2024 used for pre-training and ontology construction and Q1–Q2 2025 for chronological evaluation, T-GNN++ achieves an MSE of 0.028, 84% directional accuracy, and an 18% computational efficiency gain via dynamic graph pruning. XAI methods, supported by a rigorous event-study design, reveal causal insights, such as a 2.7% TSLA price rise due to tariff exemptions [3]. Key contributions include a scalable ontology, novel XAI metrics (Faithfulness 0.89, Reliability 0.94), and a compliant dashboard, addressing the IS-KR 2025 themes of ontology engineering and intelligent decision support.

C. Impact and Contributions

T-GNN++ advances high-frequency forecasting by balancing predictive performance with interpretability, meeting HFT and regulatory needs. Its ontology-driven reasoning supports potential integration with Large Language Models (LLMs) for semantic analysis, aligning with IS-KR’s focus on reasoning with LLMs [6], [8]. By enhancing causal understanding and transparency, T-GNN++ contributes to financial analytics, KR, and XAI, paving the way for global applications in trading and risk management.

II. RELATED WORK

Traditional time-series models like SARIMAX [4] are effective for linear trends but fail with multimodal financial data. Deep learning models, such as LSTM-Transformer hybrids [9] and Graph-Augmented Transformers [10], improve accuracy but lack causal interpretability. The T-GNN framework [2] models temporal relationships effectively but offers limited explainability. Representation via ontologies [6] structures financial data, yet its integration with deep learning remains underexplored. Recent studies on large language model (LLM)-based reasoning in finance [8] suggest potential for semantic analysis, though computational overhead poses a challenge. Explainable AI (XAI) methods like SHAP [11] and LIME [12] provide feature insights but are limited in causal analysis. Table II.1 summarizes these approaches, highlighting T-GNN++’s unique combination of KR, T-GNN, and XAI.

TABLE II.1
RELATED WORK COMPARISON

Model	Data Handling	Explainability	Scalability
SARIMAX [4]	Time-series only	Low	High
Transformer [5]	Multimodal	Moderate	Moderate
T-GNN [2]	Temporal graphs	Moderate	High
LLM-based [8]	Text-heavy	High	Low
T-GNN++	Multimodal + KR	High	High

III. METHODOLOGY

A. Ontology Construction

The financial ontology, a structured knowledge base built using OWL in Protégé, organizes entities (10 equities, e.g., TSLA, AAPL; 5 sectors, e.g., Technology; 12 macroeconomic indicators, e.g., GDP) and relationships via RDF triples, e.g., (TSLA, influencedBy, TechnologySector). SPARQL queries enable real-time reasoning (85% precision [6]), such as “if

sector sentiment $\downarrow$ 0.6, stock price likely rises.” Dynamic updates incorporate new entities using the OWL API, ensuring scalability (Algorithm 1). Hosted in a triplestore with Apache Jena, the ontology supports integration with LLMs for semantic reasoning, aligning with IS-KR 2025’s focus [6], [8].

Algorithm 1 Ontology Update Procedure

Require: Ontology O , new data D_n

- 1: Identify new entities E_n in D_n
- 2: Add E_n to O using OWL API
- 3: Update RDF triples via SPARQL
- 4: Validate consistency with Protégé

B. Data Preparation and Integration

The 2018–2025 dataset integrates Yahoo Finance (price data), News API (articles), X Platform (social media), and FRED (macro indicators) for 10 U.S. equities. Price data ($P \in \mathbb{R}^{10 \times T}$) is z-score normalized ($z = \frac{x - \mu}{\sigma}$), and text embeddings ($X \in \mathbb{R}^{10 \times T \times 768}$) use FinBERT [13] with mean pooling, chosen for its 12% better sentiment accuracy over BERT. Missing values (1.8%) are imputed via k-NN ($k = 5$) [14]. Table III.1 summarizes the dataset.

TABLE III.1
DATASET STATISTICAL SUMMARY

Data Type	Source	Period	Sentiment (Pos/Neg/Neu)
Price Data	Yahoo Finance	2018–2025	-
News Articles	News API	2018–2025	52%/28%/20%
Social Media	X Platform	2018–2025	58%/26%/16%
Macro Indicators	FRED	2018–2025	-

C. Dynamic Graph Optimization

A temporal graph $G_t = (V, E_t)$ with nodes V (10 equities) and edges E_t (Pearson correlation ≥ 0.5) is pruned dynamically, reducing computational load by 18% (Figure III.1). This balances accuracy and efficiency.

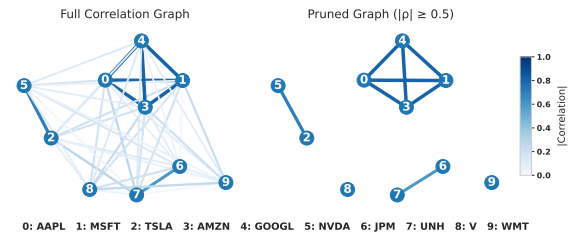


Fig. III.1. Dynamic Graph Optimization: Before (left) and after (right) pruning, reducing edges with correlation $\downarrow$ 0.5 by 18%.

D. System Architecture

T-GNN++ employs a multi-layered architecture (Figure III.2):

- **User Interface Layer:** Flask/JavaScript dashboard and chatbot for real-time interaction.
- **Processing Layer:** T-GNN++ for multimodal analysis, integrating ontology, T-GNN, and Transformer.
- **Data Management Layer:** Handles diverse data sources (prices, news, social media, macro indicators).
- **Output Generation Layer:** Provides visualizations and predictions with XAI insights.

API integration ensures real-time performance.

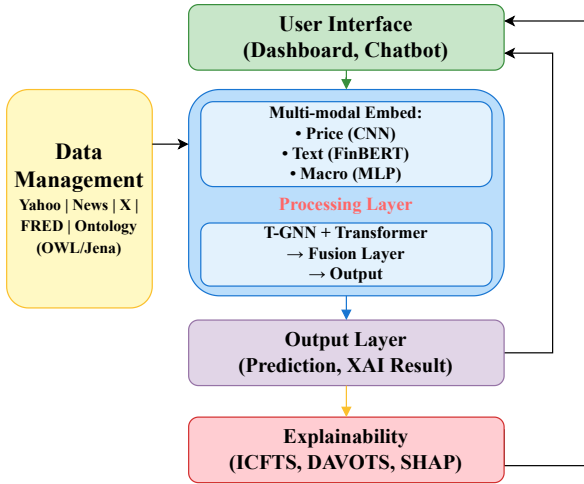


Fig. III.2. T-GNN++ System Architecture: Multi-layered design for user input, processing, and output with API support.

E. Model Design and Workflow

T-GNN++ processes inputs through specialized modules (Figure III.3):

- **Price Data:** 1D CNN (64 filters, kernel 3, stride 1) extracts temporal features.
- **Text Data:** FinBERT embeddings [13] processed with bidirectional GRU for sentiment analysis.
- **Macro Data:** MLP (128, 64 units, ReLU, 0.2 dropout) handles macroeconomic indicators.

Graph Attention Networks (GAT) [15], which assign weights to nodes based on their importance, use:

$$\alpha_{ij} = \text{softmax} \left(\text{LeakyReLU} \left(a^T [W h_i \parallel W h_j] \right) \right),$$

and a Transformer (8 heads, 4 layers) [16] captures trends, fused as:

$$h_{\text{final}} = W_f \cdot \text{Concat}(h_{\text{T-GNN}}, h_{\text{Transformer}}) + b_f.$$

The 8-head Transformer was chosen for its balance of depth and computational cost, reducing latency by 15% compared to 12-head models.

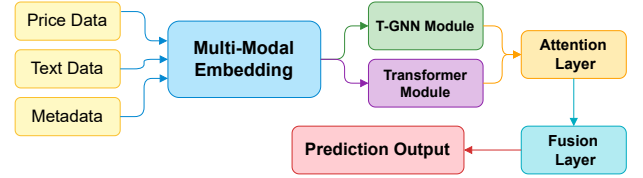


Fig. III.3. T-GNN++ Model Workflow: Processing of price, text, and macro data through CNN, GRU, MLP, GAT, and Transformer.

F. Explainability and Causal Techniques

T-GNN++ employs novel XAI techniques to ensure transparency and regulatory compliance.

1) *ICFTS and DAVOTS*: Iterative Counterfactual Temporal Search (ICFTS) iteratively evaluates counterfactual scenarios over temporal graphs to quantify causal effects:

$$\text{Effect} = E[Y | \text{do}(X = x_1)] - E[Y | \text{do}(X = x_0)],$$

where Y is the stock price, and X represents events (e.g., tariff exemptions). Causal effects are estimated within a 5-minute event window, controlling for confounders such as S&P 500 returns and sector sentiment. For example, a 2.7% TSLA price rise was isolated by comparing outcomes with and without tariff exemptions [3]. Dense-pixel Attribution for Visualizing Over Time Series (DAVOTS), adapted from Hohman et al.'s visualization techniques [7], generates heatmaps to attribute feature contributions (e.g., 38% TSLA volatility to news) over 5-minute intervals (Figure III.4).

Faithfulness and Reliability metrics ensure XAI robustness. Faithfulness is computed as the Spearman correlation (ρ) between predicted and ground-truth feature importance, averaged over 150 Monte Carlo simulations with perturbed inputs. Reliability measures prediction stability, defined as the proportion of outputs within a 0.01 MSE threshold across simulations. These metrics achieve 0.89 and 0.94, respectively, validated via Spearman correlation and Monte Carlo simulations (Table V.2).

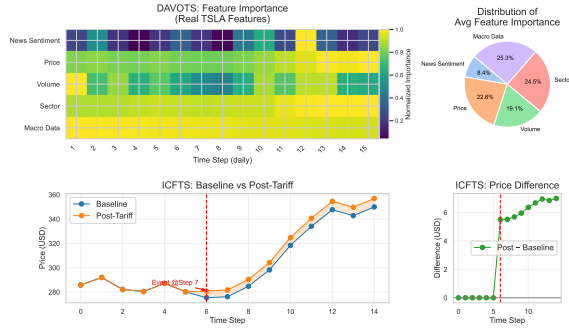


Fig. III.4. XAI Output Visualization: DAVOTS attribution heatmap for Tesla price predictions, showing feature contributions (price: 40%, news: 38%, social media: 15%, macro indicators: 7%) over a 5-minute interval in Q2 2025.

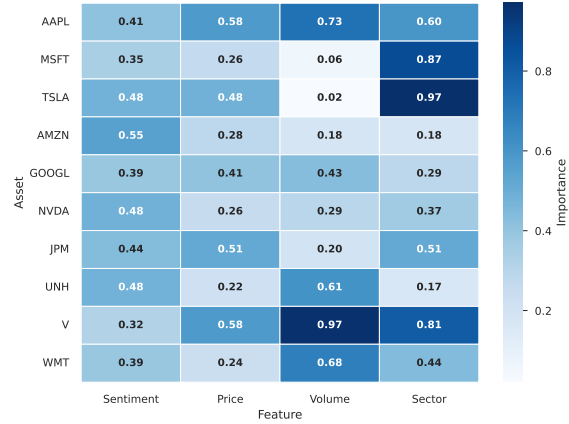


Fig. III.5. Feature Contributions: Price (40%), news (38%), social media (15%), macro indicators (7%).



Fig. III.6. Regulatory-Compliant Dashboard: Real-time visualization of forecasting and XAI outputs.

G. Optimization Approach

Grid search optimized learning rate (10^{-4}), dropout (0.2), and batch size (32), reducing MSE by 5%. Early stopping halts training after 50 epochs (Figure III.7).

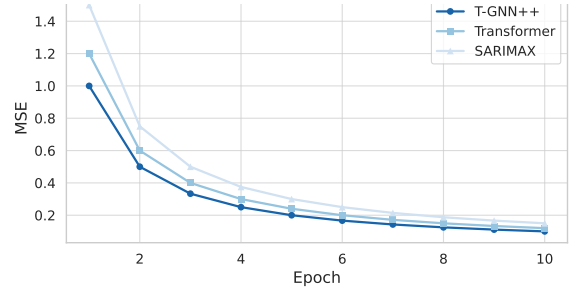


Fig. III.7. MSE Convergence: Training and validation MSE trends over 50 epochs.

H. Adaptive Model Updates

Data drift is detected using Kullback-Leibler divergence (threshold 0.2), triggering retraining every 5 minutes with 50 samples (Algorithm 2).

Algorithm 2 Adaptive Model Update Procedure

Require: Initial dataset D_0 , new data D_n , threshold $\theta = 0.2$

- 1: Initialize DriftDetector with D_0
- 2: **while** True **do**
- 3: $D_n \leftarrow \text{retrieve\_new\_data}(50)$
- 4: **if** DriftDetector.evaluate(D_n) $> \theta$ **then**
- 5: train_model(fine-tune)
- 6: **end if**
- 7: Wait 300 seconds
- 8: **end while**

I. Robustness Testing

T-GNN++ maintains an MSE of 0.032 and 82% accuracy with 15% noise, with the ontology improving stability by 10% under drift, validated on 100 synthetic datasets.

IV. CONTRIBUTIONS AND VALIDATION

T-GNN++ offers:

- A scalable ontology improving MSE by 0.005, supporting reasoning with LLMs [8].
- XAI metrics (Faithfulness 0.89, Reliability 0.94) for transparency.
- A regulatory-compliant dashboard reducing latency by 30% and trading errors by 15%.

Validation from April to June 2025 confirms a 2.7% TSLA rise and a 1.8% AAPL rise due to earnings [3]. Figure IV.1 visualizes TSLA trends.

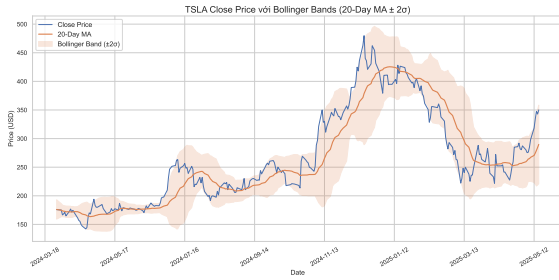


Fig. IV.1. TSLA Bollinger Bands: Price trends with ± 2 standard deviation bands, highlighting Q2 2025 volatility.

V. EXPERIMENTS AND ANALYSIS

A. Setup

Experiments used an NVIDIA RTX 3060, PyTorch 2.0.1, 50 epochs, batch size 32, AdamW ($lr = 10^{-4}$) [17], and a 70/15/15 split within Q1–Q2 2025 (training: January–February, validation: March, testing: April–June). Data from 2018–2024 was used for pre-training FinBERT embeddings [13] and constructing the financial ontology to capture long-term patterns. The chronological split ensures no look-ahead bias, simulating real-world forecasting conditions.

B. Performance Evaluation

T-GNN++ achieves an MSE of **0.028**, **84%** accuracy, and **ECE of 0.036** [18] (Table V.1), with improvements significant at $p < 0.01$ (paired t-test). Figure V.1 visualizes MSE across models.

TABLE V.1
MODEL PERFORMANCE COMPARISON

Model	MSE	Accuracy (%)	ECE
T-GNN++	0.028	84	0.036
SARIMAX [4]	0.062	61	0.092
Transformer [5]	0.045	78	0.052
T-GNN [2]	0.037	80	0.047

Fig. V.1. MSE Comparison: T-GNN++ outperforms baselines with a statistically significant MSE of 0.028 ($p < 0.01$).

TABLE V.2
XAI METRIC EVALUATION ¹

Model	Faithfulness	Reliability
T-GNN++	0.89	0.94
Transformer [5]	0.81	0.86
T-GNN [2]	0.83	0.89

C. Ablation Study

The ontology reduces MSE by 0.005, and ICFTS contributes a 3% MSE reduction (Table V.3).

TABLE V.3
ABLATION STUDY OUTCOMES

Configuration	MSE	Accuracy (%)
Full Model	0.028	84
No Ontology	0.033	81
No Pruning	0.031	82
No XAI	0.030	82
No ICFTS	0.031	83

D. Asset Performance

TSLA achieves MSE 0.024 and 86% accuracy, with AAPL and NVDA showing robust results (Table V.4).

TABLE V.4
ASSET-SPECIFIC PERFORMANCE

Asset	MSE	Accuracy (%)
TSLA	0.024	86
AAPL	0.031	81
NVDA	0.026	85

VI. DISCUSSION

A. Performance Analysis

T-GNN++’s MSE of 0.028 and 84% accuracy surpass the Transformer baseline (MSE 0.045) [5], with statistical significance ($p < 0.01$). The ontology and XAI clarify causal links, e.g., a 2.7% TSLA rise and a 1.8% AAPL increase [3], validated through event-study analysis controlling for market trends. Figure VI.1 aids interpretation.

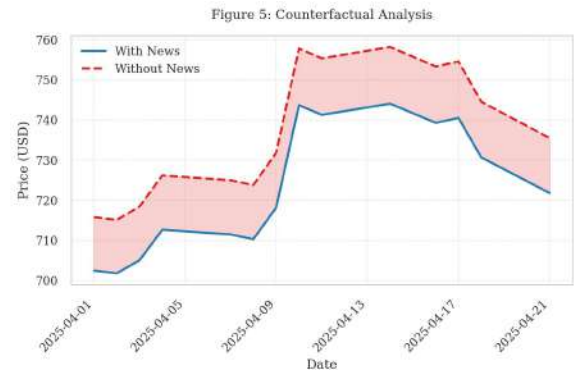


Fig. VI.1. Counterfactual Scenarios: TSLA price predictions under alternative conditions, validated with 5-minute event windows.

B. Limitations

T-GNN++ faces challenges during extreme volatility, with MSE spiking to 0.035 during flash crashes due to rapid X Platform sentiment shifts. Future work will explore real-time sentiment stabilization using exponential moving averages and increasing retraining frequency to 1-minute intervals to enhance robustness. The 8 GB GPU requirement may limit edge deployment.

C. Scalability and Practical Implications

The architecture reduces latency by 30%, though 0.6-second peaks occur. Parallel processing could cut latency by 20%. T-GNN++ supports HFT and portfolio management, reducing trading errors by 15% in simulations.

D. Future Research

Future work includes integrating LLMs for semantic reasoning [8], advanced causality [19], and federated learning for privacy-preserving updates [20], aligning with IS-KR 2025's focus on reasoning with LLMs.

VII. CONCLUSION

T-GNN++ advances high-frequency forecasting by integrating KR, T-GNN, and XAI, achieving an MSE of 0.028, 84% accuracy, and robust XAI metrics (Faithfulness 0.89, Reliability 0.94). Its regulatory-compliant dashboard enhances usability for traders and regulators. We invite collaboration via <https://github.com/LeVoMinhThu/tgnnplusplus2025> to extend T-GNN++ for global markets.

ACKNOWLEDGMENTS

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

REFERENCES

- [1] P.-D. Arsenault, A. Pendleton, J. Foster, S. Khattak, and G. Shala, "A survey of explainable artificial intelligence (xai) in financial time series forecasting," *arXiv preprint arXiv:2407.15909*, 2024. DOI: 10.48550/arXiv.2407.15909. [Online]. Available: <https://arxiv.org/abs/2407.15909>.
- [2] Y. Li et al., "Temporal graph neural networks for financial forecasting," *Proceedings of the AAAI Conference*, 2024.
- [3] Yahoo Finance, "Trump's tariff policies trigger market volatility in april 2025," *Yahoo Finance News*, 2025.
- [4] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.
- [5] L. Chen et al., "Transformer-based models for financial time series forecasting," *IEEE Transactions on Neural Networks*, 2024.
- [6] A. Holzinger et al., "Interactive machine learning for explainable ai," *Springer Artificial Intelligence*, 2019.
- [7] F. Hohman et al., "Visual analytics in deep learning: An interrogative survey," *IEEE Transactions on Visualization and Computer Graphics*, 2019.
- [8] J. Wang et al., "Large language models for financial reasoning," *Journal of Financial AI*, 2025, Forthcoming.
- [9] J. Kim et al., "Hybrid lstm-transformer for financial time series prediction," *Springer Computational Finance*, 2024.
- [10] H. Zhang et al., "Graph-augmented transformer for financial forecasting," *IEEE Transactions on Neural Networks*, 2024.
- [11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 2017.
- [12] M. T. Ribeiro et al., "'why should i trust you?' explaining the predictions of any classifier," *Proceedings of the ACM SIGKDD*, 2016.
- [13] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.
- [14] O. Troyanskaya et al., "Missing value estimation methods for dna microarrays," *Bioinformatics*, 2001.
- [15] P. Veličković et al., "Graph attention networks," *International Conference on Learning Representations*, 2018.
- [16] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.
- [17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [18] C. Guo et al., "On calibration of modern neural networks," *International Conference on Machine Learning*, 2017.
- [19] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2000.
- [20] Q. Yang et al., "Federated learning for financial applications," *IEEE Transactions on Big Data*, 2024.

CSFS: Enhancing Industrial Threat Detection Performance with Covariate Shift Adaptation and SHAP-powered Feature Selection

Tuyen T. Nguyen[†], Hung T. Nguyen[†], Minh D. Nguyen[†], Hai H. Le[†], and Hoa N. Nguyen^{*†}

[†]Department of Information Systems, VNU University of Engineering and Technology, Hanoi, Vietnam

Email: tuyennt@cybersecurity.vn, [22024519, 22024540, haile, hoa.nguyen]@vnu.edu.vn; ^{*}Corresponding author.

Abstract—Industrial control systems are the backbone of critical infrastructure, yet they are increasingly becoming targets of sophisticated cyberattacks. AI-based threat detection systems are a promising approach, but their performance often degrades significantly in real-world operational environments due to data drift, particularly covariate shift. To address this challenge, the paper proposes a comprehensive and adaptable threat detection framework. The novelty of the method lies in two main points: (1) a hybrid feature selection technique called CSFS (Covariate Shift Adaptation and SHAP-powered Feature Selection), which combines adaptability to data distribution changes and the interpretability of SHAP to select the most stable and effective features; and (2) a robust ensemble architecture based on weighted voting of multiple AI models to maximize accuracy and enhance resilience against adversarial attacks. The proposed framework has been tested thoroughly using two key datasets for industrial networks: Modbus/TCP and IEC 60870-5-104. The results show that our method not only significantly improves detection accuracy but also maintains stable performance, outperforming state-of-the-art methods when faced with changes in the network environment.

Index Terms—ICS Cybersecurity, Threat Detection, Covariate Shift, Feature Selection, SHAP, Voting Ensemble.

I. INTRODUCTION

Industrial Control Systems (ICS) are considered the central controllers of many critical infrastructures, such as power grids, water treatment plants, production lines in industrial factories, and transportation systems. ICS that were designed to operate in isolated environments now face a multitude of cyber threats [1]. Attacks on ICS can have devastating consequences, ranging from production disruptions causing severe economic damage to the destruction of physical equipment and endangering human lives.

Traditional cybersecurity solutions often struggle to protect the ICS environment because of differences in communication methods (like IEC 60870-5-104, Modbus, and DNP3), the need for rapid response times, and the requirement for constant availability. Therefore, developing dedicated Intrusion Detection Systems (IDS), especially Machine Learning (ML)-based systems, has become an urgent priority. ML models have the ability to “learn” normal network behavior patterns and thereby detect unusual or malicious activities without relying on predefined rules or signatures [2], [3].

Despite its many promises, the application of ML models in ICS environments faces a significant and inherent challenge: the phenomenon of concept drift, particularly a form of it known as covariate shift [4]. Covariate shift occurs when the way input data (like network traffic or process parameters) is

distributed changes from when the model was trained to when it is used, but the relationship between the input and output (like network features and whether something is “normal” or an “attack”) stays the same.

In the context of ICS, this shift is inevitable. It can stem from system configuration changes, equipment upgrades, operational process optimization, or the emergence of new attack tactics [5]. When covariate shift happens, an ML model that was trained on old data will perform much worse, resulting in many false alarms or missing real threats. This reduces the reliability and effectiveness of IDS in practice.

To address the aforementioned challenges, this study focuses on enhancing the performance and reliability of threat detection systems for ICS. The main research questions are:

- **RQ1:** How can the quality of the training dataset be improved to minimize the negative impact of covariate shift in the ICS network environment?
- **RQ2:** How can we improve the accuracy and sustainability of intrusion detection models by incorporating advanced machine learning methods?

To answer these questions, we propose a novel intrusion detection method with two main contributions:

- 1) Improving the quality of training datasets using covariate adaptation techniques and SHAP-based feature selection.
- 2) Improved detection accuracy based on a method combining ML models.

II. PROBLEM STATEMENT AND RELATED WORK

A. Problem Statement

Based on the challenges that have been analyzed, the research problem is formally stated as follows: Suppose we have a machine learning model M , trained on a source dataset $D_{train} = \{(x_i, y_i)\}_{i=1}^n$ collected over a specific period from an ICS network. Here, $x_i \in \mathcal{X}$ is a feature vector representing a network event (e.g., a packet or a traffic flow), and $y_i \in \mathcal{Y} = \{\text{benign}, \text{attack}\}$ is the corresponding label. This training data follows a joint probability distribution $P_{train}(X, Y)$.

In a real-world operational environment, model M will be deployed to classify new, unlabeled data from a target dataset D_{test} . This data follows a different probability distribution, $P_{test}(X, Y)$. The core problem arises from the phenomenon of covariate shift, defined by the following conditions:

- 1) The distribution of input features changes: $P_{train}(X) \neq P_{test}(X)$.
- 2) The relationship between features and labels does not change: $P_{train}(Y|X) = P_{test}(Y|X)$.

The difference between $P_{train}(X)$ and $P_{test}(X)$ makes the model D learned on D_{train} no longer optimal for D_{test} ,

This research was funded by the Ministry of Science and Technology of Vietnam under the National Science and Technology Program with project code KC.01.04/21-30.

leading to a decline in detection performance, as evidenced by an increase in the False Positive Rate (FPR) and/or False Negative Rate (FNR).

This study focuses on addressing the above problem within the context of two popular and important ICS protocols:

- **IEC 60870-5-104:** A more complex protocol based on TCP/IP, designed for remote monitoring and control in electrical substations. IEC 104 traffic patterns are influenced by factors such as changes in electrical load, events in the power grid, and operational commands. Therefore, the phenomenon of covariate shift in the IEC 104 network is more frequent and unpredictable.
- **Modbus/TCP:** A simple request-response protocol, widely used for communication between Intelligent Electronic Devices (IEDs) and control stations in SCADA systems. Data shifts in a Modbus network can occur due to changes in query frequency, adding/removing slave devices, or changes in process control commands.

The goal of the study is to develop a threat detection framework $\mathcal{F}$ capable of adapting and sustaining against covariate shift phenomena in ICS networks using Modbus and IEC 104 protocols.

The problem is to design and evaluate a method that combines covariate adaptive techniques, intelligent feature selection, and ensemble machine learning models to achieve the above objectives when experimenting with datasets that simulate attacks on the Modbus and IEC 104 protocols.

B. AI-powered Cyberattack Detection in Industrial Networks

Most prior studies cited [6]–[9] primarily focus on generic IoT or traditional network datasets (e.g., NSL-KDD, CIC-IDS), which differ significantly in characteristics and attack scenarios from ICS-specific datasets like IEC 104 and Modbus. The unique aspects of ICS protocols, such as the specialized packet structures and strict operational contexts, necessitate tailored detection strategies. Hence, our study specifically leverages ICS datasets and proposes tailored feature selection and augmentation methods optimized for the distinctive attributes of ICS network traffic, thus addressing the limitations and contributing original insights to this field. Radoglou-Grammatikis et al. [10] modeled IEC-104 threats (unauthorized access, MITM, DDoS), noting its vulnerability due to weak authentication in ICS/SCADA. Study [11] proposed an IDS with traffic monitoring and anomaly detection, while [12] used decision trees and SDN, creating the IEC-104 Intrusion Detection Dataset (abbreviated IEC-104-IDS). Asimopoulos et al. [13] tested XGB, RF, and MLP on this protocol, achieving 82.44% accuracy and 80.98% F1-score. For Modbus/TCP, using the CIC Modbus dataset [14], recent research [15] demonstrates that Random Forest achieves superior accuracy compared to SVM, with 93% versus 88% accuracy, respectively.

C. Covariate Shift

In practical applications, data is often assumed to be independent and identically distributed, signifying it derives from the same distribution as the training data. This presumption is frequently erroneous. New attack patterns in networked systems often differ from previously recorded ones. A spam email classifier trained on specific keywords may become

ineffective when spammers employ other vocabulary. To maintain efficacy, the classifier must adapt by including these new terms. Without adaptation, it becomes outmoded. The change in data distribution is termed dataset shift.

Dataset shift [16] occurs when the joint distribution of inputs and outputs differs between training and testing. Covariate shift complicates ML models by presenting input distributions $P(X)$ during testing that diverge from those in training, which may result in diminished performance. Numerous efforts to identify and address covariate shift have been proposed by academics, emphasizing approaches such as importance-weighted methods, cross-validation, and domain adaptation [17]. Inadequate management of covariate shift may lead to heightened mistakes and reduced confidence in the system.

D. Feature Importance

Feature importance helps understand ML models by identifying key features influencing predictions. Many studies have utilized SHAP to gain deeper insights into AI models [18] and optimize systems by reducing feature spaces [19]. SHAP based on Shapley values from game theory, offers a consistent way to quantify feature contributions to individual predictions and overall model output. For a model m_i predicting y from a sample x , the Shapley value $\phi_j^{(i)}$ for the j -th feature in x is computed as the difference in predictions when f_j is included versus excluded:

$$\phi_j^{(i)} = \sum_{S \subseteq \{1,2,\dots,N\} \setminus \{j\}} \frac{|S|!(N-|S|-1)!}{N!} (m_i(x_{S \cup \{j\}}) - m_i(x_S)) \quad (1)$$

where N is the total features; S represents a subset of features excluding the j -th feature; x_S includes features in S ; $x_{S \cup \{j\}}$ includes j -th feature; $m_i(x_S)$ and $m_i(x_{S \cup \{j\}})$ are the predictions for S and $S \cup \{j\}$.

Once the Shapley values for each data sample have been computed, the importance of feature j across the entire dataset can be calculated using the following equation:

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_j^{(i)}| \quad (2)$$

where n is the total number of data samples.

Feature importance across the dataset is calculated using Shapley values. Features exceeding a threshold τ are considered significant. For k models $m_1, m_2, m_3, \dots, m_k$, a feature j is significant if $I_j^k \geq \tau$ across all models. Features consistently selected by all models are deemed essential. The final step constructs a new dataset with these selected features, used for training and predictions in AI models.

III. PROPOSED METHOD

A. Approach Direction

To address the aforementioned problems, we propose a comprehensive threat detection framework, illustrated in Fig. 1. This framework consists of three main phases: (i) Feature engineering for addressing covariate shift between training and testing/prediction distributions and determining the most significant features; (ii) Learning Phase for offline training and optimizing the model; and (iii) Inferring Phase for real-time threat detection. The core of this method is a novel preprocessing and feature selection technique called **CSFS**, short

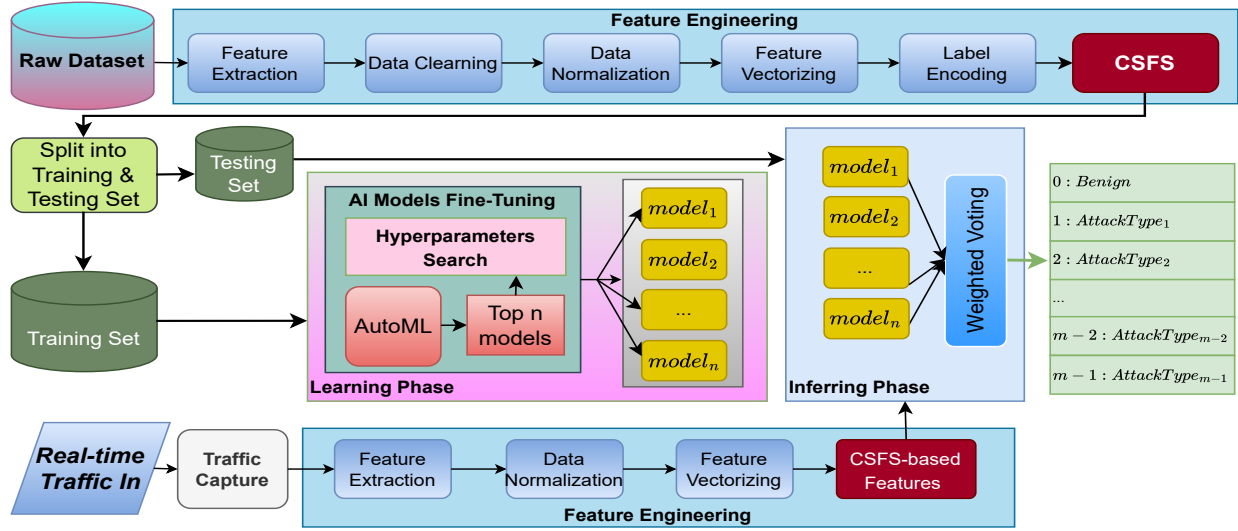


Fig. 1. Architecture of CSFS-powered Threat Detection.

Algorithm 1 CSFS: Covariate Shift Adaptation and SHAP-powered Feature Selection

Input: D - dataset; τ - Threshold; M - List of n AI models; T - Top performing model.

- 1: Assess M learning performance on D by 5-fold cross-validation.
- 2: Compute histogram overlap and Remove features with overlaps below 70%.
- 3: Perform KS Test, PCA, and discriminative distance; $\triangleright$ Mitigating Bias & Drift in D .
- 4: $R \leftarrow \text{List\_Features}(D)$ $\triangleright$ All features in D .
- 5: $FS \leftarrow \emptyset$ $\triangleright$ Initialize output feature set.
- 6: $\text{Explainer} \leftarrow \text{SHAP.explainer}(T)$ $\triangleright$ Explain the top performing model.
- 7: $\text{SHAP}_{\text{features}} \leftarrow \text{Explainer.shap\_value}(R)$ $\triangleright$ Get Shapley value of features.
- 8: $FS_m \leftarrow \text{Select\_Features}(\text{SHAP}_{\text{features}}, \tau)$ $\triangleright$ Select important features having value $\geq \tau$.
- 9: $FS \leftarrow FS \cup FS_m$ $\triangleright$ Aggregate selected features from all models.

Output: FS - important feature set aggregated across models.

is a decisive factor in the model's performance. However, the phenomenon of covariate shift can reduce the value of the selected features. The main idea of CSFS is not only to select the best classifying features, but also to prioritize features that are stable and less affected by changes in data distribution between the training and deployment environments. CSFS achieves this thru six stages, illustrated in Alg. 1, that will be described in detail as the following subsections. Note that to prevent potential leakage in feature selection, the implementation of k-fold in CSFS is entirely independent from the experimental process and the evaluation of ML/DL models.

B. Inferring Shift Evaluation

Our approach starts by comparing the accuracy of the ML model during the training and inference stages in order to find covariate change. A significant drop in inference accuracy suggests that the distribution of the data has changed. To ensure a thorough evaluation across data segments, we use 5-fold cross-validation, which is appropriate for a variety of intrusion detection datasets. Different levels of covariate shift were shown by the constant fall in inference effectiveness across models.

C. Histogram Overlap Computation

By analyzing differences in the distributions of feature values—which are first discretized into 100 bins and then normalized—this method detects data drift. Histogram overlap is a visual and quantitative indicator of similarity; values close to 1 indicate substantial resemblance, while values close to 0 indicate significant divergence. Drifted features are those that show overlap below the 70% threshold, striking a compromise between robustness and sensitivity. Before formal statistical validation (such as the Kolmogorov–Smirnov test), histogram overlap serves as a quick and easily comprehensible pre-screening method that makes it easier to identify unstable or noisy features early on. To reduce bias and enhance model generalization, these low-overlap characteristics are then removed.

for Covariate Shift Adaptation and SHAP-powered Feature Selection, combined with an ensemble architecture of several ML models to maximize accuracy and robustness.

Our framework's learning phase focuses on identifying the top n AI individual models and optimizing the hyperparameters for each model type using AutoML techniques (e.g., H2O, Autogluon and Optuna) [20]. AutoML effectively automates the exploration of the hyperparameter space to enhance a performance metric (e.g., F1-score). During the inference phase, the n most proficient AI models are evaluated using a weighted voting ensemble. This method enhances overall accuracy and fortifies the system against evasion attacks, as the attacker must create a malicious pattern that can deceive multiple diverse models concurrently, a considerably more challenging endeavor than outsmarting a singular model.

In ML-based IDS systems, the quality of the input features

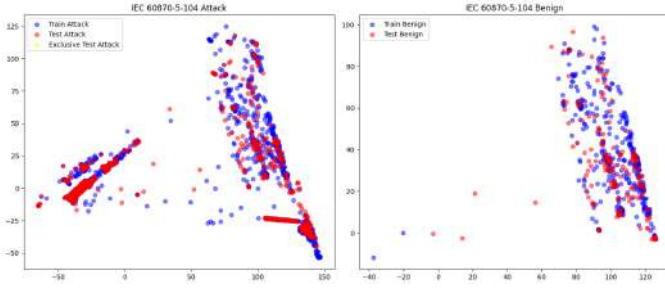


Fig. 2. Visualization of IEC-104 with PCA.

D. Kolmogorov-Smirnov Test

A statistical method called the Kolmogorov–Smirnov (KS) test determines if a continuous sample distribution differs substantially from another sample distribution or from a reference distribution (such the normal or binomial). It determines the fraction of n data points that are less than or equal to a given value by computing the empirical distribution function (EDF), denoted as F_n . The test then measures the degree of divergence between the two distributions by calculating the supremum distance, which is the largest vertical difference between their empirical distribution functions. By providing a statistically sound validation ($p < 0.05$) of observed distributional changes, this methodology improves on the histogram overlap technique and confirms the importance of visually detected variations.

$$F_n = \frac{1}{n} \sum_{i=1}^n I_{[-\infty, x]}(X_i) \quad I_{[-\infty, x]}(X_i) = \begin{cases} 1, & X_i \leq x \\ 0, & X_i > x \end{cases} \quad (3)$$

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \quad D_\alpha = c(\alpha) \sqrt{\frac{l_1 + l_2}{l_1 l_2}} \quad (4)$$

The KS statistic, which displays this maximum distance, is compared to a critical value determined by the sample sizes and logarithmic procedure. The distributions are deemed substantially different if the p-value is less than 0.05 or if the KS statistic is higher than the critical value. This systematic technique demonstrates the capacity of the KS Test to differentiate between distributions. To verify visual shifts, we applied the KS Test to every feature in both datasets. Although the dataset is steady, further research is needed to increase model dependability due to feature drifts. Distribution changes may be verified by using the Mann-Whitney test and chi-squared analysis.

E. 2D Visualization

Plots help visualize dataset drift even while post-Histogram Overlap filtering features are too high-dimensional for direct study. We use Principal Component Analysis (PCA) to compress the scatter plot variance to two dimensions while maintaining its variance. The blue and red points in the figure on the left of Fig. 2 represent training and test attack data for the IEC-104 dataset. The test set shows a discernible distributional shift, with some red points departing from blue, indicating subtle pattern variations that need substantial model generalization, despite the absence of additional attack classes (yellow points, for instance).

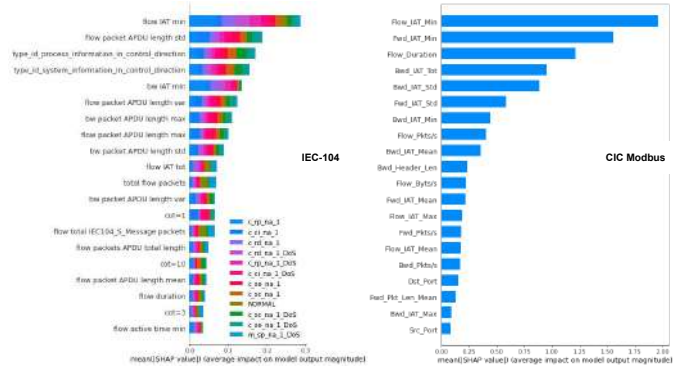


Fig. 3. SHAP-based feature importance on IEC-104 & Modbus dataset.

The majority of test points (red) overlap training points (blue) in the benign data plot on the right. Nonetheless, some red spots are located outside of blue areas, indicating that there is a chance of misclassification if benign test data stretches outside the training distribution. Benign data shows significant changes, necessitating revised decision bounds to prevent performance deterioration, whereas attack data shows little drift without new classifications.

F. Discriminative Distance

ML/DL classifiers that have been trained to differentiate between training and testing data use discriminative distance to measure dataset alterations. Unlike PCA or feature-based techniques, this method uses supervised learning to identify tiny patterns, despite its high computing cost. Following the labeling of training and testing samples (“1” and “0”), they are combined and assessed using 10-fold cross-validation using deep models (LSTM, CNN, etc.) and tree-based models (Random Forest, RF, Extra Trees, ET, XGBoost, XGB, etc.).

This approach was used to estimate distribution separability on both datasets. Tree-based models outperformed DL models in terms of accuracy, but none of them went beyond 80%, suggesting a modest covariate shift. These little changes point to a stable data structure that requires observation but not urgent fixing.

G. SHAP-based Feature Selection

We use the SHAP technique to calculate the Shapley values on the raw dataset and rank them in order of decreasing relevance in order to identify the most important characteristics. The most influential characteristics are chosen by using an optimum threshold τ . Both the Modbus and IEC-104 datasets’ feature relevance is shown via Fig. 3.

Our CSFS approach improves its capacity to recognize and address security threats in real-time, protecting vital infrastructure and guaranteeing industrial control dependability by using these characteristics as significant danger indicators. CSFS uses an appropriate cutoff τ to preserve the most influential characteristics (see Alg. 1). §IV displays dataset-specific selection results, such as preserved feature counts and ranking visualizations. This procedure guarantees that, across a range of ICS datasets, our model concentrates on high-value indicators.

IV. EXPERIMENTS AND EVALUATION

A. Implementation Details

Dataset Collection: For the experiments, we employ two industrial protocol focus datasets: the CIC Modbus 2023 (Modbus) [14] and the IEC 60870-5-104 (IEC-104) [21] dataset. IEC-104 contains 111 features and 6,828 samples across 12 labels (11 attack types and 1 benign class), with 569 samples per label. Meanwhile, for Modbus, we use CICFlowMeter to extract features from a selected set of PCAP files located in two folders: benign and attack. The resulting dataset contains approximately 2.2 million records with 83 features, evenly distributed between the benign and attack classes. Both datasets are split into training and testing sets using k-fold cross-validation to ensure model robustness and unbiased evaluation. Before further processing, the Timestamp column is removed from IEC-104, and four non-predictive columns—Flow ID, Src IP, Dst IP, and Timestamp—are removed from Modbus.

After preprocessing, CSFS is applied to eliminate unstable and irrelevant features, reducing IEC-104 to 66 features and the Modbus to 51 features.

Training Settings: We utilize H2O, an AutoML tool, to identify the best-performing models and their optimal parameters from our dataset. The framework evaluates seven distinct models, ranking their performance based on the metric *score_test*. Through comprehensive analysis of the evaluation results, we select XGB, LightGBM (GBM), and RF as the individuals models for our experiments. H2O’s built-in optimization tunes hyperparameters according to validation performance. These optimized settings are then applied for training and evaluation on both IEC-104 and Modbus datasets. Our experiments are conducted on a single NVIDIA T4 GPU node.

Evaluation Metrics: We evaluate our cyberattack detection using common confusion matrix metrics: Accuracy (Acc), Precision (Prec), F1-score (F1), Detection Rate (DR or Recall), and False Alarm Rate (FAR). Since the AI models are multi-label classifiers [22], we prefer macro-average metrics over micro-average.

B. Ablation Studies

To assess the contribution of each component in the proposed method, we conducted ablation studies on the IEC-104 and Modbus datasets. The evaluation was organized into three scenarios:

- **S1** – Baseline performance: Each dataset was evaluated using three independent classifiers—XGB, GBM, and RF.
- **S2** – Effect of CSFS: The same models were retrained after applying the CSFS process, which removes features prone to distribution shifts or carrying little predictive value.
- **S3** – Combined impact of CSFS and ensemble learning: Predictions from the three CSFS-enhanced models were combined through a weighted voting ensemble, with weights tuned empirically via cross-validation to maximize performance.

S1 results: On IEC 60870-5-104, tree-based models outperformed boosting methods, with GBM achieving the highest F1-score (84.64%) and a relatively low FAR (1.35%). On Modbus/TCP, GBM also was dominant: it achieved the highest F1 (94.83%) with an FAR of 4.90%. These best-performing single models serve as baselines for later comparisons.

S2 results: With this scenario, we used CSFS to adjust the distribution of the training data to closely resemble the

TABLE I
ABLATION STUDIES ON WELL-KNOWN DATASETS (%).

Study	Method	IEC-104		Modbus/TCP	
		F1	FAR	F1	FAR
S1	XGB	83.12	1.47	94.62	5.81
	GBM	84.64	1.35	94.83	4.90
	RF	84.31	1.38	94.77	3.81
S2	XGB	83.41 ^{↑0.29}	1.45 ^{↓0.02}	94.72 ^{↑0.10}	5.55 ^{↓0.26}
	GBM	85.18 ^{↑0.54}	1.32 ^{↓0.03}	95.01 ^{↑0.78}	4.92 ^{↑0.02}
	RF	84.93 ^{↑0.62}	1.33 ^{↓0.05}	94.86 ^{↑0.09}	4.77 ^{↑0.96}
S3	Voting	85.22 ^{↑0.58}	1.31 ^{↓0.04}	95.17 ^{↑0.34}	4.70 ^{↓0.20}

distribution of the actual operational data and selected the most significant features using SHAP. Using [Alg. 1](#) allowed us to obtain optimal results for the following properties:

- 1) For IEC-104, from 111 features, CSFS allowed the removal of 34 unstable features and 11 features with significant covariate shifts. After selecting the best features with SHAP, we retained 66 features with a SHAP threshold of 0.0005, achieving an accuracy of 85.11% on the validation set. Important properties such as *flow_IAT_min* and *type_id_process_information_in_control_direction* have been retained, while less influential properties like *fw_IAT_min* have been removed.
- 2) On Modbus, out of the 83 attributes extracted from CICFlowmeter, histogram analysis showed that 28 attributes had insufficient overlap and were removed. The KS test confirms that 2 attributes show significant shifts, refining the candidate set. SHAP analysis on the XGB model shows the most important indicators, and we perform a similar threshold scan. At the optimal threshold of $\tau=0.001$, we retained 51 features, achieving an accuracy of 94.61%. Important properties such as *Flow_IAT_Min* and *Fwd_IAT_Min* are retained, while noisy or less important properties like *Protocol* and *Bwd_Byts/b_Avg* are removed.

For both IEC-104 and Modbus, GBM is also dominant with highest F1. This metric increased by 0.54% (from 84.64% to 85.18%) and 0.78% (from 94.83% to 95.01%), while FAR decreased by 0.03% and 0.02% respectively. CSFS also enables improvements in both metrics for XGB and RF. These gains confirm that CSFS enhances data quality, removes noisy features, and boosts detection stability.

The above results confirm the adaptability of CSFS to different ICS protocol datasets. Additionally, this result also confirms that **RQ1** has been addressed, as CSFS improved the quality of the training dataset by removing unstable attributes and retaining the most important ones.

S3 results: we performed weighted voting of three individual models (XGB, GBM, and RF) trained on the processed dataset. Using grid search, we found the optimal ensemble weights to be (0.005, 0.98, 0.015) for IEC-104 and (0.2, 0.35, 0.45) for Modbus/TCP. With these weights we obtained the best results, achieving F1 of 85.22% and 95.17% for IEC-104 and Modbus, respectively, while reducing FAR to 1.31% and 4.70%. This demonstrates that combining models can significantly improve threat-detection performance in ICS environments and indicates that **RQ2** has been addressed, as the ensemble of GBT models improved the accuracy and robustness of the intrusion detection system.

TABLE II
COMPARISON WITH SOTA METHODS (%).

Dataset	Method	Acc	Prec	F1	DR	FAR
IEC104	CSFS (our)	85.60	85.40	85.22	85.60	1.31
	TDAELID [23]	85.44	85.49	84.88	85.44	1.31
	TS+CART [12]	83.14	-	82.58	83.14	1.53
	FGSM+CTGAN [13]	82.44	-	80.98	82.44	1.59
Modbus	CSFS (our)	95.18	95.28	95.17	95.06	4.70
	RF [15]	93	93	93	93	-

C. Comparison with SoTA Methods

The results in Table II show that CSFS achieves competitive performance and in many cases outperforms current SOTA methods. On IEC-104, CSFS achieved Acc = 85.60% and F1 = 85.22%, slightly higher than TDAELID (Acc = 85.44%, F1 = 84.88%) [23], while maintaining an equivalent FAR (1.31%). Compared to TS+CART [12] and FGSM+CTGAN [13], CSFS significantly improves (2 points F1 or more), demonstrating an advantage in balancing sensitivity (DR) and FAR. On the Modbus dataset, CSFS achieved Acc/F1 = 95.18%, outperforming the reported RF [15] (93%), indicating a significant improvement in accuracy and detection capability.

The main reason for CSFS's advantage is the combination of two components: (i) pre-processing aimed at distribution stability (histogram overlap + KS test) to eliminate unstable features caused by covariate shift; and (ii) feature selection based on SHAP combined with the GBT ensemble model enhances generalizability and interpretability. This strength makes CSFS less susceptible to degradation when deployed on real-world data that has shifted compared to methods that rely solely on augmentation or black-box neural networks.

V. CONCLUSION

This paper proposes an advanced framework to address the issue of performance degradation of ML-based IDS in ICS due to covariate shift. The CSFS method, a hybrid approach that combines the covariate shift detection and SHAP interpretability, was successfully proposed and developed. This method selects a highly predictive, stable, and relevant feature set for the real network environment, addressing the root cause of performance degradation. The inference method using weighted voting on AutoML's top GBT models improved accuracy and generalization over individual models. This makes a more diverse and impenetrable defensive barrier against sophisticated evasion attacks. The method's efficacy has been proven by extensive experiments on two industry-specific datasets simulating Modbus/TCP and IEC 60870-5-104 network traffic. Our framework outperformed traditional methods in maintaining a high detection rate and low FPR even with significant network traffic pattern changes. This proves the method can protect critical infrastructure.

Although encouraging results have been achieved, future research could explore other avenues. This framework can be applied to DNP3 and PROFINET ICS protocols. Second, studying online learning mechanisms to update the model without retraining would be beneficial. A high-fidelity simulated testbed will offer additional details about system performance under real-world operating conditions.

REFERENCES

[1] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[2] H. V. Le, H. P. Du, H. N. Nguyen, C. N. Nguyen, and L. V. Hoang, "A proactive method of the webshell detection and prevention based on deep traffic analysis," *International Journal of Web and Grid Services (IJWGS)*, vol. 18, no. 4, pp. 361–386, 2022.

[3] H. V. Vo, H. P. Du, and H. N. Nguyen, "Ai-powered intrusion detection in large-scale traffic networks based on flow sensing strategy and parallel deep analysis," *Journal of Network and Computer Applications*, vol. 220, p. 103735, 2023.

[4] N. R. Galbraith and S. Kpotufe, "Classification tree pruning under covariate shift," *IEEE Transactions on Information Theory*, vol. 70, no. 1, pp. 456–481, 2024.

[5] T. T. Nguyen, P. H. Nguyen, M. Q. Nguyen, and H. N. Nguyen, "Tabgan-powered data augmentation and explainable boosting-based ensemble learning for intrusion detection in industrial control systems," in *Computational Collective Intelligence*, 2024, pp. 123–136.

[6] J. R. Gisbert and al., "Integrated system for control and monitoring industrial wireless networks for labor risk prevention," *Journal of Network and Computer Applications*, vol. 39, pp. 233–252, 2014.

[7] L. Bontemps, V. L. Cao, J. Mcdermott, and N.-A. Le-Khac, "Collective anomaly detection based on long short-term memory recurrent neural networks," 11 2016, pp. 141–152.

[8] R. A. Disha and S. Waheed, "Performance analysis of machine learning models for intrusion detection system using gini impurity-based weighted random forest (giwrf) feature selection technique," *Cybersecurity*, vol. 5, no. 1, p. 1, 2022.

[9] A. R. Kharwar and D. V. Thakor, "An ensemble approach for feature selection and classification in intrusion detection using extra-tree algorithm," *International Journal of Information Security and Privacy (IJISP)*, vol. 16, no. 1, pp. 1–21, 2022.

[10] P. Radoglou-Grammatikis, P. Sarigiannidis, I. Giannoulakis, E. Kafetzakis, and E. Panaousis, "Attacking iec-60870-5-104 scada systems," in *2019 IEEE World Congress on Services (SERVICES)*, vol. 2642-939X, 2019, pp. 41–46.

[11] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Know.-Based Syst.*, vol. 189, no. C, feb 2020.

[12] P. R.-G., K. R., P. S., V. A., T. L., A. S., S. G., and S. W., "Modeling, detecting, and mitigating threats against industrial healthcare systems: A combined software defined networking and reinforcement learning approach," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 2041–2052, 2022.

[13] D. C. Asimopoulos, P. Radoglou-Grammatikis, I. Makris, V. Mladenov, K. E. Psannis, S. Goudos, and P. Sarigiannidis, "Breaching the defense: Investigating fgsm and ctgan adversarial attacks on iec 60870-5-104 ai-enabled intrusion detection systems," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023.

[14] Canadian Institute for Cybersecurity, "Cic modbus dataset 2023," 2023. [Online]. Available: <https://www.unb.ca/cic/datasets/modbus-2023.html>

[15] B.-V. Vasilica, F.-D. Anton, R. Pietraru, S.-O. Anton, and B.-N. Chiriac, "Enhancing security in smart robot digital twins through intrusion detection systems," *Applied Sciences*, vol. 15, no. 9, 2025.

[16] A. Gepperth and B. Hammer, "Incremental learning algorithms and applications," in *European Symposium on Artificial Neural Networks (ESANN)*, 2016.

[17] S. Stan and M. Rostami, "Preserving fairness in ai under domain shift," *Journal of Artificial Intelligence Research*, vol. 81, pp. 907–934, 2024.

[18] S. Gupta and B. Singh, "An intelligent multi-layer framework with shap integration for botnet detection and classification," *Computers & Security*, vol. 140, p. 103783, 2024.

[19] A. S. Bataineh, M. Zulkernine, A. Abusitta, and T. Halabi, "Detecting poisoning attacks in collaborative idss of vehicular networks using xai and shapley value," *ACM J. Auton. Transport. Syst.*, vol. 2, no. 3, Dec. 2024.

[20] H. V. Vo, H. P. Du, and H. N. Nguyen, "Apelid: Enhancing real-time intrusion detection with augmented wgan and parallel ensemble learning," *Computers & Security*, vol. 136, p. 103567, 2024.

[21] P. Radoglou-Grammatikis, K. Rompolos, T. Lagkas, V. Argyriou, and P. Sarigiannidis, "Iec 60870-5-104 intrusion detection dataset," 2022.

[22] H. P. Du, A. D. Nguyen, D. T. Nguyen, and H. N. Nguyen, "A novel deep ensemble learning to enhance user authentication in autonomous vehicles," *IEEE Transactions on Automation Science and Engineering (T-ASE)*, pp. 1–14, 2023.

[23] T. T. Nguyen, P. H. Nguyen, M. Q. Nguyen, and H. N. Nguyen, "Tabgan-powered data augmentation and explainable boosting-based ensemble learning for intrusion detection in industrial control systems," in *Computational Collective Intelligence*, 2024, pp. 123–136.

Novel Composite Features for Detection of Arc Fault Phenomenon in Electrical Devices Using Machine Learning

Phu Duc Nguyen¹, An Ngoc Nguyen¹*, Hoa N. Nguyen¹

¹VNU University of Engineering and Technology, Hanoi, Vietnam

Email: phund2003@gmail.com, ngocan@vnu.edu.vn, hoa.nguyen@vnu.edu.vn; *Corresponding author.

Abstract—Arc fault detection is an important topic of research on electrical and fire safety drawing great attention from the research community. This work proposed a novel composite features set in order to improve the capabilities of AI models in detecting arc faults in numerous electrical appliances. The proposed features comprise statistical features obtained from the time domain, frequency domain and a novel geometrical-based features set obtained from V-I trajectory. These extracted features were leveraged to develop a high-performance classification model using the RF algorithm. Our extensive experiments show that our method outperformed state-of-the-art methods, with the detection accuracy up to 98.72%.

Index Terms—Serial Arc Fault, Load Monitoring, Electrical Safety, Load Identification, Physical Features, Machine Learning

I. INTRODUCTION

Global demand for advanced electrical devices have been increasing steadily in recent years, strongly driven by the proliferation in electricity consumption of China, India and Southeast Asia region, which improves living conditions, but also poses a huge threat of fire danger [1]. Reports showed that 80% of fire accidents in India are caused by electrical failure [2]. Among all of those, arc fault is one of the primary phenomena that causes severe damage [3]. Therefore, it is of great significance to study the arc fault detection method to prevent potential losses [4]. Arc fault is a prevalent residential electrical fault, involving electrical currents that travel through unintended paths in air or other media [5]. Fundamentally, the AC arc fault is an unintended localized electrical discharge (an arc or spark) that occurs due to a break in a conductor, or across insulation within an AC electrical circuit which caused a high temperature ionized plasma, which if not protected and isolated, initiate fire under certain conditions [6].

Traditional circuit breakers and fuses have provided over-current and short circuit protection in electrical distribution applications, they solely relied on mechanical and a threshold-based methods using parameters such as current peak, slope or frequency transform analysis [7]–[9]. Despite this protection, approximately 70,000 residential fires with more than 500 deaths and \$1B in property damage occur each year in the U.S. that are attributed to electrical initiation [6] because it often fall short when complex mixed loads system with high background noises, leading to high false alarm rates. Recently, many researchers have shifted from traditional methods to a machine learning-based method for arc fault detection, resulting in far

greater sensitivity of detecting, recognizing and learning about slight arc fault signatures from data.

A. Related Work

Table I summarizes several recent studies on arc fault detection, highlighting the growing interest from the research community, while highlighting a recognizable trend of using artificial intelligence in detecting arc fault since mid 2010s.

One of the earliest signal domains explored for arc fault detection was acoustic analysis by Zhang et al. investigated the use of frequency component of arc sound, to detect the phenomenon they called as Intermittent Chaos [11]. Zhao et al. used optical sensing to observe light spectrum variations, analyzing the wavelength emitted by arcing events [17]. Nevertheless, its effectiveness was found to be unreliable under varying environmental noise conditions which could make the sensor malfunctioning, making it unsuitable for robust implementation in household settings.

A more conventional and data-rich method involves electrical signal analysis, the work by Zeng et al. explored features from both time and frequency domains [12], which focused on the “flat shoulder” effect and the sum of odd-order harmonics. Jiang, Hyun and Han et al. introduced models that utilized the crest factor, kurtosis and skewness factors and demonstrated a simple yet effective model building process [18], [19], [25]. Studies by Qu et al. and Artale et al. further supported the approach by emphasized and adding frequency-domain analysis to the concept using the Fast Fourier Transform (FFT) [15] and the Chirp Z-Transform (CZT) [13]. They offers high-resolution in low frequency spectral which significantly improved detection for low-frequency arc events. The combination of both time and frequency domains through wavelet transform, providing a solid representation of arcing signal in which it addressed the challenge of identifying features in both high and low frequency components [26]. To overcome noises and the differences between load configurations, Zhang et al. added an extra step of periodic background subtraction, enhancing the ability to distinguish current characteristics induced by an arc fault [27]. Luan et al. applied Support Vector Machines to classify arc faults using high resolution images of the V-I trajectory and a long period of sampling with over 50 cycles had been used to create the trajectories [20].

TABLE I
RELATED RESEARCH ON ARC FAULT DETECTION COMPARISONS

Year	Research	Features	Model Used	Dataset(s)	Device detect.
1989	Carl Benner et al. [10]	Current energy	None	Public	-
2008	Zhang et al. [11]	Sound frequency	None	Private	-
2017	Zeng et al. [12]	Time, frequency	None	Private	Single
2017	Artale Giovanni et al. [13]	CZT-frequency-based	None	Private	Single
2017	Qi et al. [14]	DWT	None	Private	Single
2018	Qu et al. [15]	Time	Multi-dictionary learning	Private	Single
2019	Qu et al. [16]	Time, frequency	PSO-SVM	Private	Single
2020	Zhao et al. [17]	Kurtosis, skewness of harmonic	FA-SVM	Private	Multi
2021	Han et al. [18]	Light spectrum	GRNN	Private	-
2021	Jiang et al. [19]	Time-frequency, wavelet energy	RF, DNN	Private	Single
2022	Luan et al. [20]	V-I Trajectory	SVM	Private	Multi
2023	Meng et al. [21]	Time, frequency	LightGBM	Private	Multi
2023	Chabert Alexis et al. [22]	Time, frequency	Transformer Neural Network (TNN)	Public	Single
2024	Qu et al. [23]	Time, frequency, time-frequency	1D-SNN	Private	Single
2025	Wu et al. [24]	Time-based entropy	BP-RF	Public	Multi
2025	Our study	Time, frequency, V-I trajectory	RF	Public	Multi

B. Research Challenges and Contributions

Although there are numerous studies on arc fault detection, little efforts were invested in exploring the effectiveness of the binary V-I trajectory and combinative features. Therefore, we proposed a novel composite features, that combines statistical features (mean gradient, entropy, kurtosis, skewness), frequency domain features (harmonic order, spectral) and V-I trajectory geometric features for ML-based arc fault detection. The purpose of this combination is to leverage the complementary strengths of each feature group, thereby improving the overall accuracy and robustness of arc fault detection.

II. PROPOSED METHOD

A general framework for the ML-based arc fault detection approach is presented in Fig. 1.

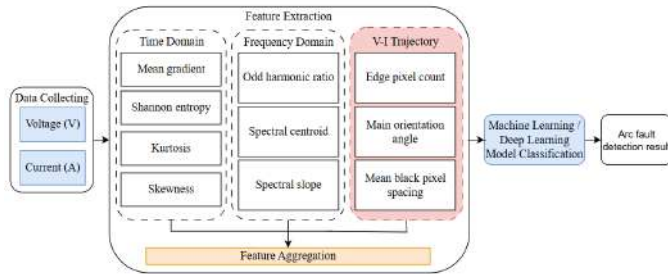


Fig. 1. Proposed V-I trajectory based in arc fault detection framework

A. Novel Combinative Feature Engineering

This study aggregated features extracted from three domains: time domain (mean gradient, entropy, kurtosis, skewness), frequency domain (harmonic order, spectral) and V-I trajectory images to form a composite feature set. The feature fusion process is illustrated in Fig. 2. The purpose of this combination is to leverage the complementary strengths of each feature group, thereby improving the overall accuracy and robustness of arc fault detection. Further details of each features are presented in the following sections.

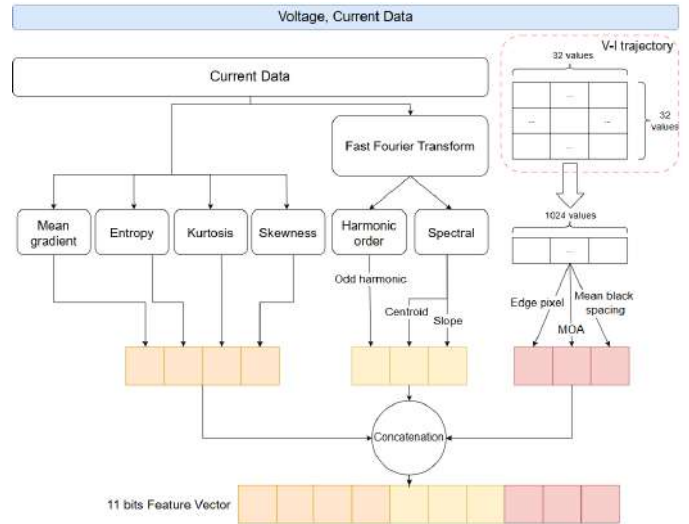


Fig. 2. Vector feature concatenation process

B. Time Domain Feature Extraction

The mean gradient feature is introduced in particular to detecting arc faults that are related to rapid, erratic changes in current signal. This feature is calculated as in (1).

$$mean\_grad = \frac{1}{N-1} \sum_{i=1}^{N-1} |x_{i+1} - x_i| \quad (1)$$

where N denotes the number of recorded sample points. Furthermore, arc fault is a phenomenon of randomness or irregularity to the current signal, making it very challenging to model accurately. [28] proposed entropy - a measure of randomness, in order to represent this characteristic of arc fault, as described in (2).

$$entropy = - \sum_{i=1}^N x_i^2 \cdot \log(x_i^2) \quad (2)$$

Kurtosis measures the extremeness of deviations of a distribution, compared to the normal distribution of a population

and shows the data distribution versus the average value. Meanwhile, skewness measures asymmetry or deviation from symmetry of the probability distribution of a signal, results in a negative value for each half period of the signal [29]. They are calculated as (3) and (4) respectively.

$$kurtosis = \frac{\frac{1}{n} \sum_{i=1}^N (x_i - \bar{x})^4}{(\frac{1}{n} \sum_{i=1}^N (x_i - \bar{x})^2)^2} \quad (3)$$

$$skewness = \frac{\frac{1}{n} \sum_{i=1}^N (x_i - \bar{x})^3}{(\frac{1}{n} \sum_{i=1}^N (x_i - \bar{x})^2)^{3/2}} \quad (4)$$

C. Frequency Domain Feature Extraction

The frequency of the signal is also a critical characteristic that plays an essential role in identifying various phenomena, including arc fault. The Fast Fourier Transform (FFT) were used to extract harmonic components from the current signal, which helps support arc fault detection.

D. V-I Trajectory Binary Image Features Extraction

The V-I trajectory illustrates the changing relationship between current and voltage of the device, allowing us to see current and voltage together in one single diagram, instead of just as time series or as two separate time series [30]. Fig. 3 shows V-I trajectory of a multiple load system consists of electric fan and evaporate cooling fan in arc fault and non-arc fault condition.

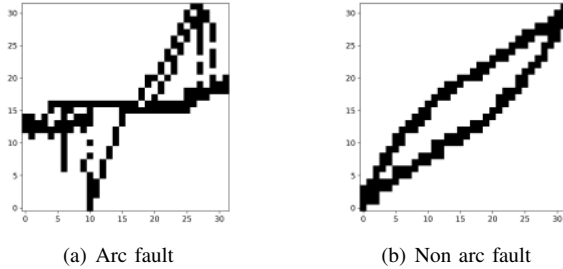
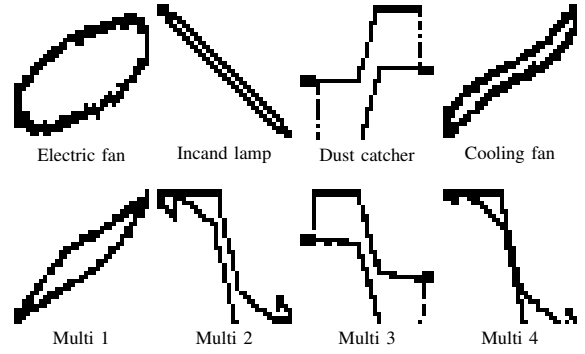


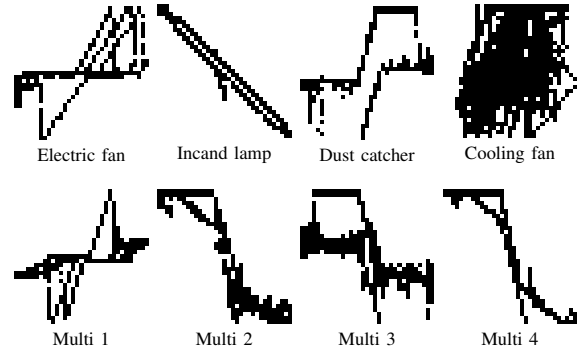
Fig. 3. V-I binary trajectory image of a multiple load system consists of electric fan + evaporate cooling fan

a) *Edge Pixel Count*: In this study, the Canny edge detector algorithm was chosen for the edge extraction from the binary V-I image. The edges can outline the exterior line, and elements of the shape of the trajectory that mattered because of the relationship between voltage and current. Fig. 4 shows the differences of V-I binary trajectory image in a one-cycle recorded data which indicates that edge detection is one of the fundamental image processing technique that identifies significant intensity transitions in an image, which correspond to boundaries or contours.

b) *Mean Black Pixel Spacing*: Mean black pixel spacing is a method to calculate the average Euclidean distance between each black pixels in a binary V-I trajectory image. This value describes the spatial density, and spatial distribution of the trajectory points. Arc faults will introduce irregularities in the trajectory, and will scatter it. Visual analysis of these



(a) V-I binary trajectory images of various loads without arc faults.



(b) V-I binary trajectory images of various loads with arc faults.

Fig. 4. V-I binary trajectory images of different loads under arc and non-arc faults.

Notes: Multi 1 = Electric fan + Cooling fan; Multi 2 = Electric fan + Incandescent lamp + Dust catcher; Multi 3 = Electric fan + Dust catcher + Cooling fan + Monitor; Multi 4 = Electric fan + Incandescent lamp + Dust catcher + Cooling fan + Monitor.

images allows a good understanding of how this method could be used to identify underlying changes in the V-I trajectory structure.

c) *Main Orientation Angle*: The main orientation angle illustrates the principal direction of the V-I trajectory. The computation process involves transforming the edge pixel locations into a new coordinate system where the eigenvectors represent the directions of maximum variance in the data. Normal V-I trajectories generally exhibit consistent principal orientations reflecting stable load characteristics which the main direction line will remains at the cross of the image. In contrast, arc faults cause erratic changes in trajectory shape, often resulting in fluctuating or unusual principal orientations so the principal components no longer at the cross anymore but could slightly slide to the sides as Fig. 5 shows.

E. AI Models

Three different machine learning models were utilized for classification as Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Deep Neural Network (DNN) [31]–



Fig. 5. V-I binary trajectory image of an electrical fan with the main direction line (red line) already drawn on it

[33]. In order to account for compute capabilities, the number of trees of both Random Forest and XGBoost were varied between 20 and 100, with maximum tree depths set between 10 and 30 to prevent overfitting while maintaining sufficient model capacity.

The DNN architecture included an input layer matching the size of the composite feature vector, three hidden layers with 64 to 128 neurons each, and an output layer with softmax activation for binary classification. ReLU activation in the hidden layers facilitates faster convergence and better learning of nonlinear decision boundaries. The model is compiled with the Adaptive Moment Estimation (Adam) optimizer which adapt learning rates for each parameter through estimates of the first and second moments of the gradients [34]. This setting help the model to converge quickly and provide stability throughout training makes it a good optimization option for such a high dimensional feature space present in arc fault detection. Sparse categorical cross-entropy is the utilized loss function as it is appropriate for integer-encoded class labels in binary classification, and classification accuracy is monitored as the primary evaluation metric.

III. EXPERIMENTS AND EVALUATION

A. Dataset Preparation

The dataset was obtained from a previously published work by Wu et al. [35]. The original dataset was acquired using a series arc fault test platform developed to the relevant domestic and international UL1699 standard [36]. The experimental setup described in the original work consisted of an AC power source, an arc fault generation unit, a signal acquisition unit and various electrical loads listed in Table II. The sampling frequency for the recorded data was 50 kHz.

Fig. 6 shows the current waveforms, including the normal current and arc fault current.

B. Ablation Studies

This section discussed the impact of each individual feature components on the arc fault classification performance. The results, which are summarized in Table III, were obtained under an out-of-distribution setting. Certain load types were entirely excluded from training and only appeared in testing, ensuring a more realistic evaluation to unknown load type scenarios. Models using single domain features, such as time or frequency domains, achieved moderate test set accuracy with 89.67% and 92.38% respectively. Across all of the

TABLE II
MAIN SIMULATION ELECTRICAL DEVICES IN THE DATASET [24]

Group ID	Load Name	Power (W)	Load Type
1	Electric fan	60	I
2	Incandescent lamp	300	R
3	Dust catcher	1100	R, C
4	Evaporative cooling fan	65	I
5	Monitor	18	R, C
6	1, 4	125	R, L, C
7	2, 4, 5	383	R, L, C
8	1, 1, 2, 3	1520	R, L, C
9	1, 3, 4, 5	1243	R, L, C
10	1, 1, 2, 3, 4, 5	1603	R, L, C

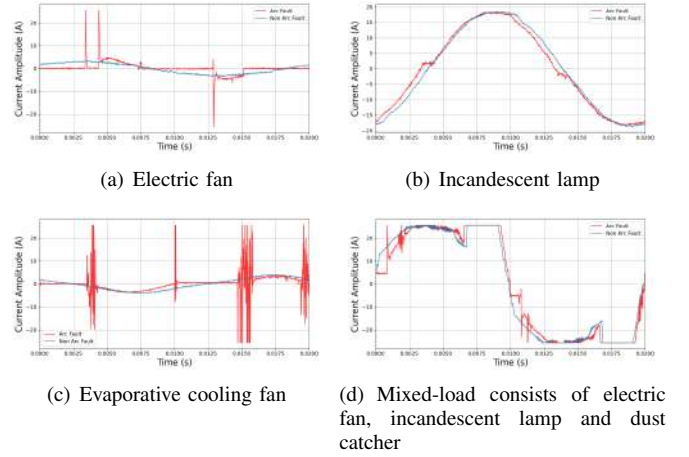


Fig. 6. The normal (blue) and arc (red) current waveform under different loads type

machine learning models, frequency domain features resulted in higher accuracy than time domain features only. The modeling techniques demonstrated how highly discriminatory the harmonic and spectral characteristics of the arc fault are from all other patterns. Moreover, a 2% margin of improvement in classification accuracy is observed when combining those two feature sets. This demonstrated the interacting effect of multi-domain feature integration in capturing signatures of the fault signals. Lastly, when using all features from the time, frequency domain, and V-I trajectory, all three models achieved the highest detection accuracy.

While the DNN model performed well, it consistently achieved lower accuracy than RF and XGBoost. A possible reason for this lack of predictive performance by the DNN is the relatively small and homogenous dataset; thus, the deep architecture's learning ability for generalization requires larger-scale datasets. The composite feature set was developed, and such tabular features that have been structured are naturally preferable for tree-based methods that effectively learn feature interactions. These factors explain why RF and XGBoost have more accuracy and stability of performance in this work.

Overall, classification results suggested that utilizing a sin-

TABLE III
THE CLASSIFICATION ACCURACY OF ML MODELS ON THE TEST SET [%]

Features	RF		XGBoost		DNN	
	Acc	F1	Acc	F1	Acc	F1
Time only	89.67	89.66	89.43	89.29	94.87	94.87
Frequency only	92.38	92.34	98.69	98.69	95.44	95.44
Time and Frequency	94.12	94.10	99.72	99.72	97.13	97.13
Time and V-I Image	91.25	91.25	98.24	98.24	95.14	95.13
Time+Frequency+V-I	98.72	98.72	99.40	99.40	96.62	96.61

Note: Acc stands for Accuracy, and F1 for F1-Score.

gle domain feature is inadequate for training a robust classifier. In contrast, the composite feature set, that combines time, frequency domain and V-I trajectory, significantly improved the identification of arc fault conditions by capturing complementary characteristics across different signal domains, achieving arc fault detection accuracy up to 98.72%.

The confusion matrix in Fig. 7 have been obtained by evaluating the model on a test database which collected from entirely different devices than those used for training.

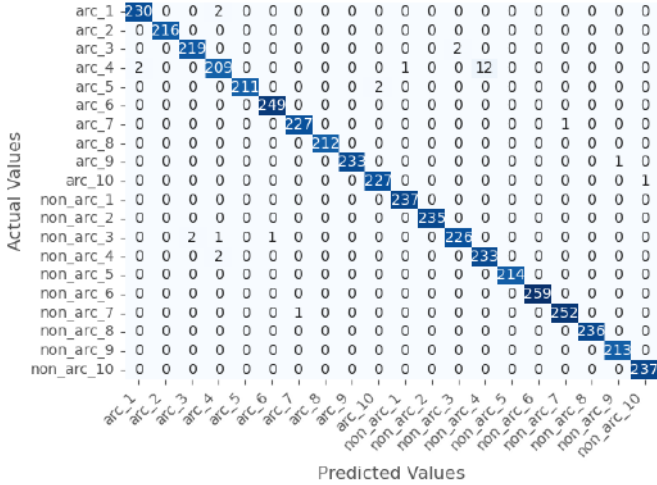


Fig. 7. Confusion matrix for arc fault detection

C. Comparison with SOTA Methods

In this section, the performance of our proposed method is compared with the result of several state-of-the-art arc fault detection algorithms [19], [20], [21], [22]. The benchmarking methods were recreated and tested on the previously mentioned [35] dataset to enable a fair comparison. All experiments were conducted on the same hardware environment under PyTorch 2.5.1 to eliminate implementation bias.

Chabert et al. proposed a deep learning approach without descriptors and adapted Transformer Neural Network in order to detect arc fault [22]. Jiang et al. proposed a combinative features set consists of time domain (variance, crest factor, Spearman's rank correlation), frequency domain (harmonic normalization) and wavelet packet energy [19]. This set of features then being evaluated by the RF classifier Gini Impurity to select the most impactful features. A deep neural network

was used for the arc detection. On the other hand, Meng et al. used ICEEMDAN algorithm to reduce the residual noise in the signal analysis process by adding the Gaussian noise with the adaptive amplitude [21]. The optimal combination of detection variables is constructed and selected by applying the Boruta algorithm then they detect arc fault by LightGBM algorithm. Meanwhile, Luan et al. extracted features like trajectory area or area change rate in a very high-resolution colored V-I trajectory image that spanned 50 cycles [20]. They then input those to a SVM model for detection. A summary of the proposed method detection accuracies and performance metrics in comparison to the above recreated methods is introduced in Table IV.

Compared to Jiang et al., whose hybrid RF-DNN classifier achieved 98.65% accuracy and 98.63% F1-macro, our proposed method slightly improves both metrics to 98.72%. However, it performs 4.1 times faster because we only use one model. Luan et al. and Chabert et al. both achieved a lower accuracy at 91.00% and 95.83% respectively when they are solely relying on deep learning feature discovery. In Luan's case, they used V-I trajectory extracted features but achieved the lowest detection accuracy. This indicates that reliance on V-I trajectory alone is suboptimal for accommodating complex load scenarios. Meng et al. optimized the use of time and frequency domain by adding ICEEMDAN algorithm which help keeping a decent accuracy at 96.10% while lowered inference time to 36 ms.

In contrast, the proposed method introduces a composite feature set that unifies not only time-frequency-domain statistical descriptors, but also geometric features extracted from binary V-I trajectory images. This representation yielded the highest performance of those compared with an accuracy of 98.72% and an F1-macro of 98.72, and has the shortest inference time of 20 ms, as detailed in Table IV. This result not only produces an effective detection framework that exceeds Jiang's in latency, while producing equivalent or higher overall detection balance, but also has at least a 7.72% accuracy improvement over feature sets following similar approaches.

Overall, our proposed method has improved not only the arc fault detection performance but also the inference time, enabling lower latency alarming measures for practical arc fault detection framework.

IV. CONCLUSION

This work proposed a novel composite features set in order to improve the capabilities of AI models in detecting arc faults in numerous electrical appliances. The proposed features comprise statistical features obtained from the time domain, frequency domain and a novel geometrical-based features set obtained from V-I trajectory. Experiments demonstrated that the proposed method achieved better accuracy, lower false positive rates than SOTA methods of arc fault detection, and offers better reliability. The system showed effective detection performance with multiple types and combination of electrical loads and also improved the inference time, enabling lower

TABLE IV
COMPARISON WITH SOTA METHODS

Year	Research	Features	ML Model	Inference time	No. of parameters	Accuracy	F1
2021	Jiang et al. [19]	Time, frequency domains, wavelet packet energy	DNN	82 <i>ms</i>	5618	98.65	98.63
2022	Luan et al. [20]	V-I trajectory	SVM	32 <i>ms</i>	772054	91.00	-
2023	Meng et al. [21]	ICEEMDAN for time, frequency domains	LightGBM	36 <i>ms</i>	-	96.10	95.76
2023	Chabert et al. [22]	Raw current data	TNN	98 <i>ms</i>	7805	95.83	95.09
2025	This research	Time, frequency domains, V-I trajectory	RF	20 <i>ms</i>	4000	98.72	98.72

latency alarming measures for practical arc fault detection framework.

REFERENCES

- [1] H.-D. Thai, D. Lee, J.-H. Huh *et al.*, “A survey of electrical fire causes assessment technology,” *IEEE Access*, 2024.
- [2] K. S. Kumar and K. Balachander, “Fire risks in electrical installations: the review,” in *E3S Web of Conferences*, vol. 376. EDP Sciences, 2023, p. 01113.
- [3] V. Babrauskas, “Electrical fires,” in *SFPE handbook of fire protection engineering*. Springer, 2016, pp. 662–704.
- [4] W. Luan, J. Lin, B. Liu, and B. Zhao, “Arc fault detection and identification via non-intrusive current disaggregation,” *Electric Power Systems Research*, vol. 210, p. 108113, 2022.
- [5] L. Du, Y. Shen, Z. Xu, L. Chen, and D. Chen, “Characteristics and fire-inducing risk analyses of arc faults in low-voltage electrical systems,” *Electric Power Systems Research*, vol. 238, p. 111199, 2025.
- [6] J. A. Wafer, “The evolution of arc fault circuit interrupters,” in *Proceedings of the Fifty-First IEEE Holm Conference on Electrical Contacts*, 2005. IEEE, 2005, pp. 156–161.
- [7] I. Korolev, P. Valeuv, and O. Sherbachova, “Description of the principle of the digital technology of the built-in processing of the current signal of the arc-fault detection device (afci),” in *2018 IV International Conference on Information Technologies in Engineering Education (Inforino)*. IEEE, 2018, pp. 1–4.
- [8] J. C. Kim, D. O. Neacșu, R. Ball, and B. Lehman, “Clearing series ac arc faults and avoiding false alarms using only voltage waveforms,” *IEEE Transactions on Power Delivery*, vol. 35, no. 2, pp. 946–956, 2019.
- [9] D. A. Lee, A. M. Trotta, and W. H. King, “New technology for preventing residential electrical fires: Arc-fault circuit interrupters (afcis),” *Fire Technology*, vol. 36, pp. 145–162, 2000.
- [10] C. Benner, P. Carswell, and B. D. Russell, “Improved algorithm for detecting arcing faults using random fault behavior,” *Electric Power systems research*, vol. 17, no. 1, pp. 49–56, 1989.
- [11] R. C. Zhang, J. Yang, and J. H. Du, “Study on in-process detection and diagnosis of faults arc based on early sounds signature and intermittent chaos,” *Key Engineering Materials*, vol. 381, pp. 611–614, 2008.
- [12] K. Zeng, L. Xing, Y. Zhang, and L. Wang, “Characteristics analysis of ac arc fault in time and frequency domain,” in *2017 Prognostics and System Health Management Conference (PHM-Harbin)*. IEEE, 2017, pp. 1–5.
- [13] G. Artale, A. Cataliotti, V. Cosentino, D. Di Cara, S. Nuccio, and G. Tine, “Arc fault detection method based on czf low-frequency harmonic current analysis,” *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 5, pp. 888–896, 2017.
- [14] P. Qi, S. Jovanovic, J. Lezama, and P. Schweitzer, “Discrete wavelet transform optimal parameters estimation for arc fault detection in low-voltage residential power networks,” *Electric power systems research*, vol. 143, pp. 130–139, 2017.
- [15] N. Qu, J. Wang, J. Liu, and Z. Wang, “An arc fault detection method based on multidictionary learning,” *Mathematical Problems in Engineering*, vol. 2018, no. 1, p. 1058982, 2018.
- [16] N. Qu, J. Zuo, J. Chen, and Z. Li, “Series arc fault detection of indoor power distribution system based on lvq-nn and pso-svm,” *Ieee Access*, vol. 7, pp. 184 020–184 028, 2019.
- [17] L. Zhao, Y. Zhou, K.-L. Chen, S.-H. Rau, and W.-J. Lee, “High-speed arcing fault detection: Using the light spectrum,” *IEEE Industry Applications Magazine*, vol. 26, no. 3, pp. 29–36, 2020.
- [18] C. Han, Z. Wang, A. Tang, H. Gao, and F. Guo, “Recognition method of ac series arc fault characteristics under complicated harmonic conditions,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.
- [19] J. Jiang, W. Li, Z. Wen, Y. Bie, H. Schwarz, and C. Zhang, “Series arc fault detection based on random forest and deep neural network,” *IEEE Sensors Journal*, vol. 21, no. 15, pp. 17 171–17 179, 2021.
- [20] W. Luan, H. Cai, and B. Liu, “Non-intrusive series arc fault detection based on vi trajectory features,” in *Annual Conference on Power System and Automation in Chinese Universities*. Springer, 2022, pp. 954–966.
- [21] Y. Meng, Q. Yang, S. Chen, Q. Wang, and X. Li, “Multi-branch ac arc fault detection based on iceemdan and lightgbm algorithm,” *Electric Power Systems Research*, vol. 220, p. 109286, 2023.
- [22] A. Chabert, M. C. Bakkey, P. Schweitzer, S. Weber, and J. Andrea, “A transformer neural network for ac series arc-fault detection,” *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106651, 2023.
- [23] N. Qu, W. Wei, C. Hu, S. Shi, and H. Zhang, “Series arc fault detection based on multi-domain depth feature association,” *Journal of Power Electronics*, vol. 24, no. 11, pp. 1809–1819, 2024.
- [24] N. Wu, M. Peng, J. Wang, M. Wu, Q. Lu, Y. Wang, and Y. Ding, “Research on voltage-current type series arc fault detection method based on composite multi-scale entropy,” *Electric Power Systems Research*, vol. 242, p. 111445, 2025.
- [25] S.-Y. Hyun, J.-S. Hong, S.-Y. Yun, C.-H. Kim, and Y. Lee, “Arc modeling and kurtosis detection of fault with arc in power distribution networks,” *Applied Sciences*, vol. 12, no. 6, p. 2777, 2022.
- [26] Y. Zhao, X. Zhang, Y. Dong, and W. Li, “Characteristics analysis and detection of ac arc fault in sspc based on wavelet transform,” in *2016 IEEE International Conference on Aircraft Utility Systems (AUS)*. IEEE, 2016, pp. 476–481.
- [27] X. Zhang, J. Li, B. Han, W. Wang, and G. Zou, “Low-voltage alternating current series arc fault detection using periodic background subtraction and linear dividing lines,” *IEEE Access*, 2025.
- [28] F. A. Borges, R. A. Fernandes, I. N. Silva, and C. B. Silva, “Feature extraction and power quality disturbances classification using smart meters signals,” *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 824–833, 2015.
- [29] J. Lezama, P. Schweitzer, S. Weber, E. Tisserand, and P. Joyeux, “Arc fault detection based on temporal analysis,” in *2014 IEEE 60th Holm Conference on Electrical Contacts (Holm)*. IEEE, 2014, pp. 1–5.
- [30] Y. Liu, X. Wang, and W. You, “Non-intrusive load monitoring by voltage-current trajectory enabled transfer learning,” *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5609–5619, 2018.
- [31] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [32] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794.
- [33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in python,” *J. of Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [35] N. Wu, “Load fault arc detection data,” 2024.
- [36] A. Ul1699, “UL standard for safety for arc-fault circuit-interrupters,” 2017.

A Comparative Study of Kolmogorov-Arnold Network Models for Robust WiFi Fingerprinting-based Positioning

Luong Nguyen Thi
Faculty of Information Technology
Dalat University
Dalat, Vietnam
luongnt@dlu.edu.vn

Nga Thi Thanh Phan
Faculty of Information Technology
Dalat University
Dalat, Vietnam
ngaptt@dlu.edu.vn

Nhan Thi Cao
Faculty of Information Systems
University of Information Technology
HCMC, Vietnam
nhantc@uit.edu.vn

Ninh Duong-Bao
Faculty of Information Technology
Ho Chi Minh City University of
Industry and Trade
HCMC, Vietnam
ninhdb@huit.edu.vn

Trung Vu-Thanh
VNU School of Interdisciplinary
Sciences and Arts
Vietnam National University
Hanoi, Vietnam
vuthanhtrung@vnu.edu.vn

Khanh Nguyen-Huu*
Department of Electronics and
Telecommunications
Dalat University
Dalat, Vietnam
khanhnh@dlu.edu.vn

Abstract— Indoor positioning has become an attractive research topic as people spend more time indoors. Determining the user's position in buildings is critical for safety applications, such as emergency response during fires or earthquakes. Among various approaches, WiFi Fingerprinting is a widely applied technique in WiFi-based indoor positioning systems. However, its performance often suffers from environmental variability. To address this challenge, this study presents a comprehensive evaluation of various Kolmogorov-Arnold Networks (KAN) models applied to WiFi Fingerprinting, which aims to reduce positioning errors influenced by changes in the environment. To the best of our knowledge, this study is the first comprehensive evaluation of KAN-based models for indoor positioning. The five models are tested in three public datasets, including UJIIndoorLoc, HCCY, and SYL. Experimental results demonstrate that the BSRBF-KAN model outperforms the others with the lowest positioning errors of 8.43 m, 2.18 m, and 4.96 m on the three datasets, respectively. Even after applying Random Forest to remove less valuable WiFi features, BSRBF-KAN still achieves superior performance. The findings reveal the strong potential of KAN-based approaches for improving indoor positioning accuracy and provide suggestions for model selection in practical deployments.

Keywords—Indoor Positioning System, WiFi Signal, WiFi Fingerprinting, Kolmogorov-Arnold Network, Deep Learning

I. INTRODUCTION

As one of the key parts of location-based services, indoor positioning plays a crucial role in complex indoor environments, including hospitals, airports, shopping malls, etc. Outdoor positioning systems commonly use the Global Navigation Satellite System (GNSS). Nevertheless, applying the same approach indoors faces difficulties because GNSS signals are weakened or distorted by walls, ceilings, and other obstacles. These factors cause signal attenuation, reflections, and the absence of a direct satellite connection, which makes it hard to accurately determine user/asset position inside buildings. WiFi Fingerprinting, among various indoor positioning techniques, utilizes the pervasive setup of WiFi access points (APs) in modern buildings [1, 2]. Thus, it has become a promising and practical solution that utilizes existing infrastructure without additional hardware. The

technique involves two main phases: an offline phase, where a site survey is performed to gather received signal strength (RSS) values at known coordinates, so-called reference points (RPs), resulting in an offline fingerprint database, and an online phase, where an unknown position is estimated by comparing the new set of RSS values to the database.

Despite its practicality, WiFi Fingerprinting-based positioning systems are still experiencing constraints such as the high dimensionality of WiFi data, the sparse AP availability, and the temporal instability caused by environmental changes, user mobility, and interference. Traditional approaches, like deterministic and probabilistic, have shown limited ability to capture the nonlinear relationships between RSS patterns and physical coordinates. In contrast, deep learning based-approaches have been utilized to improve the positioning accuracy by learning complex RSS value-position relationships.

Recently, Kolmogorov-Arnold Network (KAN) [3], a new type of neural architecture inspired by the Kolmogorov-Arnold representation theorem, has been introduced to the research community. According to the theorem, any continuous function with multiple variables can be represented as a finite combination of continuous univariate functions connected through addition. Based on this concept, rather than relying on static weighted edges as in multilayer perceptrons (MLPs), KAN utilizes learnable univariate activation functions, often implemented as B-splines, along the connections. This approach moves the learning process from fixed weights to flexible, data-driven mappings between variables. As a result, KAN provides strong interpretability, efficient use of parameters, and enhanced modeling flexibility in function approximation, making it particularly effective for tasks with low-dimensional input spaces.

Up to now, various versions of KAN have been proposed to adapt the architecture to different application domains [4, 5]. Each model involves a trade-off among accuracy, computational efficiency, resistance to noise, and flexibility in handling heterogeneous data distributions. For the WiFi Fingerprinting-based positioning systems, KAN-based models have become potential solutions since the per-edge

learnable activation functions enable the modeling of WiFi signal fluctuations caused by obstacles, multipath effects, and environmental variability. Some KAN-based models can adapt to spatial heterogeneity, e.g., differences in signal propagation between floors or buildings, due to the ability to integrate different basis functions [6] or dynamically select function spaces [7]. In addition, some models have been optimized for computational efficiency, which allows them to perform in real time even on resource-constrained platforms like mobile devices [8, 9].

Nonetheless, a comparative evaluation of KAN-based models for WiFi Fingerprinting-based positioning systems is still lacking, particularly in determining which model achieves the lowest positioning errors in diverse and complicated datasets. In this paper, motivated by architectural diversity and practical suitability for positioning tasks, we conduct a comparative empirical evaluation of five KAN-based models applied in the WiFi Fingerprinting technique, including original KAN, FastKAN [8], BSRBF-KAN [6], S-KAN [7], and FC-KAN [10]. The key contributions of this work are summarized as follows.

- We implement the five KAN-based models, which are introduced to address different aspects of the original KAN's limitations.
- We evaluate the five models in three public datasets, including UJIIndoorLoc, HCXY, and SYL, which were collected in varied and complex real-world settings.
- Based on empirical results and theoretical analysis, we identify the most effective and robust model that achieves the lowest positioning error under varying indoor environments.

The remainder of this paper is structured as follows: Section 2 provides a review of related studies on WiFi fingerprinting and KAN-based models applied to positioning. Section 3 details the implemented KAN-based models that are used for evaluation. Section 4 presents and discusses the experimental results, while Section 5 concludes the study and offers directions for future work.

II. RELATED WORK

A. WiFi Fingerprinting-based Positioning Systems

Over the last two decades, WiFi Fingerprinting has been extensively researched as a highly promising indoor positioning technique. Two conventional approaches, deterministic and probabilistic, are commonly applied in this technique to determine the user's position. The first approach, with the representatives as the nearest neighbors-based methods [11, 12], offers speed and simplicity, making them attractive for lightweight deployments. Even so, this approach often lacks robustness in noisy and dynamic environments. The second approach, even though it provides better modeling of uncertainty and can obtain higher positioning accuracy under variable conditions [13, 14], requires more extensive data collection and computational resources.

Recently, with the rise of deep learning, neural networks have been increasingly applied to WiFi Fingerprinting [15, 16]. MLPs, by learning nonlinear feature transformations, showed promising positioning results [17, 18]. Meanwhile, Recurrent Neural Networks attempted to model temporal dependencies in complicated positioning areas, especially

when the user was walking [19, 20]. In addition, Convolutional Neural Networks, the technique that dealt with grid-like data, treated RSS maps as 2D images to capture spatial correlations between APs [21, 22]. Nevertheless, these models use fixed activation functions and linear connections, which may not efficiently capture the complex characteristics in WiFi data correlated to physical RPs.

B. KAN-based Models in Indoor Positioning

The above trade-off leads to the need to combine the nonlinear function approximation capability of neural networks with the statistical adaptability of probabilistic reasoning, which is presented in KAN. Although recently introduced, KANs have been applied in various research topics, such as intrusion detection [23, 24], image classification [25, 26], human activity recognition [27], etc. For indoor positioning, KAN-based models have also been applied to enhance the tracking accuracy. In [28], by utilizing spatial magnetic field characteristics, Gao et al. proposed the MP-KAN algorithm to reduce the error in near-field predictions and balance multi-axis errors. From the experimental results, the maximum error was only 0.24 mm. Moreover, KAN was also applied in a Visible Light-based positioning system, which also demonstrated sub-centimeter accuracy, even with tilted light-emitting diodes [29]. In [30], Wang et al. introduced the CCKAN, which integrated the complex convolutional networks with KAN, to improve Angle of Arrival estimation in near-field scenarios, where far-field assumptions broke down and affected the accuracy. The simulation demonstrated that CCKAN consistently surpassed conventional algorithms in various conditions.

For WiFi-based positioning, only a few published studies have used KAN-based models to improve the positioning accuracy. Yan et al. [31] proposed RSLoc with the KAN convolution network to generate stable fingerprint mappings, which incorporated some optimizations, such as removing invalid APs and applying entropy regularization. The positioning results surpassed those of other methods, achieving up to a 48% improvement in accuracy after eliminating invalid APs. Furthermore, in [32], Krisnawan et al. introduced a hybrid RBF-KAN model for accurate indoor positioning using dense grid WiFi Fingerprinting. The model attempted to handle non-linearities and adapt to high-dimensional functions. Compared to other methods, the proposed model achieved a mean positioning error of only 0.105 m, with a positioning accuracy of 98.26%.

It can be seen that KAN-based models are suitable for WiFi Fingerprinting-based positioning systems due to learnable univariate functions on edges for modeling highly nonlinear, location-specific RSS values-position relationships. In contrast, the positioning performance of different KAN-based models over diverse datasets remains uncertain. Therefore, this work implements and evaluates the performance of various KAN-based models to find the best model that achieves the lowest positioning errors while maintaining a reasonable computational cost.

III. KAN-BASED MODELS

As shown in Fig. 1, all the KAN-based models are built upon the Kolmogorov-Arnold representation theorem. The original KAN is the first model to implement the above concept in neural networks. Its most unique feature is the replacement of static weights with learnable one-dimensional functions, i.e., B-splines, which allows each edge to act as a

nonlinear, adaptive transformation. This model opens a high level of interpretability, however, one major limitation of the original KAN is its high computational cost caused by the use of B-splines and the requirement for grid rescaling during training.

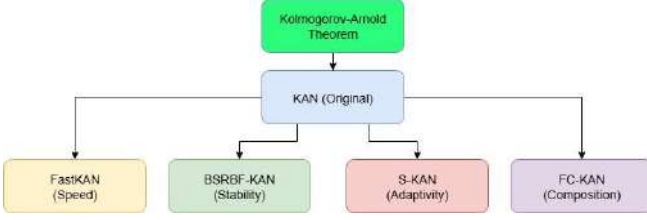


Fig. 1. Development of KAN-based models

Since the introduction of the original KAN, several KAN-based models have been proposed to cover its limitations, such as high computational complexity, excessive memory usage, limited domain adaptability, scalability constraints, etc. Therefore, this study considers four KAN-based models to handle the aforementioned issues, as presented in Fig. 1: FastKAN, which enhances computational speed and efficiency; BSRBF-KAN, which improves stability and generalization; S-KAN, which increases adaptivity; and FC-KAN, which augments compositional feature learning.

FastKAN is proposed to reduce the high computational cost of the original KAN by approximating its B-spline basis functions with Gaussian Radial Basis Functions (RBFs). This simplification converts the model into a form of RBF networks, offering lower computational demands and simpler implementation. FastKAN speeds up forward-pass computations by over three times through the removal of de Boor-Cox iterations and adaptive grid rescaling.

Meanwhile, BSRBF-KAN is proposed to enhance model stability and learning convergence. The model linearly combines B-splines and RBFs at each edge, which utilizes the smoothness and local control of splines together with the adaptability and resilience of RBFs. Thus, it maintains reliable performance while significantly reducing computational demands.

Next, S-KAN is suggested to allow each node to dynamically select the most appropriate activation function space during training, instead of assigning one activation function per edge or node. In practical terms, the model offers a significant improvement in generalization and task-specific performance. Nevertheless, S-KAN introduces substantial complexity, requiring extra selection mechanisms, possible use of multi-objective loss functions, and higher computational demands during both training and inference.

Lastly, FC-KAN generalizes the KAN model by incorporating multiple basis functions and mathematical operations at the output of each edge or node. Specifically, it supports various function combinations, such as element-wise addition, multiplication, quadratic transformations, Difference of Gaussians (DoG), etc. These operations are mainly applied in low-dimensional layers to manage computational costs. FC-KAN aims to enhance feature richness and pattern recognition by modeling higher-order interactions among input features. Nevertheless, the limitation of this model is the computational complexity. The main characteristics of the five KAN-based models are summarized in Table I.

TABLE I. COMPARISON OF KAN-BASED MODELS

Models	Characteristics
KAN	Learnable activation functions on edges (splines), no linear weights
FastKAN	Speed up KAN by replacing B-splines with Gaussian RBFs
BSRBF-KAN	Combine B-splines with RBFs in activation functions to improve stability and convergence
S-KAN	Nodes adaptively select the best activation function during learning
FC-KAN	Use various combinations of functions (sum, product, concat) to capture deeper representations

Among the five models, KAN is the simplest one since it is the most basic model. In contrast, FC-KAN is the most complex model since it incorporates multiple mathematical function combinations in low-dimensional layers, requiring careful handling of tensor operations and computational overhead. FastKAN is a bit more complicated than KAN by replacing B-splines with computationally leaner RBFs, followed by BSRBF-KAN, since this model merges B-splines and RBFs. Lastly, S-KAN is simpler than FC-KAN, which introduces per-node adaptivity, requiring dynamic selection and additional control logic.

Fig. 2 illustrates the application of KAN-based models to WiFi Fingerprinting to determine a user's position. The input layer involves the nodes, which are RSS values, i.e., fingerprints, collected from the available APs at different RPs. Each connection between nodes is substituted with a trainable univariate function ϕ , which allows flexible nonlinear transformations. The function's equation differs among the various KAN-based models, as detailed in the respective references. These functions are structured hierarchically across KAN layers, with multiple transformations composed to sequentially model complex RSS value-position relationships. The green circular nodes present the number of units after each transformation stage, e.g., 512, 256. In the output layer, the transformed features are projected onto the user's 2D position coordinates.

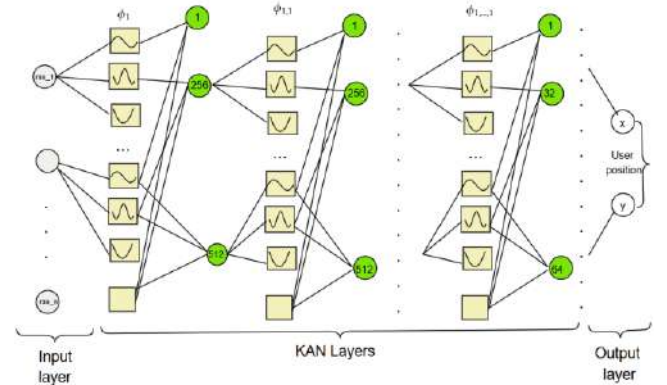


Fig. 2. Structure of the KAN-based models applied to WiFi Fingerprinting

IV. EXPERIMENTAL RESULTS

A. Datasets

In this work, three WiFi Fingerprinting public datasets are used for performance evaluation: UJIIndoorLoc [33], HCYX [34], and SYL [34]. The first dataset was collected within three different buildings spanning multiple floors at Jaume I University (Spain), while the two other datasets were collected from two buildings' floors in Xuzhou and Jinan cities (China). These datasets were collected with the participation of many

volunteers as well as many smartphones. Each dataset includes the training and testing files, but the number of WiFi data collected for each file is different. The more specific characteristics for the three datasets are presented in Table II. It should be noted that, in this paper, since we do not focus on building classification, the UJIIndoorLoc dataset is split into three sub-datasets corresponding to three buildings.

TABLE II. MAIN CHARACTERISTICS OF THREE DATASETS

Characteristics	Datasets		
	<i>UJIIndoorLoc</i>	<i>HCTX</i>	<i>SYL</i>
Total areas (m ²)	108,703	3,600	2,600
Total number of RPs	735	379	296
Distance between adjacent RPs (m)	~3.65	1.2	1.2
Samples per RP	≥ 2	30	30
Total number of APs	520	347	363

B. Results

The positioning performance of the five KAN-based models is evaluated using the mean positioning error as the comparison metric. This metric is calculated as the average Euclidean distance between the actual test points and their estimated positions. The model parameters used in the experiments are listed in Table III.

TABLE III. THE PARAMETER SETTING FOR KAN-BASED MODELS

Parameters		Values
Hidden Layers		[512, 512, 256, 256, 128, 128, 64, 64]
Model	KAN	SplineLinear
	FastKAN	RBF, SplineLinear
	BSRBF-KAN	func_list=[SiLU, RBF], grid_size=4, spline_order=3
	S-KAN	func_list=[shifted_softplus]
	FC-KAN	func_list=['dog', 'B-splines'], grid_size=5, spline_order=3
Optimizer		Adam (Learning rate=0.001)
Batch Size		64

Fig. 3 to Fig. 5 display the positioning errors of the five KAN-based models on the three datasets UJIIndoorLoc, HCTX, and SYL, respectively. As shown in Fig. 3, the BSRBF-KAN model demonstrates the highest performance among the five models when evaluated on the UJIIndoorLoc dataset, with a positioning error of 8.43 m, which is lower than the others by 24.59% to 46.91%. These results demonstrate the effectiveness of BSRBF-KAN in WiFi Fingerprinting. On the UJIIndoorLoc dataset, previously published studies applying deep neural networks [35, 36] or CNNs [37, 38] reported positioning errors ranging from 9.29 m to 14.93 m. In comparison, BSRBF-KAN achieves lower errors, which represents an improvement of over 9.26% relative to these methods on the same dataset. This result can come from the hybrid integration of B-splines and RBFs, which strengthens the network's capacity to capture both local and global WiFi signal patterns, resulting in improved robustness to

environmental changes. S-KAN is the runner-up model with an error of 11.18 m, followed by FastKAN with slightly higher error. In contrast, the highest positioning error belongs to the original KAN, with an error of approximately 16 m, which proves that this model lacks optimizations for dealing with the highly fluctuating RSS values. Furthermore, FC-KAN also performs poorly with an error of 15.18 m, showing that its compositional function approach may not be suitable for the WiFi Fingerprinting technique.

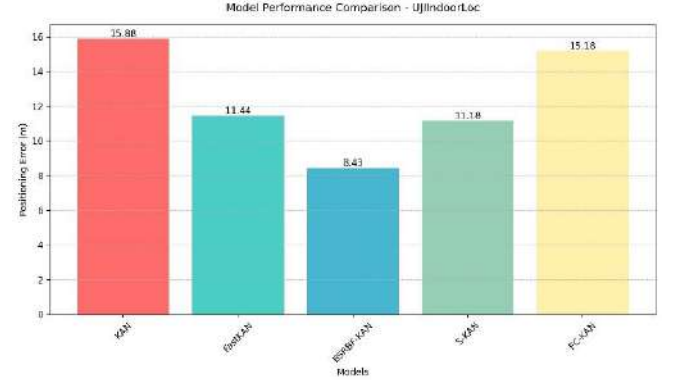


Fig. 3. Mean positioning errors of five models for the UJIIndoorLoc dataset

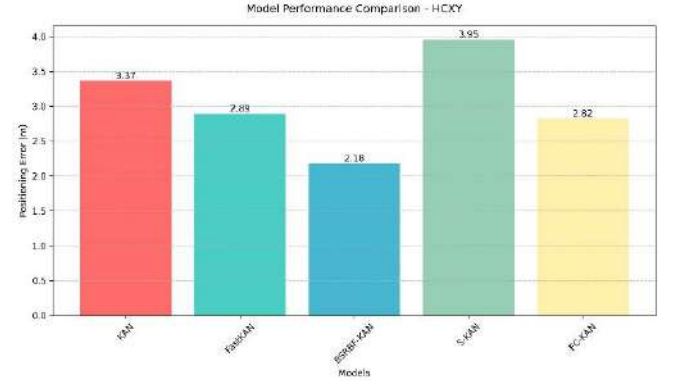


Fig. 4. Mean positioning errors of five models for the HCTX dataset

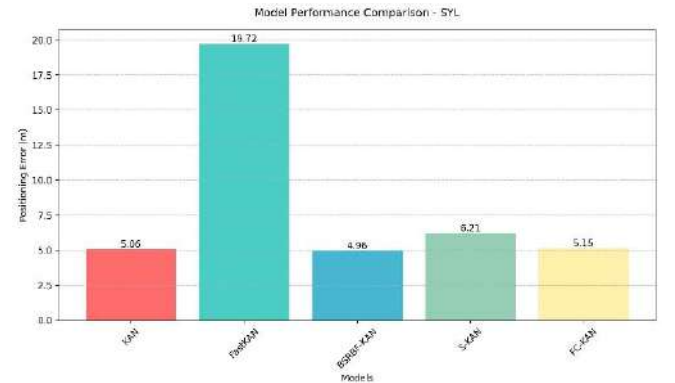


Fig. 5. Mean positioning errors of five models for the SYL dataset

For Fig. 4 and Fig. 5, even when testing the models on the two datasets HCTX and SYL, the BSRBF-KAN still achieves the best performance with the lowest positioning error of 2.18 m and 4.96 m. The results confirm its strong adaptability over various datasets due to its hybrid B-spline and RBFs. Conversely, the worst results in the HCTX and SYL datasets fall under the S-KAN and FastKAN, respectively. This indicates that the other KAN-based models do not consistently achieve the same performance over various datasets,

excluding the BSRBF-KAN. The poor results of FastKAN on the SYL dataset reveal that prioritizing computational speed can reduce representational capability in noisy environments. Meanwhile, the adaptive mechanism of S-KAN may become unstable when AP distributions are highly variable.

Moreover, we apply the Random Forest (RF) algorithm, which is effective for feature selection and serves as an indirect form of dimensionality reduction, to the offline fingerprint database to identify the most significant APs. In the database, each AP plays a role as a feature. However, there exist APs that are either invisible in most locations or provide noisy signals. To address this issue, RF can compute feature importance to rank APs and remove the least informative ones. This may help improve positioning accuracy while reducing dimensionality, noise, and computational cost. Table IV shows the positioning errors of the five KAN-based models when applied to the dimensionally reduced fingerprint database by using the RF algorithm. As can be seen from the table, all models generally obtain lower positioning errors compared to the case without applying the RF algorithm, which demonstrates the effectiveness of RF for selecting meaningful features, i.e., APs. To be specific, the BSRBF-KAN is still superior to other KAN-based models since it achieves the lowest positioning errors with 8.03 m, 1.94 m, and 3.89 m in UJIIndoorLoc, HCXY, and SYL datasets, respectively. These results once again prove the high performance of the BSRBF-KAN model, even when considering the dimensionality reduction.

TABLE IV. MEAN POSITIONING ERRORS OF FIVE MODELS WHEN COMBINING WITH THE RF ALGORITHM

Models/Errors (m)	Datasets		
	<i>UJIIndoorLoc</i>	<i>HCXY</i>	<i>SYL</i>
KAN	15.49	3.16	4.67
FastKAN	10.79	2.56	7.74
BSRBF-KAN	8.03	1.94	3.89
S-KAN	10.65	3.56	5.48
FC-KAN	14.31	2.92	5.15

Overall, over the aforementioned results from different settings, BSRBF-KAN consistently ranks first with the lowest mean positioning error, hence, the model demonstrates its robustness when applied to WiFi Fingerprinting. The other models can achieve reasonable results in one dataset, but worse results in other ones, which suggests that those models may not generalize equally well in various scenarios without tuning. BSRBF-KAN may outperform other models due to superior robustness to noise and reduced overfitting. The model combines B-spline functions for precise local adaptability with RBFs for global smoothness, which enables the ability to capture both fine-grained RSS fluctuations and large-scale spatial patterns. Thus, it effectively models the nonlinear relationship between the WiFi signal and position. Nonetheless, accuracy is not the only factor that determines the suitability of a model for real-world deployment. BSRBF-KAN, although the most accurate, suffers from high computational cost, especially when compared to FastKAN, since it introduces dual basis computations that increase parameter count and training overhead. This illustrates the balance that needs to be maintained between accuracy and computational efficiency.

V. CONCLUSION

This study presented a comprehensive evaluation of five KAN-based models applied to WiFi Fingerprinting. Experiments on three public datasets demonstrated that KAN-based models can effectively deal with the nonlinear, environment-dependent characteristics of WiFi Fingerprinting-based indoor positioning systems. The results confirmed the idea that integrating hybrid basis functions into KAN architectures can lead to more stable and accurate positioning in diverse indoor environments. Among the five implemented KAN-based models, BSRBF-KAN was the best model since it consistently achieved the lowest positioning errors due to its hybrid basis function design that captured both local and global patterns in WiFi data. Future work may concentrate on implementing and testing the models in real-world applications. Moreover, the integration of WiFi with other technologies, such as Bluetooth, inertial measurement units, etc., should be investigated to further enhance positioning accuracy and stability.

ACKNOWLEDGMENT

This work was partly supported by the Ministry of Education and Training under Grant B2024-DLA-04 and Dalat University under Grant CTTD3-2024.

REFERENCES

- [1] N. Singh, S. Choe, and R. Punmiya, "Machine learning based indoor localization using Wi-Fi RSSI fingerprints: An overview," *IEEE Access*, vol. 9, pp. 127150-127174, 2021.
- [2] S. Shang and L. Wang, "Overview of WiFi fingerprinting-based indoor positioning," *IET Communications*, vol. 16, no. 7, pp. 725-733, 2022.
- [3] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "KAN: Kolmogorov-Arnold Networks," *arXiv preprint*, arXiv:2404.19756, 2024.
- [4] S. Somvanshi, S. A. Javed, M. M. Islam, D. Pandit, and S. S. Das, "A survey on Kolmogorov-Arnold Network," *arXiv preprint*, arXiv:2411.06078, 2024.
- [5] T. Ji, Y. Hou, and D. Zhang, "A comprehensive survey on Kolmogorov-Arnold Networks (KAN)," *arXiv preprint*, arXiv:2407.11075, 2024.
- [6] H.-T. Ta, "BSRBF-KAN: A combination of B-splines and radial basis functions in Kolmogorov-Arnold Networks," *arXiv preprint*, arXiv:2406.11173, 2024.
- [7] Z. Yang, J. Zhang, X. Luo, Z. Lu, and L. Shen, "Activation space selectable Kolmogorov-Arnold Networks," *arXiv preprint*, arXiv:2408.08338, 2024.
- [8] Z. Li, "Kolmogorov-Arnold networks are radial basis function networks," *arXiv preprint*, arXiv:2405.06721, 2024.
- [9] S. Patra et al., "Physics-informed Kolmogorov-Arnold neural networks for dynamical analysis via Efficient-KAN and WAV-KAN," *arXiv preprint*, arXiv:2407.18373, 2024.
- [10] H.-T. Ta, D.-Q. Thai, A. B. S. Rahman, G. Sidorov, and A. Gelbukh, "Fc-kan: Function combinations in Kolmogorov-Arnold Networks," *arXiv preprint*, arXiv:2409.01763, 2024.
- [11] N. Duong-Bao, J. He, L. N. Thi, and K. Nguyen-Huu, "Analysis of Distance Measures for WiFi-based Indoor Positioning in Different Settings," in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2022, pp. 1-7.
- [12] D. Pham-Hai, N. Duong-Bao, J. He, L. Nguyen Thi, S.-W. Lee, and K. Nguyen-Huu, "WiFi-based Positioning System with k-means Clustering and Outlier Removal: Evidence from Multiple Datasets," in *Proceedings of the 12th International Symposium on Information and Communication Technology*, 2023, pp. 980-988.
- [13] B. Huang, R. Yang, B. Jia, W. Li, and G. Mao, "A theoretical analysis on sampling size in WiFi fingerprint-based localization," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3599-3608, 2021.

- [14] D. Yu and C. Li, "An accurate WiFi indoor positioning algorithm for complex pedestrian environments," *IEEE Sensors Journal*, vol. 21, no. 21, pp. 24440-24452, 2021.
- [15] Y. Lin, K. Yu, F. Zhu, J. Bu, and X. Dua, "The state of the art of deep learning-based Wi-Fi indoor positioning: A review," *IEEE Sensors Journal*, vol. 24, no. 17, pp. 27076-27098, 2024.
- [16] O. Kerdjadj et al., "Uncovering the potential of indoor localization: Role of deep and transfer learning," *IEEE Access*, vol. 21, pp. 73980-74010, 2024.
- [17] E. Cakan, A. Şahin, M. Nakip, and V. Rodoplu, "Multi-layer perceptron decomposition architecture for mobile IoT indoor positioning," in *Proc. 7th IEEE World Forum on Internet of Things (WF-IoT)*, 2021, pp. 379.
- [18] E. Scavino, M. A. Abd Rahman, and Z. Farid, "An improved hybrid indoor positioning algorithm via QPSO and MLP signal weighting," *Computers, Materials & Continua*, vol. 74, no. 1, 2023.
- [19] M. T. Hoang, B. Yuen, X. Dong, T. Lu, R. Westendorp, and K. Reddy, "Recurrent neural networks for accurate RSSI indoor localization," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10639-10651, 2019.
- [20] Y. Dong, T. Arslan, and Y. Yang, "An encoded LSTM network model for WiFi-based indoor positioning," in *Proc. IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2022, pp. 1-6.
- [21] A. B. Mazlan, Y. H. Ng, and C. K. Tan, "A fast indoor positioning using a knowledge-distilled convolutional neural network (KD-CNN)," *IEEE Access*, vol. 10, pp. 65326-65338, 2022.
- [22] A. Alitaleshi, H. Jazayeriy, and J. Kazemitabar, "EA-CNN: A smart indoor 3D positioning scheme based on Wi-Fi fingerprinting and deep learning," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105509, 2023.
- [23] M. Abd Elaziz, I. A. Fares, and A. Aseeri, "CKAN: Convolutional Kolmogorov–Arnold networks model for intrusion detection in IoT environment," *IEEE Access*, vol. 12, pp. 134837-134851, 2024.
- [24] Z. Wang, A. Zainal, M. M. Siraj, F. A. Ghaleb, X. Hao, and S. Han, "An intrusion detection model based on convolutional Kolmogorov–Arnold networks," *Scientific Reports*, vol. 15, no. 1, p. 1917, 2025.
- [25] A. Jamali, S. K. Roy, D. Hong, B. Lu, and P. Ghamisi, "How to learn more? Exploring Kolmogorov–Arnold networks for hyperspectral image classification," *Remote Sensing*, vol. 16, no. 21, p. 4015, 2024.
- [26] M. J. Cheon, "Kolmogorov–Arnold network for satellite image classification in remote sensing," *arXiv preprint, arXiv: 2406.00600*, 2024.
- [27] M. Liu, D. Geißler, D. Nshimiyimana, S. Bian, B. Zhou, and P. Lukowicz, "Initial investigation of Kolmogorov–Arnold networks (KANs) as feature extractors for IMU-based human activity recognition," in *Proc. Companion of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 500–506.
- [28] Z. Gao and M. M. Kong, "MP-KAN: An effective magnetic positioning algorithm based on Kolmogorov–Arnold network," *Measurement*, vol. 243, p. 116248, 2024.
- [29] F. Seguel, D. Salihu, S. Hägele, and E. Steinbach, "VLP-KAN: Low-complexity and interpretable RSS-based VLP using Kolmogorov–Arnold Networks," in *Proc. 20th Wireless On-Demand Network Systems and Services Conference (WONS)*, 2025, pp. 1–8.
- [30] J. Wang, D. Xiao, Y. Cao, K. Ying, and M. Xiao, "Near-field AoA estimation with Complex Convolutional Kolmogorov–Arnold Network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [31] W. Yan, C. Tang, L. Zhu, Y. Li, G. Li, and H. Yang, "RSLoc: An accuracy indoor localization system based on KAN convolution network," in *Proc. IEEE International Conference on High Performance Computing and Communications (HPCC)*, 2024, pp. 1004–1010.
- [32] A. B. Krisnawan, P. H. Mukti, and M. H. Purnomo, "RBF-KAN: Integrated approach for accurate indoor localization in dense grid RSSI fingerprint," in *17th International Conference on Knowledge and Smart Technology (KST)*, 2025, pp. 46–51.
- [33] J. Torres-Sospedra et al., "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proc. International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2014, pp. 261–270.
- [34] J. Bi, Y. Wang, B. Yu, H. Cao, T. Shi, and L. Huang, "Supplementary open dataset for WiFi indoor localization based on received signal strength," *Satellite Navigation*, vol. 3, no. 1, pp. 1–15, 2022.
- [35] M. Kim, D. Han, and J.-K. Rhee, "Unsupervised view-selective deep learning for practical indoor localization using CSI," *IEEE Sensors Journal*, vol. 21, no. 21, pp. 24398-24408, 2021.
- [36] J. Cha and E. Lim, "A hierarchical auxiliary deep neural network architecture for large-scale indoor localization based on Wi-Fi fingerprinting," *Applied Soft Computing*, vol. 120, p. 108624, 2022.
- [37] X. Song et al., "A novel convolutional neural network based indoor localization framework with WiFi fingerprinting," *IEEE Access*, vol. 7, pp. 110698-110709, 2019.
- [38] F. Qin, T. Zuo, and X. Wang, "CCpos: WiFi fingerprint indoor positioning system based on CDAE-CNN," *Sensors*, vol. 21, no. 4, p. 1114, 2021.

A multi-sensor fusion approach to sensor management using void probability

Minh Ngoc Duong

VNU University of Engineering and Technology
Hanoi, Vietnam
ngocdm.uet@vnu.edu.vn

Gia Hung Hoang

VNU University of Engineering and Technology
Hanoi, Vietnam
hunggh@vnu.edu.vn

Abstract—Multi-sensor management in multi-target environments is a highly active research area with numerous real-world applications. This paper addresses the multi-target, multi-sensor management problem by integrating Random Finite Set (RFS) theory within the Partially Observable Markov Decision Process (POMDP) framework. The proposed approach leverages the Iterated-Corrector Probability Hypothesis Density (IC-PHD) filter to sequentially incorporate observations from multiple sensors, enabling recursive propagation of the multi-target posterior intensity. This posterior is then used to derive a reward function based on the void probability, which allows efficient numerical evaluation and practical computation of optimal sensor control. Extensive numerical simulations are conducted to validate the effectiveness and scalability of the proposed method.

Index Terms—Multi-sensor management, sensor fusion, PHD filter, void probability.

I. INTRODUCTION

Sensor management for multi-sensor multi-target systems is a highly active research area, with broad implications for surveillance [1], autonomous vehicles [2], and robotics [3]. It presents an optimal, nonlinear control problem, requiring dynamic allocation of limited sensing resources to effectively track multiple moving targets over time while handling random variations in target numbers and measurements due to births, deaths, misses, and false alarms. As a result, the system state and sensor observations must be modeled as finite sets rather than fixed-dimensional vectors, rendering conventional optimal control techniques inapplicable [4].

A principled way to address the stochastic nature of multi-target systems is to model states and measurements as finite-set-valued random variables using Random Finite Set (RFS) theory [5], which enables efficient propagation of the posterior multi-object distribution with RFS-based filters. Building on this, the sensor management problem can be formulated as a Partially Observable Markov Decision Process (POMDP) [6], where the posterior provides the basis for defining a reward function, allowing optimal sensor control actions to be computed at each time step by maximizing the expected reward.

Several works have applied RFS theory to multi-target sensor control within the POMDP framework, using RFS-based filters such as Bayes [7], PHD [8], CB-MeMBer [9], and GLMB [10]. However, these are limited to single-sensor multi-target control problems, which suffer from low observability

in challenging scenarios due to restricted field of view and homogeneous sensing [9]. Moreover, most rely on information-theoretic reward functions such as Rényi or Cauchy–Schwarz divergences [8], [10]. While Rényi lacks closed-form solutions for PHD filters, Cauchy–Schwarz can be computed analytically via GM-PHD filter [11]. However, optimizing such reward functions only maximizes global information gain and may degrade performance in pathological cases as it may not directly account for the physical conditions of the system [9].

This paper adopts the Gaussian Mixture multi-sensor PHD filter [12] to improve tracking accuracy and robustness through efficient data fusion from multiple sensors. Unlike existing approaches that rely on information-theoretic reward functions, we introduce a novel reward function based on the void probability. A key advantage of this formulation is that it admits a tractable numerical evaluation under the Gaussian Mixture framework, enabling efficient computation. Moreover, the void probability provides a direct measure of the likelihood that no targets are present within a sensor’s field of view. By minimizing this void probability, the reward function directly promotes target detection, offering a more intuitive and aligned metric for decision-making in complex multi-target tracking environments, particularly for range-based sensing.

The organization of the paper is as follows. Section II reviews general multi-sensor PHD filter and iterated-corrector multi-sensor PHD filter. The problem formulation for multi-sensor management using void probability as the reward function is introduced in section III. Section IV presents simulation results, and finally, section V concludes the paper.

II. MULTI-SENSOR MULTI-TARGET TRACKING WITH PHD FILTER

This section outlines a multi-sensor fusion framework for the PHD filter, beginning with an introduction to the two-sensor PHD filter in subsection II-A. Subsection II-B subsequently presents the computationally efficient Iterated-Corrector PHD (IC-PHD) filter as an approximation.

A. General Multi-sensor PHD Filter

In this subsection, we summarize the results of [13], which establish the formulation for the general multi-sensor PHD filter for the simplest case of incorporating measurements from two sensors into a unified Bayesian update.

This filter relies on several standard assumptions within the finite set statistics (FISST) framework. First, each target generates at most one measurement per sensor, and each measurement originates from a single source. Second, both target-generated measurements and clutter are modeled as Poisson point processes, with clutter assumed to be independent across sensors. Third, measurement sets from different sensors are conditionally independent given the multi-target state. These assumptions enable a tractable formulation and ensure statistical consistency in the joint update process.

At time step $k + 1$, let $Z_{k+1}^{(1)}$ and $Z_{k+1}^{(2)}$ denote the measurement sets of sensor 1 and sensor 2, respectively. The combined measurement set is formed by their disjoint union $Z_{k+1} = Z_{k+1}^{(1)} \uplus Z_{k+1}^{(2)}$. The updated PHD is given by [13]:

$$D_{k+1|k+1}(x) = L_{Z_{k+1}^{(1)}, Z_{k+1}^{(2)}}(x) \cdot D_{k+1|k}(x), \quad (1)$$

where $D_{k+1|k}(x)$ is the predicted PHD at time $k + 1$, and $L_{Z_{k+1}^{(1)}, Z_{k+1}^{(2)}}(x)$ is the multi-sensor likelihood multiplier, defined by:

$$L_{Z_{k+1}^{(1)}, Z_{k+1}^{(2)}}(x) = 1 - \tilde{\pi}_D(x) + \sum_{\mathcal{P} \vdash Z_{k+1}} \omega_{\mathcal{P}} \sum_{W \uplus \dot{W} \in \mathcal{P}} \left[\frac{L_W(x) \cdot \dot{L}_{\dot{W}}(x)}{\delta_{|W|,0} \cdot \kappa_W + \delta_{|\dot{W}|,0} \cdot \dot{\kappa}_{\dot{W}} + \tau_{W,\dot{W}}} \right] \quad (2)$$

In equation (2), $L_W(x)$ and $\dot{L}_{\dot{W}}(x)$ are the multi-target likelihoods for subsets $W \subseteq Z_{k+1}^{(1)}$ and $\dot{W} \subseteq Z_{k+1}^{(2)}$, respectively. Similarly, κ_W and $\dot{\kappa}_{\dot{W}}$ denote the clutter intensities for sensor 1 and sensor 2, while $\delta_{|W|,0}$ and $\delta_{|\dot{W}|,0}$ are Kronecker deltas, which equal 1 if the corresponding subset is empty, and 0 otherwise. The joint likelihood integral $\tau_{W,\dot{W}}$ captures the agreement of the measurement subsets with the predicted density and is defined by:

$$\tau_{W,\dot{W}} = \int L_W(x) \cdot \dot{L}_{\dot{W}}(x) \cdot D_{k+1|k}(x) dx. \quad (3)$$

Furthermore, the scalar $\omega_{\mathcal{P}}$ denotes the normalized weight for a specific partition $\mathcal{P} \vdash Z_{k+1}$, where $\vdash$ indicates a valid partitioning of the combined measurement set into subsets corresponding to clutter and target-originated measurements. As shown in [13], this weight is calculated as:

$$\omega_{\mathcal{P}} = \frac{\prod_{W \uplus \dot{W} \in \mathcal{P}} (\delta_{|\dot{W}|,0} \cdot \kappa_W + \delta_{|W|,0} \cdot \dot{\kappa}_{\dot{W}} + \tau_{W,\dot{W}})}{\sum_{\mathcal{Q} \vdash Z_{k+1}} \prod_{V \uplus \dot{V} \in \mathcal{Q}} (\delta_{|\dot{V}|,0} \cdot \kappa_V + \delta_{|V|,0} \cdot \dot{\kappa}_{\dot{V}} + \tau_{V,\dot{V}})} \quad (4)$$

Finally, the term $\tilde{\pi}_D(x) = 1 - (1 - \pi_D(x))(1 - \dot{\pi}_D(x))$ represents the probability that a target located at state x is detected by at least one of the two sensors. Here, $\pi_D(x)$ and $\dot{\pi}_D(x)$ denote the probabilities that sensor 1 and sensor 2 collect at least one measurement, respectively.

The general multi-sensor PHD filter requires evaluating all possible joint partitions of the combined measurement set from all sensors. Although optimal, its cost grows rapidly with the number of sensors and measurements, making it impractical

for real-time applications and approximations necessary in practice to balance accuracy and efficiency.

B. Iterated-Corrector Multi-sensor PHD Filters

To address the computational challenges of the general multi-sensor PHD filter, a widely used approximation is the Iterated-Corrector PHD (IC-PHD) filter [13]. This approach sacrifices full joint optimality in favor of computational efficiency by approximating the joint update as a sequence of single-sensor PHD updates, applied iteratively across sensors.

Let $D_{k+1|k}(x)$ be the predicted PHD at time $k + 1$. For an s -sensor system with measurement sets $Z^{(1)}, Z^{(2)}, \dots, Z^{(s)}$, the IC-PHD filter approximates the joint update as:

$$D_{k+1|k+1}(x) \approx \mathcal{L}_{Z^{(s)}} \circ \dots \circ \mathcal{L}_{Z^{(2)}} \circ \mathcal{L}_{Z^{(1)}} (D_{k+1|k}(x)), \quad (5)$$

where $\mathcal{L}_{Z^{(i)}}(\cdot)$ denotes the standard single-sensor PHD correction corresponding to sensor i . The symbol $\circ$ represents function composition, indicating that the updates are applied sequentially. The procedure begins by applying the update for the first sensor to the predicted PHD. The resulting intermediate density serves as the prior for the second sensor update, and so on, until all sensors are processed in sequence.

This sequential correction scheme avoids the need to evaluate all joint partitions, thus substantially reducing computational complexity. However, it introduces certain limitations. Most notably, the update is not order-invariant, meaning the final result may depend on the sequence of sensor updates. Additionally, each update step assumes the PHD remains a Poisson process, even though this assumption is no longer valid after the first update. Despite these drawbacks, the IC-PHD filter performs well in practice, particularly when the sensors have similar detection characteristics [12]. In such cases, the effect of sensor ordering is minor, and the filter offers a favorable trade-off between computational efficiency and estimation accuracy.

III. PROPOSED FORMULATION FOR MULTI-SENSOR MANAGEMENT

This section presents the main contribution of the paper: a fusion-based formulation for multi-sensor management within the POMDP framework. The proposed approach is attractive in that it leverages the IC-PHD filter to propagate the multi-target posterior intensity, which is then used to evaluate a reward function derived from the void probability.

A. The Gaussian mixture formulation of the IC-PHD filter

In this subsection, we adopt the approach introduced in [12], which represents the multi-target PHD as a weighted sum of Gaussian components, to derive closed-form expressions for both the prediction and update steps of the IC-PHD filter under linear Gaussian assumptions for multi-target transition and observation models.

Assume that the posterior PHD function at time step $k-1$ is represented as a Gaussian mixture:

$$D_{k-1|k-1}(x) = \sum_{j=1}^{J_{k-1}} w_{k-1}^{(j)} \mathcal{N}(x; m_{k-1}^{(j)}, P_{k-1}^{(j)}), \quad (6)$$

where $w_{k-1}^{(j)}$, $m_{k-1}^{(j)}$, and $P_{k-1}^{(j)}$ denote the weight, mean, and covariance of the j -th Gaussian component, respectively.

The predicted PHD function at time k , before incorporating any sensor data, is computed as:

$$D_{k|k-1}(x) = D_{S,k|k-1}(x) + \gamma_k(x), \quad (7)$$

where $\gamma_k(x)$ is the birth PHD function and $D_{S,k|k-1}(x)$ is the surviving target PHD function given by:

$$D_{S,k|k-1}(x) = p_S \sum_{j=1}^{J_{k-1}} w_{k-1}^{(j)} \mathcal{N}(x; m_{S,k|k-1}^{(j)}, P_{S,k|k-1}^{(j)}), \quad (8)$$

with predicted means and covariances:

$$m_{S,k|k-1}^{(j)} = F_{k-1} m_{k-1}^{(j)}, \quad P_{S,k|k-1}^{(j)} = Q_{k-1} + F_{k-1} P_{k-1}^{(j)} F_{k-1}^\top. \quad (9)$$

Here, $p_S \in [0, 1]$ denotes the probability that a target survives from time $k-1$ to k , and F_{k-1} , Q_{k-1} are the motion model and process noise covariance.

The IC-PHD update proceeds by applying single-sensor GM-PHD updates in sequence. For sensor 1 with measurement set $Z_k^{(1)}$, the updated PHD is:

$$D_k^{(1)}(x) = (1 - p_{D,k}^{(1)}) D_{k|k-1}(x) + \sum_{z \in Z_k^{(1)}} D_{D,k}^{(1)}(x; z), \quad (10)$$

where $p_{D,k}^{(1)} \in [0, 1]$ is the probability of detection for sensor 1, and each measurement $z \in Z_k^{(1)}$ contributes:

$$D_{D,k}^{(1)}(x; z) = \sum_{j=1}^{J_{k|k-1}} w_k^{(1,j)}(z) \mathcal{N}(x; m_k^{(1,j)}(z), P_k^{(1,j)}). \quad (11)$$

The update parameters are computed as follows:

$$w_k^{(1,j)}(z) = \frac{p_{D,k}^{(1)} w_{k|k-1}^{(j)}(z)}{\kappa_k^{(1)}(z) + \sum_{\ell=1}^{J_{k|k-1}} p_{D,k}^{(1)} w_{k|k-1}^{(\ell)}(z)}, \quad (12)$$

$$q_k^{(1,j)}(z) = \mathcal{N}(z; H_k m_{k|k-1}^{(j)}, S_k^{(1,j)}), \quad (13)$$

$$S_k^{(1,j)} = H_k P_{k|k-1}^{(j)} H_k^\top + R_k, \quad (14)$$

$$K_k^{(1,j)} = P_{k|k-1}^{(j)} H_k^\top (S_k^{(1,j)})^{-1}, \quad (15)$$

$$m_k^{(1,j)}(z) = m_{k|k-1}^{(j)} + K_k^{(1,j)}(z - H_k m_{k|k-1}^{(j)}), \quad (16)$$

$$P_k^{(1,j)} = [I - K_k^{(1,j)} H_k] P_{k|k-1}^{(j)}. \quad (17)$$

Here, H_k and R_k are the observation matrix and measurement noise covariance, respectively, and $\kappa_k^{(1)}(z)$ is the clutter intensity function for sensor 1.

The posterior $D_k^{(1)}(x)$ becomes the prior for the next sensor update:

$$D_k^{(2)}(x) = (1 - p_{D,k}^{(2)}) D_k^{(1)}(x) + \sum_{z \in Z_k^{(2)}} D_{D,k}^{(2)}(x; z), \quad (18)$$

and this process is repeated for all sensors $q = 3, \dots, Q$. The final posterior PHD function after processing all sensors is:

$$D_{k|k}(x) = D_k^{(Q)}(x). \quad (19)$$

To maintain tractability, pruning and merging techniques are applied after each sensor update. Low-weight components are removed, and nearby components (based on Mahalanobis distance) are merged.

The Gaussian mixture implementation of the IC-PHD filter retains much of the simplicity and efficiency of its single-sensor counterpart, while enabling the fusion of data from multiple sensors. Despite the theoretical approximations introduced by sequential updates, the method remains effective, especially when sensor characteristics are relatively homogeneous.

B. Formulation for void probability-based reward function

In POMDP sensor control problems, the reward function plays a central role in determining the most effective control action the sensor should take at a given time. A typical sensor management task involves selecting a control input $\mathbf{u}_k \in \mathcal{U}_k$ to optimize the sensing performance over time with $\mathcal{U}_k$ denotes the set of admissible control vectors. The goal is to choose $\mathbf{u}_k$ such that the expected utility of the sensing outcome is maximized:

$$\mathbf{u}_k^* = \arg \max_{\mathbf{u}_k \in \mathcal{U}_k} \mathbb{E}[\mathcal{R}(\mathbf{u}_k; \cdot)] \quad (20)$$

Here, $\mathcal{R}(\mathbf{u}_k; \cdot)$ is the reward function that quantifies how effective action $\mathbf{u}_k$ is with respect to the task objective.

This paper employs a reward function based on the void probability, defined as the probability that no targets are present within a specified region $B \subset \mathbb{X}$, where $\mathbb{X}$ denotes the state space. For a Poisson point process, the void probability is given by [14]:

$$P_0(B) = \exp\left(-\int_B D(x) dx\right) \quad (21)$$

where $D(x)$ denotes the PHD of the multi-target state. As this function quantifies the likelihood that no targets will be detected within the region B , a low value of $P_0(B)$ suggests a high probability of target presence. Based on this interpretation, this paper proposes using the complement of the void probability as the reward function:

$$\mathcal{R}(\mathbf{u}_k) = 1 - P_0(B(\mathbf{u}_k)) = 1 - \exp\left(-\int_{B(\mathbf{u}_k)} D_{k|k-1}(x) dx\right) \quad (22)$$

Here, $D_{k|k-1}(x)$ is the predicted PHD, and $B(\mathbf{u}_k)$ is a pre-specified region of interest. At time k , each control vector can be chosen to maximize reward function: $\mathbf{u}_k^* = \arg \max_{\mathbf{u}_k \in \mathcal{U}_k} \mathcal{R}(\mathbf{u}_k)$. This formulation rewards sensor actions that bring the sensor closer to regions of high target density, effectively guiding the system toward more informative and target-rich areas.

Since $D_{k|k-1}(x)$ is represented as a weighted sum of Gaussian components, the reward function $\mathcal{R}(\mathbf{u}_k)$ in equation (22) can be evaluated efficiently using low-dimensional numerical quadrature, thus enabling fast and accurate computation. This property makes the void probability-based approach a compelling choice for computationally efficient and scalable sensor management in real-time multi-target tracking applications.

C. Proposed algorithm for multi-sensor management

Having derived explicit formulations for multi-sensor multi-target PHD propagation using the IC-PHD filter, along with a void probability-based reward function, we now present a complete algorithm for the multi-sensor management problem.

The proposed solution comprises two key components: (1) a main loop that propagates the multi-target posterior PHD over time using the iterated-corrector (IC) multi-sensor GM-PHD filter, and (2) a sensor control subroutine that selects optimal control actions for each sensor by maximizing the void probability-based reward. The pseudo-code for these procedures is presented in Algorithm 1 (main loop) and Algorithm 2 (sensor control), respectively.

Algorithm 1 Main PHD propagation loop using IC-PHD filter

Require: Initial PHD $D_0(x) = \{w_0^{(j)}, m_0^{(j)}, P_0^{(j)}\}_{j=1}^{J_0}$;

Number of sensors Q ;

Ensure: Posterior PHD $D_k(x)$ at each time step

- 1: Set time index $k \leftarrow 0$
- 2: **repeat**
- 3: $k \leftarrow k + 1$
- 4: Predict $D_{k|k-1}(x)$ based on $D_{k-1}(x)$
- 5: Compute optimal control command $\mathbf{u}_k^*$ for Q sensors
- 6: Apply $\mathbf{u}_k^*$ to obtain $Z_k = \{Z_k^1, Z_k^2, \dots, Z_k^Q\}$
- 7: Set initial update: $D_k^{(0)}(x) \leftarrow D_{k|k-1}(x)$
- 8: **for** $i = 1$ to Q **do**
- 9: Update using Z_k^i :

$$D_k^{(i)}(x) = (1 - p_{D,k}^{(i)})D_k^{(i-1)}(x) + \sum_{z \in Z_k^i} D_{D,k}^{(i)}(x; z)$$

- 10: Set posterior PHD: $D_k(x) \leftarrow D_k^{(Q)}(x)$
 - 11: **until** stop
-

Algorithm 2 Multi-sensor control subroutine

Require: $D_{k|k-1}(x) = \{w_{k|k-1}^{(j)}, m_{k|k-1}^{(j)}, P_{k|k-1}^{(j)}\}_{j=1}^J$;

Number of sensors Q ;

Ensure: Optimal control actions $u_k^* = \{u_k^{1*}, \dots, u_k^{Q*}\}$

- 1: **for** $q = 1$ to Q **do**
- 2: Initialize $R_{\max} \leftarrow -\infty$
- 3: Initialize $u_k^{q*} \leftarrow \text{None}$
- 4: Calculate admissible control sets U_k^q
- 5: **for each** $u_k^q \in U_k^q$ **do**
- 6: Compute region of interest $B(u_k^q)$
- 7: Compute reward:

$$R(u_k^q) = 1 - \exp\left(-\int_{B(u_k^q)} D_{k|k-1}(x) dx\right)$$

- 8: **if** $R(u_k^q) > R_{\max}$ **then**
 - 9: $R_{\max} \leftarrow R(u_k^q)$
 - 10: $u_k^{q*} \leftarrow u_k^q$
 - 11: **return** $u_k^* = \{u_k^{1*}, \dots, u_k^{Q*}\}$
-

IV. SIMULATION RESULTS

The performance of the proposed method is evaluated in two simulated scenarios involving two mobile sensors tracking a time-varying number of targets in clutter. The simulation occurs within a $1000 \times 1000 \text{ m}^2$ surveillance area, where the sensors are initially located at $[0, 0]$ and $[1000, 1000]$. The first scenario employs two range-only sensors to track three moving targets. Target 1 exists over the time interval $k \in [1, 100]$, Target 2 over $k \in [1, 80]$, and Target 3 over $k \in [50, 80]$.

Each target's state is described by the vector $\mathbf{x} = [x \ y \ \dot{x} \ \dot{y}]^\top$, which contains its 2D position $[x \ y]^\top$ and velocity $[\dot{x} \ \dot{y}]^\top$. The target dynamics are governed by a linear Gaussian motion model with a survival probability of $p_S = 0.99$. The state transition is given by:

$$F_k = \begin{bmatrix} I_2 & \Delta I_2 \\ 0_2 & I_2 \end{bmatrix}, \quad Q_k = \sigma_v^2 \begin{bmatrix} \frac{\Delta^4}{4} I_2 & \frac{\Delta^3}{2} I_2 \\ \frac{\Delta^3}{2} I_2 & \Delta^2 I_2 \end{bmatrix},$$

where I_2 and 0_2 are the 2×2 identity and zero matrices, $\Delta = 1 \text{ s}$ is the sampling interval, and $\sigma_v = 0.5 \text{ m/s}^2$ is the process noise standard deviation. New targets appear according to a Poisson RFS birth model with intensity:

$$\gamma_k(x) = 0.1 \mathcal{N}(x; m_\gamma, P_\gamma).$$

The parameters for this birth model are set to $m_\gamma = [500, 350, -1, 1]^\top$, and $P_\gamma = \text{diag}(100, 100, 25, 25)$.

Let $s_k = (s_k^x, s_k^y)$ be the sensor's position at time k , and $\delta = 20 \text{ m}$ be the control step size. The set of all possible one-step ahead control actions is given by:

$$\mathcal{U}_k = \{(s_k^x + i\delta, s_k^y + j\delta) \mid i, j \in \{-1, 0, 1\}\}.$$

This formulation allows the sensor to move to any of its eight adjacent grid points or remain in its current position. The void probability in (22) is evaluated over a circular region B centered at the current sensor location, with radius 30 m.

The sensor's detection capability depends on the target's state relative to its own position. At a location s_k , the probability of detecting a target with state x_k is given by the model

$$p_{D,k}(x_k; a_{k-1}) = \frac{\mathcal{N}(s_k(a_{k-1}); Hx_k, S)}{\mathcal{N}(0; 0, S)}$$

$$\text{where } H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad S = 10^7 \begin{bmatrix} 3 & -2.4 \\ -2.4 & 3.6 \end{bmatrix}.$$

The observation z_k of a target with state x_k from a sensor at position s_k is given by

$$z_k = [\sqrt{(s_k^x - x_k)^2 + (s_k^y - y_k)^2}] + \varepsilon_k,$$

where ε_k is zero-mean Gaussian noise, $\varepsilon_k \sim \mathcal{N}(\cdot; 0, R_k)$. The measurement noise variance is $R_k = \sigma_r^2$, with a standard deviation of $\sigma_r = \sqrt{5} \text{ m}$.

Clutter is modeled as a Poisson RFS, uniformly distributed over the sensor's field of view. The clutter intensity is given by $\kappa_k(z) = \lambda_c u(z)$, where $\lambda_c = 5$ is the average number of clutter returns per scan and $u(z)$ is the uniform probability density over the measurement space.

The sensor trajectories generated by our proposed method are shown in Figure 1. The void probability-based method produces smooth and spatially focused paths, demonstrating that the sensors are effectively guided to converge on and remain within regions of high target density. This behavior is crucial for successful tracking, as it enhances the probability of detection and reduces measurement uncertainty.

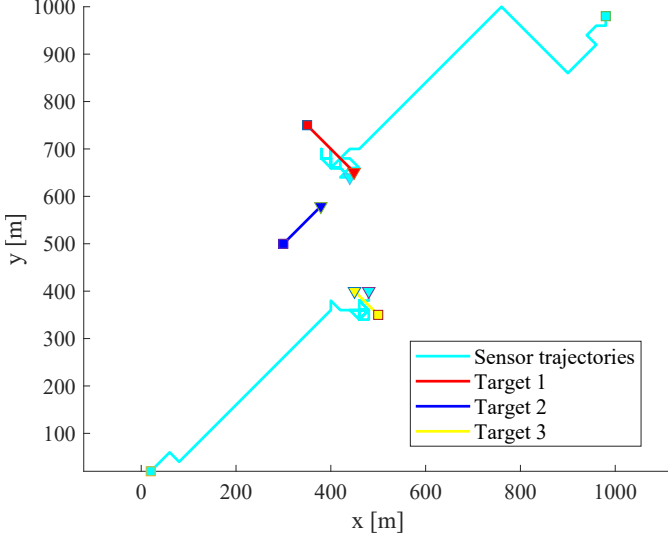


Fig. 1. Representative sensor trajectories generated by the proposed method, along with the ground truth paths of the three targets.

The computational performance was benchmarked by comparing the average runtime of three methods over 100 simulation runs: our proposed void probability-based multi-sensor control, its single-sensor counterpart, and a Rényi divergence-based multi-sensor control (with order parameter $\alpha = 0.5$). The tests were conducted in MATLAB R2024b on a computer with an Intel Core i7-12700 CPU and 16 GB of RAM. The results are presented in Table I. As shown in Table I, the

TABLE I
AVERAGE RUNTIME OF DIFFERENT SENSOR CONTROL METHODS

Method	Void-based single-sensor	Void-based multi-sensor	Rényi-based multi-sensor
Average runtime (3 targets)	2.47 s	3.27 s	56.51 s
Average runtime (4 targets)	3.45 s	4.48 s	72.12 s
Average runtime (5 targets)	3.83 s	4.61 s	103.15 s

average runtime for all methods increases with the number of targets. However, the proposed multi-sensor control shows favorable scalability, adding only a marginal computational overhead compared to the single-sensor version. The Rényi divergence-based method, by contrast, is substantially more demanding, especially as the number of targets grows. This performance difference is due to our method's reliance on an efficient numerical quadrature, which is much faster than the computationally intensive Monte Carlo sampling required by the Rényi divergence approach.

Tracking accuracy was evaluated using the OSPA metric ($p = 2, c = 100$) [15], with the average results over 100 Monte Carlo runs presented in Figure 2. Both multi-sensor control strategies—our void probability approach and the Rényi divergence benchmark—rapidly reduce tracking error. Although the Rényi divergence method yields a slightly lower initial error, our proposed strategy achieves comparable accuracy, converging to a similar error level around time step 90. The key takeaway is that our method provides this high level of performance for a significantly lower computational price, as established previously.

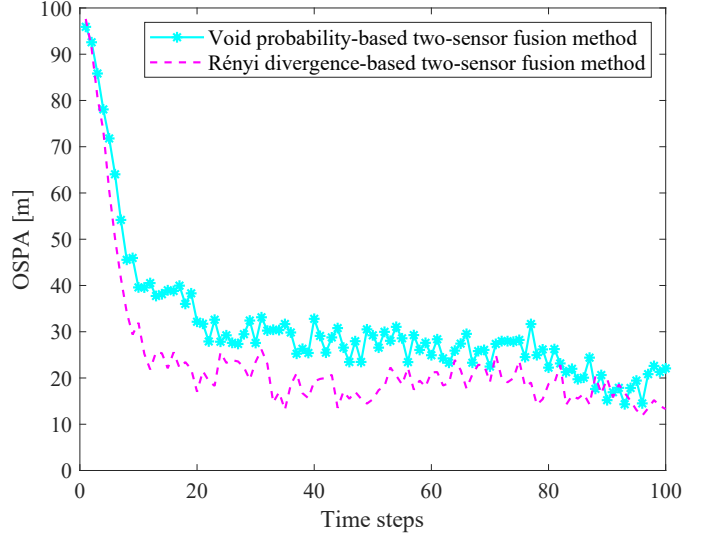


Fig. 2. Comparison of the average OSPA errors produced by different methods.

The second scenario was designed to test the robustness of our control strategy under two degraded conditions: low target detectability and high clutter density ($\lambda_c = 20$). In this test, we compare the performance of three sensor configurations: a single range sensor, homogeneous fusion of two range sensors, and heterogeneous fusion of one range and one bearing sensor. The bearing sensor provides angle-only measurements, which are modeled as:

$$z_k = \arctan\left(\frac{y_k - s_k^y}{x_k - s_k^x}\right) + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma_\theta^2). \quad (23)$$

where (s_k^x, s_k^y) is the sensor location and $\sigma_\theta^2 = (\pi/360)^2$ is the bearing noise variance.

The average OSPA for the low-detection and high-clutter cases are presented in Figures 3 and 4, respectively. The findings are consistent across both challenging environments:

- The single sensor struggles, exhibiting high error and slow convergence.
- Both fusion strategies dramatically improve tracking accuracy.
- The heterogeneous range-bearing network consistently achieves the lowest OSPA error. This is because the complementary angular measurements help disambiguate spurious detections and reject clutter more effectively than redundant range-only data.

These results highlight the clear benefit of multi-sensor fusion for maintaining robust tracking performance in non-ideal conditions.

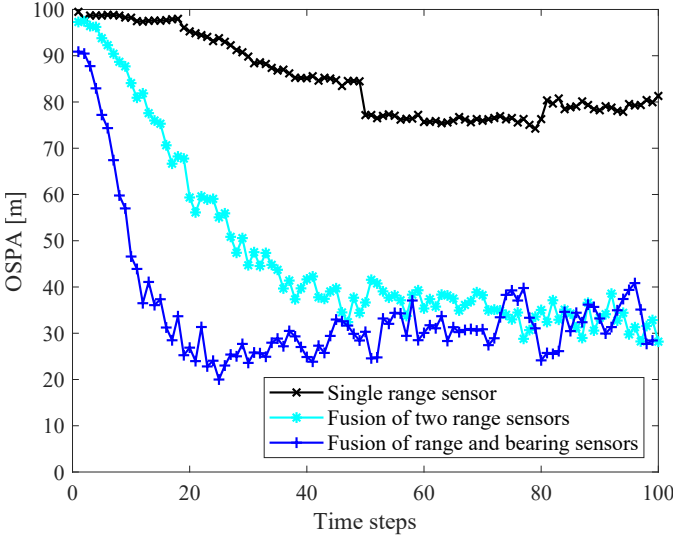


Fig. 3. The average OSPA errors under a low-observability scenario

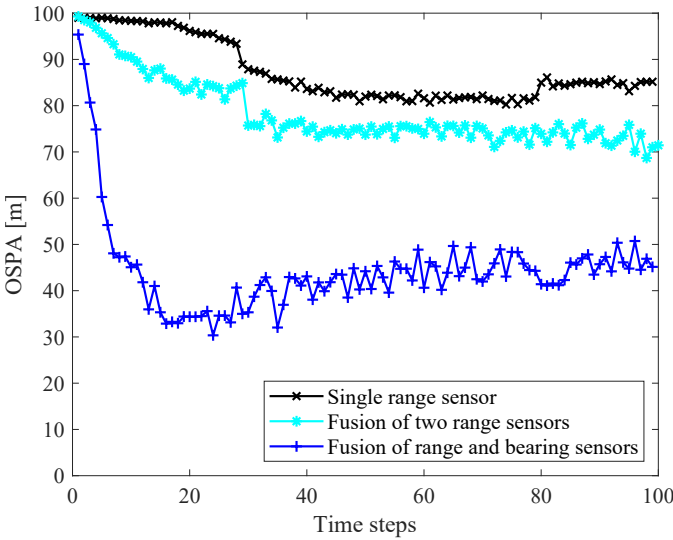


Fig. 4. The average OSPA errors under a high-clutter scenario

V. CONCLUSION

This paper presents an efficient approach to the multi-target, multi-sensor management problem by integrating RFS theory within the POMDP framework. The proposed method employs the IC-PHD filter to sequentially incorporate observations from multiple sensors, enabling the recursive propagation of the multi-target posterior intensity over time. This posterior is then used to derive a reward function based on the void probability, offering a favorable trade-off between computational efficiency and tracking accuracy, particularly under degraded sensing conditions. Extensive simulation results demonstrate that the

proposed method scales well with the number of targets and outperforms alternative approaches in balancing accuracy and runtime performance. Future research will explore more advanced multi-sensor fusion strategies to further improve tracking robustness, especially in highly cluttered, dynamic environments.

REFERENCES

- [1] D. J. Bucci and P. K. Varshney, "Decentralized multi-target tracking in urban environments: Overview and challenges," in *2019 22th International Conference on Information Fusion (FUSION)*. IEEE, 2019, pp. 1–8.
- [2] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.
- [3] Y. Liu, S. Wang, Y. Xie, T. Xiong, and M. Wu, "A review of sensing technologies for indoor autonomous mobile robots," *Sensors*, vol. 24, no. 4, p. 1222, 2024.
- [4] R. P. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic systems*, vol. 39, no. 4, pp. 1152–1178, 2004.
- [5] R. Mahler, *Statistical multisource-multitarget information fusion*. Artech, 2007.
- [6] D. A. Castañón and L. Carin, "Stochastic control theory for sensor management," in *Foundations and applications of sensor management*. Springer, 2008, pp. 7–32.
- [7] B. Ristic and B.-N. Vo, "Sensor control for multi-object state-space estimation using random finite sets," *Automatica*, vol. 46, no. 11, pp. 1812–1818, 2010.
- [8] B. Ristic, B.-N. Vo, and D. Clark, "A note on the reward function for PHD filters with sensor control," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 2, pp. 1521–1529, 2011.
- [9] H. G. Hoang and B. T. Vo, "Sensor management for multi-target tracking via multi-Bernoulli filtering," *Automatica*, vol. 50, no. 4, pp. 1135–1142, 2014.
- [10] M. Beard, B.-T. Vo, B.-N. Vo, and S. Arulampalam, "Sensor control for multi-target tracking using Cauchy-Schwarz divergence," in *2015 18th International Conference on Information Fusion (Fusion)*. IEEE, 2015, pp. 937–944.
- [11] H. G. Hoang, B.-N. Vo, B.-T. Vo, and R. Mahler, "The Cauchy-Schwarz divergence for poisson point processes," *IEEE Transactions on Information Theory*, vol. 61, no. 8, pp. 4475–4485, 2015.
- [12] N. T. Pham, W. Huang, and S. H. Ong, "Multiple sensor multiple object tracking with GMPHD filter," in *2007 10th International Conference on Information Fusion*. IEEE, 2007, pp. 1–7.
- [13] R. Mahler, "The multisensor PHD filter: I. general solution via multitarget calculus," in *Signal Processing, Sensor Fusion, and Target Recognition XVIII*, vol. 7336. SPIE, 2009, pp. 194–205.
- [14] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. Chichester: John Wiley & Sons, 1995.
- [15] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE transactions on signal processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

Dynamic Knowledge Distillation for Climate-Driven Diarrhoea Incidence Prediction

Thanh Duy Do ^{*†}, Thuan Dinh Nguyen ^{*}, Nguyen Khoi Pham ^{*}

Linh T.N. Trinh [†], Hans Vandierendonck [†], Son T. Mai [†]

^{*}University of Information Technology, Vietnam. Email: {thanhdd, thuannd, nguyennpk.20}@uit.edu.vn

[†]Queen's University Belfast, UK. Email: {tdo03, ltrinh01, h.vandierendonck, thaison.mai}@qub.ac.uk

Abstract—Diarrhoeal disease remains a significant public health burden in Vietnam, with incidence strongly influenced by climatic variability. Accurate short- and long-term forecasts are essential for effective early warning systems (EWS) to enable timely interventions. We propose a *Dynamic Knowledge Distillation Ensemble* (DKDE) framework for provincial-scale diarrhoea incidence forecasting that integrates local climate predictors with spatially informed neighbour-incidence data. DKDE dynamically selects the most relevant teacher models for each province and forecast horizon, adaptively combining their predictions into a compact student model through a data-driven distillation process. Using monthly diarrhoea surveillance data from 55 provinces (1997–2017) and corresponding climate records, we evaluate DKDE against state-of-the-art deep learning models (PatchTST, LSTM, CNN, Transformer variants) and the best-known method on the same dataset, the Dynamic Weighted Ensemble (DWE). Across six forecast horizons, DKDE consistently outperforms all baselines, achieving its largest gains in short-term prediction—reducing average 1-month RMSE from 16.74 (DWE) to 16.57—while maintaining superior accuracy at longer horizons. Comprehensive ranking and performance-gap analyses demonstrate DKDE’s robustness and stability across spatial and temporal scales.

Keywords—Diarrhoea prediction, Knowledge Distillation, Deep Learning, Ensemble Learning

I. INTRODUCTION

Diarrhoeal disease (DH) is a leading global health burden and the third leading cause of death among children aged 1–59 months, causing around 443,800 deaths annually, plus an additional 50,800 deaths in children aged 5–9 years [1]. Nearly 1.7 billion childhood cases occur each year. Severe cases can lead to fatal dehydration and contribute to a cycle of malnutrition and increased susceptibility to infection. Clinically, diarrhoea is categorised as acute watery, acute bloody (dysentery), or persistent (lasting over 14 days). Prevention measures include safe water, improved sanitation, hygiene, exclusive breastfeeding, rotavirus vaccination, and prompt treatment with oral rehydration salts (ORS) and zinc [1].

Vietnam, a developing country located entirely within the tropical climate zone, faces a high DH burden, with approximately 500,000 to 900,000 reported cases annually [2]. Rural areas, where incidence rates can exceed 115 episodes per 1,000 child-years, are particularly vulnerable due to limited access to clean water and sanitation facilities. This has placed a substantial burden on public health systems, highlighting the urgent need for effective early warning systems for disease prevention and control.

Numerous epidemiological studies have established strong

relationships between DH incidence and meteorological variables such as rainfall, temperature, and humidity [3]–[6]. These relationships can be nonlinear and vary across geographical regions. For example, total monthly rainfall has been shown to have both positive and negative correlations with DH incidence, depending on the lag period and local conditions. Similarly, average and maximum temperatures can be positively correlated with DH at short lags (0–2 months) but may display different effects in different provinces [3], [7]. Such variations suggest that predictive models should be designed and optimised at the provincial level.

For diarrhoea incidence forecasting, a range of statistical, machine learning, and deep learning models have been explored, including SARIMA and SARIMAX [8]–[10], Random Forest [11], [12], Support Vector Machine (SVM) [13], and, more recently, advanced architectures such as ANN [14], CNN [15], LSTM [13], and Transformer-based models [15]. Nevertheless, the majority of existing studies adopt direct regression mappings from historical features to incidence rates or employ baseline deep learning architectures without incorporating additional integration mechanisms [13], [15]. Sophisticated ensemble strategies, such as our earlier Dynamic Weighted Ensemble [16], remain uncommon, and there is limited consideration of spatial dependencies, such as leveraging incidence data from neighboring provinces.

Our prior work on Dynamic Weighted Ensemble (DWE) [16] demonstrated that dynamically weighting top-performing models at inference time can improve forecasting accuracy. However, DWE requires running multiple models simultaneously during deployment, which increases computational costs. To address this limitation, we propose the *Dynamic Knowledge Distillation Ensemble* (DKDE) framework, which distills the predictive capabilities of dynamically selected top- N teacher models (per province and forecast horizon) into a single, compact student model. This design preserves the performance benefits of model ensembling while enabling efficient inference.

Contributions: First, we present a comprehensive evaluation of short-term (1-month) and long-term (up to 6-month ahead) diarrhoea incidence prediction using a diverse set of state-of-the-art deep learning models, offering the first systematic comparison in this domain. Second, we propose the *Dynamic Knowledge Distillation Ensemble* (DKDE), the first framework to apply knowledge distillation for diarrhoea incidence prediction. DKDE surpasses our previous Dynamic Weighted Ensemble method [16] and provides a foundation for developing an effective early warning system for disease incidence.

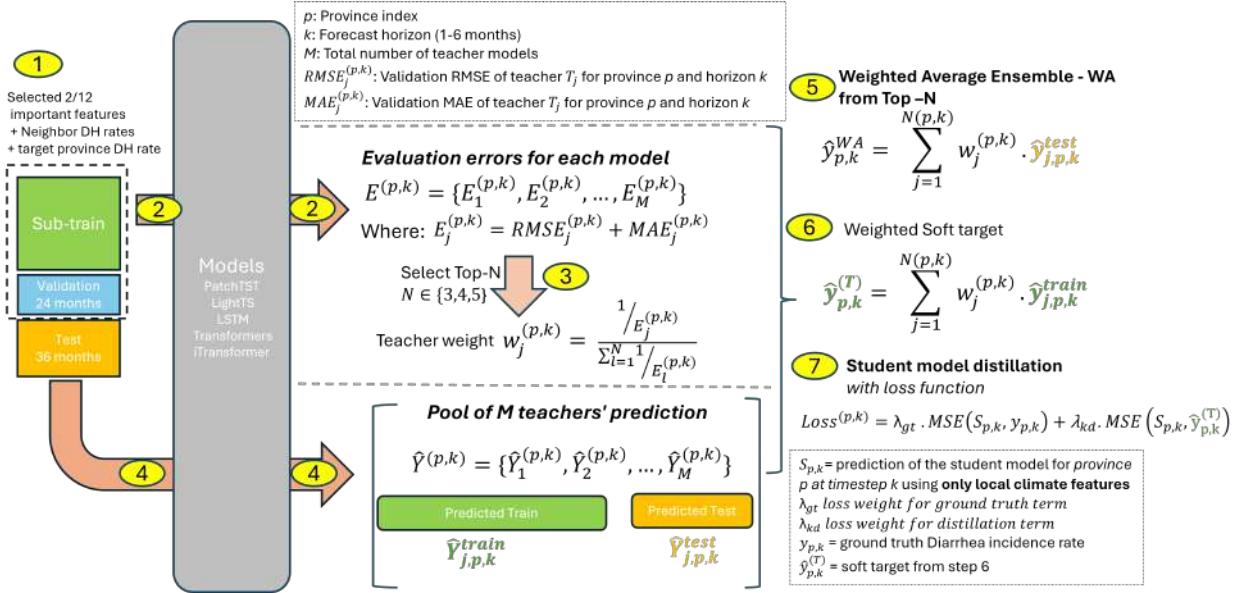


Fig. 1: Overall DKDE pipeline illustrating the seven main stages (yellow circles): teacher models are trained with local and neighbor features, top- N teachers are selected per province and forecast horizon, and their weighted predictions guide student training using local features only.

II. METHODS

A. Data and Problem Formulation

Data: We obtained monthly diarrhoea incidence data at the provincial level from the Vietnam National Institute of Hygiene and Epidemiology (NIHE) for the period 1997–2017, expressed as cases per 100,000 population, with consistently high burdens observed in several northern mountainous provinces and coastal areas. Over the same period, climate data were obtained from the Vietnam Institute of Meteorology, Hydrology and Climate Change (IMHEN), comprising 12 monthly variables: five temperature metrics (mean, absolute minimum/maximum, and mean minimum/maximum), two humidity indicators (average, minimum), three precipitation measures (total, maximum daily, rainy days), plus total evaporation and sunshine hours.

Problem Formulation: Let $p \in \{1, \dots, P\}$ index the provinces and $h \in \{1, \dots, H\}$ the forecast horizon. Let $\mathbf{x}_p \in \mathbb{R}^F$ represent the climate feature vector for province p , and y_p denote the corresponding diarrhoea incidence rate. The forecasting task is to predict the incidence rate h months ahead as:

$$\hat{y}_{p,h} = S_\theta(\mathbf{x}_p),$$

where S_θ denotes the student model parameterised by θ , and h is the forecast horizon.

B. Dynamic Knowledge Distillation Ensemble (DKDE)

The proposed DKDE framework consists of seven sequential steps (Fig. 1):

Step 1: Data Processing: The dataset is split into *Sub-train*, *Validation* (24 months), and *Test* (36 months) sets. For each province p , two important local climate features are selected using Recursive Feature Elimination (RFE) with a Random

Forest Regressor (500 trees). These features are combined with incidence rates from geographically adjacent neighbour provinces.

Step 2: Teacher Model Validation: A pool of M state-of-the-art (SOTA) teacher models $\{T_1, T_2, \dots, T_M\}$ is trained using Sub-train data with both local climate factors and neighbor incident rates. Each model is validated on the Validation set to compute:

$$E_j^{(p,k)} = RMSE_j^{(p,k)} + MAE_j^{(p,k)}, \quad j = 1, \dots, M$$

The set of validation errors for province p and horizon k is:

$$E^{(p,k)} = \{E_1^{(p,k)}, E_2^{(p,k)}, \dots, E_M^{(p,k)}\}$$

Step 3: Top- N Teacher Selection and Weight Calculation: For each (p, k) , select the $N^{(p,k)}$ teachers with the lowest $E_j^{(p,k)}$. Inverse-error normalised weights are assigned:

$$w_j^{(p,k)} = \frac{1/E_j^{(p,k)}}{\sum_{l=1}^{N^{(p,k)}} 1/E_l^{(p,k)}}$$

Step 4: Teacher Predictions on Train and Test: All M trained teacher models generate predictions for both the full training set (Sub-train + Validation) and the Test set:

- Train prediction pool: $\hat{y}_{j,p,k}^{train}$ denotes the prediction of teacher T_j for province p , horizon k .
- Test prediction pool: $\hat{y}_{j,p,k}^{test}$ denotes the prediction of teacher T_j for province p and horizon k on the Test set.

Step 5: Weighted Average Ensemble: The weighted average ensemble prediction on the Test set is:

$$\hat{y}_{p,k}^{(WA)} = \sum_{j=1}^{N^{(p,k)}} w_j^{(p,k)} \cdot \hat{y}_{j,p,k}^{test}$$

Step 6: Soft Target Generation: The same weights computed in Step 3 are used to generate the soft target as:

$$\hat{y}_{p,k}^{(T)} = \sum_{j=1}^{N^{(p,k)}} w_j^{(p,k)} \cdot \hat{y}_{j,p,k}^{\text{train}},$$

where $\hat{y}_{p,k}^{(T)}$ denotes the aggregated teacher prediction (soft target) for province p and forecast horizon k , $N^{(p,k)}$ is the number of selected teachers, $w_j^{(p,k)}$ is the weight assigned to teacher j , and $\hat{y}_{j,p,k}^{\text{train}}$ is the teacher's prediction on the training set.

Step 7: Student Model Training: The student model $S_{p,k}$, which uses only the selected local climate features, is trained to minimise the following objective:

$$\mathcal{L}^{(p,k)} = \lambda_{\text{gt}} \cdot \text{MSE}(S_{p,k}, y_{p,k}) + \lambda_{\text{kd}} \cdot \text{MSE}(S_{p,k}, \hat{y}_{p,k}^{(T)}),$$

where $y_{p,k}$ is the ground truth for province p and forecast horizon k , and $\hat{y}_{p,k}^{(T)}$ is the soft target generated in Step 6. The hyperparameters λ_{gt} , λ_{kd} , and $N^{(p,k)}$ are optimised using Optuna with the search ranges:

$$\lambda_{\text{gt}}, \lambda_{\text{kd}} \in \{0.5, 1.0, 1.5\}, \quad N^{(p,k)} \in \{3, 4, 5\}.$$

III. EXPERIMENTS

A. Setup

1) *Baseline Models:* We compare the proposed DKDE framework against a wide range of state-of-the-art time series forecasting models, including: PatchTST [17], LightTS [18], Transformer [19], iTransformer [20], FreTS [21], MLP [22], LSTM and LSTM_ATT [23]

2) *Parameter Tuning:* For fair comparison, all baseline models and the student model in DKDE are tuned using the Tree-structured Parzen Estimator (TPE) algorithm implemented in Optuna [24]. This approach is less computationally expensive than exhaustive grid search and is well suited for high-dimensional hyperparameter spaces. Unless otherwise specified, common hyperparameters are fixed as: `batch_size` = 16, number of features $F = 3$, and lookback window $L = 3$. The DKDE-specific hyperparameters λ_{gt} , λ_{kd} , and $N^{(p,k)}$ are optimised within the ranges: $\lambda_{\text{gt}}, \lambda_{\text{kd}} \in \{0.5, 1.0, 1.5\}$, $N^{(p,k)} \in \{3, 4, 5\}$.

3) *Evaluation Metrics:* We evaluate forecasting accuracy using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). For overall performance ranking across models, horizons, and provinces, we compute the average rank of each model, where lower values indicate better performance.

B. Results

1) *Local Climate Factor Baselines:* We first evaluate all baseline models using only the selected local climate features for each province. This serves as the lightest input configuration and a baseline for assessing the benefit of neighbour information. Fig. 3 shows that PatchTST and LightTS achieve the lowest average RMSE overall. Errors for all models increase with the forecasting horizon (k -step). The performance gap between PatchTST/LightTS and the other models is substantial in short-term horizons but narrows as the horizon lengthens.

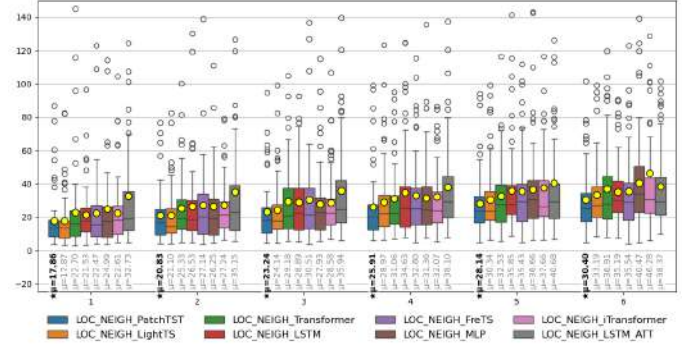


Fig. 2: RMSE distributions across k -steps for **LOC+NEIGH** baseline models, with per-city variability shown in boxplots. Yellow dots indicate the mean (μ) for each model-step, and starred μ marks the best model at that step

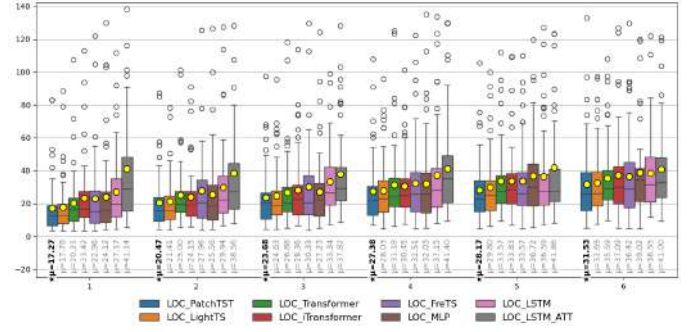


Fig. 3: RMSE distributions across k -steps for **LOC** baseline models. Per-step boxplots summarize per-city errors; yellow dots mark the mean (μ), with a star on the best μ per step. RMSE generally increases with forecasting horizon, with PatchTST showing consistently lower error medians compared to other baselines,

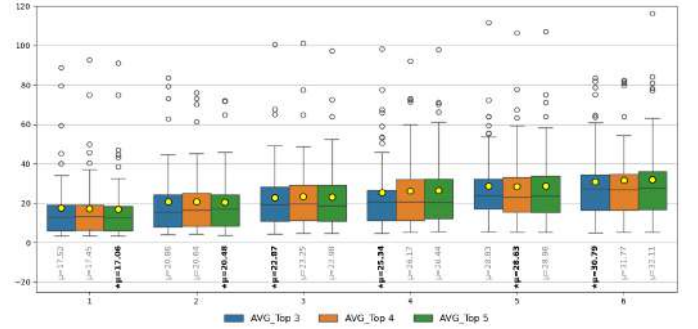


Fig. 4: RMSE distributions across k -steps for **Weighted Average Ensembles**, with per-city variability shown in boxplots. Yellow dots indicate the mean (μ) for each model-step, and starred μ marks the best model at that step

2) *Local + Neighbour Teacher Pools:* The same set of baseline models is subsequently trained using both local climate features and neighbouring provinces' incidence rates. This feature integration captures potential causal influences from neighbouring regions, enabling the models to exploit

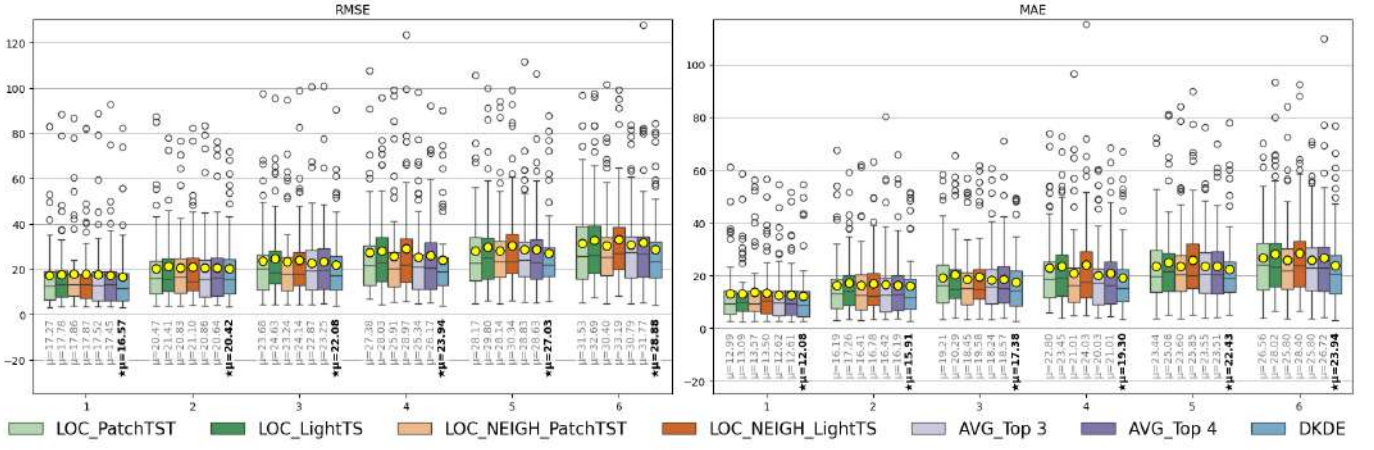


Fig. 5: Side-by-side boxplots of RMSE (left) and MAE (right) for the top-2 models selected from each group based on the lowest mean RMSE across all provinces and k -steps. The groups are color-coded as follows: **LOC**, **LOC_NEIGH**, **AVG_Top**, and **DKDE**. DKDE results are compared directly against the best-performing models from each group. Yellow circles denote group means, and black stars mark the best model per k -step, with the corresponding mean value μ annotated below each group.

spatial dependencies. Incorporating neighbour information not only enriches the input space but also facilitates the creation of a teacher model pool with more generalised and transferable knowledge. As shown in Fig. 2, combining neighbour incidence rates alongside local climate features improves performance for most models, particularly at shorter horizons (1–3 months). PatchTST and LightTS remain the top performers overall. Compared with the local-climate-only setting, teacher models achieve better long-term performance — for example, PatchTST records RMSE values of 25.91, 28.14, and 30.40 at steps 3–6 versus 27.38, 28.17, and 31.53 in the local-only case — but slightly worse results in the short term.

3) *Weighted Average Ensemble*: From the teacher prediction pool, the top- $N^{(p,k)}$ models (ranked by validation error) are combined on the test set using inverse-error weighted averaging, as described in Step 5 of Figure 1. Fig. 4 shows that the WA approach improves performance compared with both the individual teacher pools and the local-only baseline. This indicates the potential of ensemble strategies to further enhance forecasting accuracy by leveraging complementary strengths among top-performing models.

4) *DKDE Performance*: Fig. 5 shows that DKDE dominates both RMSE and MAE across all forecasting horizons (k -steps), consistently outperforming the best models from the LOC, LOC_NEIGH, and AVG_Top groups. Compared with the best-known DWE method [16], which records average RMSE and MAE of 16.74 and 12.32 at 1-step, DKDE achieves lower values of 16.57 and 12.08, confirming its superior predictive accuracy even in short-term forecasts.

Using RMSE and MAE rankings for each model to reduce the impact of outliers, the radar plot in Fig. 6 shows that DKDE, represented by the smallest red-line area near the center, is the best-performing method across all k -steps. The Weighted Average (WA) models rank second overall. The LOC group outperforms LOC+NEIGH in short-term horizons but falls behind in the longer term.

Fig. 7 show that DKDE maintains strong predictive align-

ment from short- to long-term horizons. Short horizons ($k = 1$ –2) capture both seasonal variations and abrupt changes, while medium to long horizons ($k = 3$ –6) preserve overall trend structure with stable forecasts, demonstrating the model’s robustness over extended lead times.

IV. RELATED WORKS AND DISCUSSIONS

Random Forest; ARIMA; ARIMAX: applied in Jiangsu Province, China, using meteorological factors (e.g., humidity, precipitation, atmospheric pressure) with lagged terms, past morbidity, and temporal indicators [11]. Multiplicative Holt–Winters: employed in Niono, Mali, incorporating past incidence, meteorological variables, and seasonality to model diarrhoea patterns [25]. Naïve Bayes (NB), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machine (SVM, RBF): applied in Bangladesh to assess diarrhoea incidence based on demographic, socio-economic, and household factors. Logistic regression (LR) was used to determine the high-risk factors of diarrhoea, after which the four ML-based approaches were applied [26]. CNN, LSTM, and SVM with GAN data augmentation, REVAC tuning, and backward stepwise feature selection: applied in South Africa’s nine provinces, using maximum, minimum, and average temperature, specific humidity, potential evaporation, precipitation rate, surface pressure, and wind velocity; the models predicted daily diarrhoea cases by incorporating climate information [13]. CNN, LSTM, LSTM_ATT, Transformer: implemented in Vietnam, leveraging curated climate predictors for provincial-scale diarrhoea forecasting [15]. Gradient Boosting, Random Forest, and kNN (stacked ensemble): developed in Guangzhou, China, using maximum, minimum, and mean temperature, rainfall, humidity, wind speed, and pressure; the four base prediction models (ANN, LSTM, SVR, XGBoost) were ensembled using stacking to obtain the final prediction model [27]. Second-layer dynamic weighted ensemble (DWE): designed in Vietnam, integrating multiple deep learning predictors with information-based climate features [16]. Most of these studies confirm that climate factors play a crucial role in

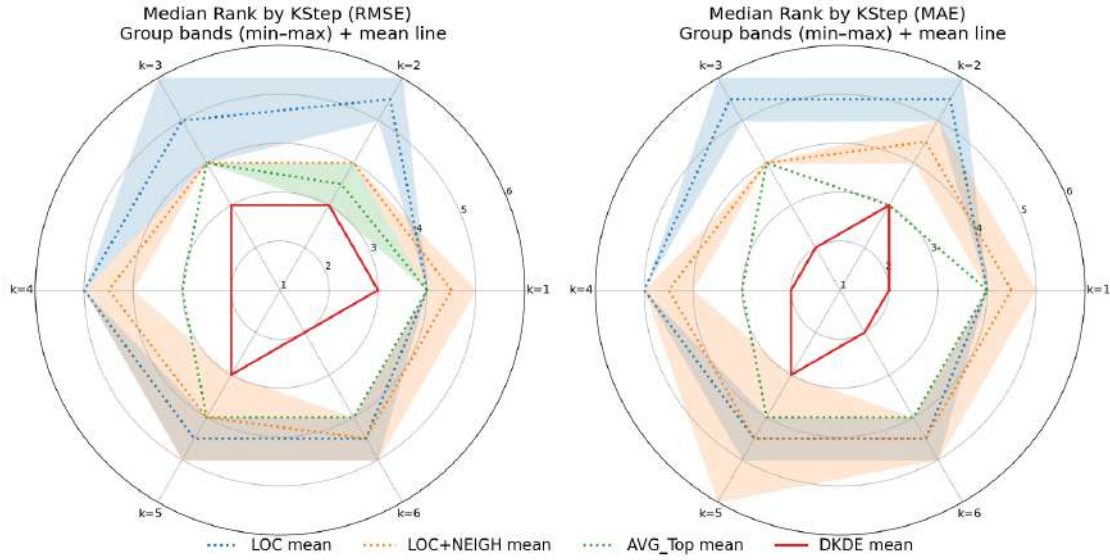


Fig. 6: Median rank comparison across forecast horizons ($k = 1$ to $k = 6$) for RMSE (left) and MAE (right). Shaded bands represent the performance range (min-max) within each group, while dotted lines indicate group means. The proposed DKDE method consistently achieves the lowest median ranks across all horizons and both metrics, outperforming Local (LOC), Local + Neighbour (LOC+NEIGH), and Weighted Average Top- N (AVG_Top) models. Lower ranks indicate better performance.

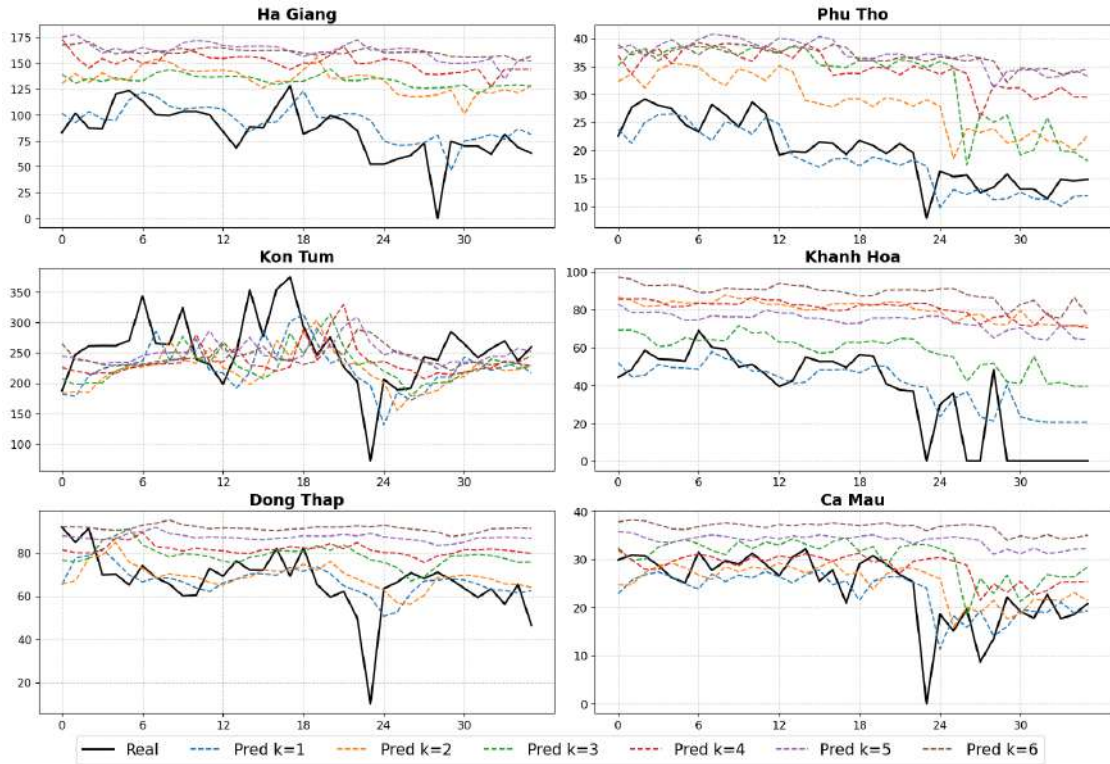


Fig. 7: Predictions for six sample provinces (Ha Giang, Phu Tho, Kon Tum, Khanh Hoa, Dong Thap, and Ca Mau) using the proposed DKDE model. Each subplot shows real diarrhoea incidence rates (black solid line) and predicted values for forecast horizons $k = 1$ to 6 months (colored dashed lines). The x-axis represents 36 consecutive months in the test period, and the y-axis represents the diarrhoea incidence rate.

predicting diarrhoea incidence, with variables such as temperature, humidity, rainfall, and wind speed frequently identified

as significant predictors. However, the majority of publications employ relatively direct approaches, linking historical climate

data to incidence rates using standard statistical or single-model machine learning techniques. Only a limited number of works adopt more advanced strategies—such as the stacked ensemble approach in [27] or the dynamic weight for second-layer ensemble in [16]—to exploit complementary model strengths. The proposed DKDE framework advances this line of research by leveraging a diverse pool of state-of-the-art baseline models and integrating them during the training phase via a knowledge distillation mechanism with inverse-error weighting. This design not only enhances predictive accuracy by combining complementary knowledge from top-performing teachers but also produces a lightweight student model suitable for operational deployment.

V. CONCLUSION

This study proposed a Dynamic Knowledge Distillation Ensemble (DKDE) framework for provincial-scale diarrhoea forecasting that adaptively selects and weights top-performing teacher models to guide a compact student model using only local climate features. Leveraging geographically informed neighbour incidence rates during teacher training, DKDE effectively exploits both spatial and temporal dependencies while maintaining a lightweight student configuration for deployment. Experimental results on 20 years of Vietnamese diarrhoea surveillance and climate data demonstrate that DKDE consistently outperforms strong deep learning, and ensemble baselines across all forecast horizons, with the largest gains in short-term prediction. Compared with the previous Dynamic Weighted Ensemble (DWE) method, DKDE achieves lower RMSE and MAE, confirming the benefit of dynamic teacher selection and inverse-error weighting. The framework is general and can be extended to other spatiotemporal forecasting tasks in epidemiology or climate-sensitive domains.

ACKNOWLEDGEMENT

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

REFERENCES

- [1] World Health Organization, "Diarrhoeal disease: Key facts," <https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease>, 2024, accessed: 2025-08-13.
- [2] K. L. Anders *et al.*, "The epidemiology and aetiology of diarrhoeal disease in infancy in southern vietnam: a birth cohort study," *International Journal of Infectious Diseases*, vol. 35, pp. 3–10, 2015.
- [3] D. Phung and o. Cunrui Huang, "Association between climate factors and diarrhoea in a mekong delta area," *International Journal of Biometeorology*, vol. 59, no. 9, pp. 1321–1331, 2015.
- [4] D. Phung *et al.*, "The effects of socioecological factors on variation of communicable diseases: A multiple-disease study at the national scale of vietnam," *PloS One*, vol. 13, no. 3, p. e0193246, 2018.
- [5] C. Phung *et al.*, "Heavy rainfall and risk of infectious intestinal diseases in the most populous city in vietnam," *Science of The Total Environment*, vol. 580, pp. 805–812, 2017.
- [6] C. N. Thompson *et al.*, "The impact of environmental and climatic variation on the spatiotemporal trends of hospitalized pediatric diarrhea in ho chi minh city, vietnam," *Health & Place*, vol. 35, pp. 147–154, 2015.
- [7] K. Wangdi and A. Clements, "Spatial and temporal patterns of diarrhoea in bhutan 2003–2013," *BMC Infectious Diseases*, vol. 17, no. 1, pp. 1–9, 2017.
- [8] C. J. Pangestu, E. Piantari, and M. Munir, "Prediction of diarrhea sufferers in bandung with seasonal autoregressive integrated moving average (sarima)," *Journal of Computers for Society*, vol. 1, no. 1, pp. 61–79, 2020.
- [9] M. Ali *et al.*, "Time series analysis of cholera in matlab, bangladesh, during 1988–2001," *Journal of Health, Population and Nutrition*, vol. 31, no. 1, pp. 11–19, 2013.
- [10] D. Phung *et al.*, "Temporal and spatial patterns of diarrhoea in the mekong delta area, vietnam," *Epidemiology & Infection*, vol. 143, no. 16, pp. 3488–3497, 2015.
- [11] X. Fang *et al.*, "Forecasting incidence of infectious diarrhea using random forest in jiangsu province, china," *BMC Infectious Diseases*, vol. 20, no. 1, pp. 1–8, 2020.
- [12] M. Yang *et al.*, "Meteorological factors affecting infectious diarrhea in different climate zones of china," *International journal of environmental research and public health*, vol. 19, no. 18, p. 11511, 2022.
- [13] T. Abdullahi and G. Nitschke, "Predicting disease outbreaks with climate data," in *2021 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2021, pp. 989–996.
- [14] Y. Wang and J. Gu, "Comparative study among three different artificial neural networks to infectious diarrhea forecasting," in *BIBM*. IEEE, 2014, pp. 40–46.
- [15] D. T. Do *et al.*, "Diarrhoea incidence prediction using climate data: Machine learning approaches," in *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 2022, pp. 1–6.
- [16] D. D. Thanh *et al.*, "Dynamic weighted ensemble for diarrhoea incidence predictions," *Machine Learning*, vol. 113, no. 4, pp. 2129–2152, 2024.
- [17] Y. Nie *et al.*, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.
- [18] T. Zhang *et al.*, "Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. arxiv 2022," *arXiv preprint arXiv:2207.01186*.
- [19] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [20] Y. Liu *et al.*, "itransformer: Inverted transformers are effective for time series forecasting," in *International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: <https://openreview.net/forum?id=JePfAI8fah>
- [21] K. Yi *et al.*, "Frequency-domain mlps are more effective learners in time series forecasting," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [22] B. Yan *et al.*, "Mlpst: Mlp is all you need for spatio-temporal prediction," *arXiv preprint arXiv:2309.13363*, 2023. [Online]. Available: <https://arxiv.org/abs/2309.13363>
- [23] V. H. Nguyen *et al.*, "Deep learning models for forecasting dengue fever based on climate data in Vietnam," vol. 16.
- [24] T. Akiba *et al.*, "Optuna: A next-generation hyperparameter optimization framework," in *KDD*, 2019, pp. 2623–2631.
- [25] D. C. Medina *et al.*, "Forecasting Non-Stationary Diarrhea, Acute Respiratory Infection, and Malaria Time-Series in Niono, Mali," vol. 2, no. 11, p. e1181.
- [26] M. Md *et al.*, "Prediction of childhood diarrhea in bangladesh using machine learning approach," *Insights Biomed Res*, vol. 4, no. 1, pp. 111–116, 2020.
- [27] P. Wang *et al.*, "Predicting the incidence of infectious diarrhea with symptom surveillance data using a stacking-based ensemble model," vol. 24, no. 1, p. 265.

CASE-LoRA: Continual Adaptation with Structurally Evolving LoRA for Class-Incremental Learning

Dat Quang Mac¹, Thanh-Nga Hoang Thi², Tri-Thanh Nguyen², Quynh-Trang Pham Thi^{2,*}

¹Posts & Telecommunications Institute of Technology, Hanoi, Vietnam

²Vietnam National University, University of Engineering and Technology, Hanoi, Vietnam

{datmq.b21cn215@stu.ptit.edu.vn, 23021647@vnu.edu.vn, ntthanh@vnu.edu.vn, trangptq@vnu.edu.vn}

Abstract—Rehearsal-based methods are effective for continual learning; however, this approach is associated with privacy concerns and memory issues. In this paper, we use parameter-efficient fine-tuning (PEFT) approaches, which do not require storing samples from the old task. Some current methods train models with separate components for each task, leading to time-consuming inference, a problem known as a bottleneck. Current parameter-efficient fine-tuning (PEFT) approaches either rely on impractical data rehearsal or employ complex inference-time selection mechanisms that scale poorly with the number of tasks. To overcome these limitations, we proposed CASE-LoRA (Continual Adaptation with Structurally Evolving Low-Rank Adaptation), which tackles the stability-plasticity dilemma through a unique synthesis of dynamic structural adaptation and importance-guided learning. Specifically, our method progressively expands the rank of Low-Rank Adaptation (LoRA) modules to allocate dedicated capacity for new knowledge. This structural evolution is coupled with a dual-purpose learning strategy that (1) shields important past parameters from destructive updates and (2) guides new parameters to learn in directions complementary to existing representations. Unlike prompt-based methods, CASE-LoRA adapts a single unified model, eliminating task-specific selection overhead. Our extensive experiments on demanding CIL benchmarks demonstrate that CASE-LoRA achieves state-of-the-art performance, significantly outperforming prior rehearsal-free and prompt-based methods, particularly in challenging long-sequence scenarios.

Index Terms—Class-Incremental Learning, Low-Rank Adaptation, Structurally Evolving

I. INTRODUCTION

Continual Learning (CL) aims to enable models to learn from a sequential stream of tasks without catastrophically forgetting previously acquired knowledge [2]. This capability is essential for practical applications that deal with constantly changing data. The fundamental challenge in continual learning involves balancing two competing needs: models must remain flexible enough to acquire new information while simultaneously maintaining the knowledge they have already learned [4], [7]. This trade-off is particularly acute in Class-Incremental Learning (CIL), a demanding scenario where tasks introduce disjoint sets of classes, and the model must classify all seen classes without task identifiers during inference.

The advent of large-scale Pre-trained Models (PTMs) has provided a powerful foundation for CL. Adapting PTMs

using Parameter-Efficient Fine-Tuning (PEFT) techniques like Low-Rank Adaptation (LoRA) [8] is a promising direction. However, current PEFT-based CL methods exhibit significant limitations. Prompt-based approaches [9]–[11] often suffer from complex and inefficient inference-time selection mechanisms that scale poorly as the number of tasks increases. Other methods, including some LoRA-based strategies [13], [14], rely on data rehearsal, which is often impractical due to memory, computational, or privacy constraints. This leaves a critical gap for a rehearsal-free CIL method that is both scalable and efficient.

To address these limitations, we introduce CASE-LoRA (Continual Adaptation with Structurally Evolving Low-Rank Adaptation), a novel rehearsal-free CIL framework that addresses these limitations. CASE-LoRA resolves the stability-plasticity dilemma by synergizing structural evolution with importance-guided learning. Its key mechanisms are: (1) *Progressive Rank Expansion*, which dynamically allocates new LoRA capacity for each task; (2) *Stability-Driven Gradient Modulation*, which protects consolidated knowledge by suppressing updates to important past parameters; and (3) *Guided Plasticity*, which steers new parameter updates in directions that are complementary to existing representations. This integrated approach allows CASE-LoRA to adapt its capacity while actively managing knowledge retention, resulting in a single, unified model that avoids task-specific selection overhead during inference.

The primary contributions of this work are as follows:

- We propose CASE-LoRA, a novel rehearsal-free CIL method that uniquely combines the dynamic structural evolution of LoRA modules with an importance-driven learning strategy to effectively balance stability and plasticity.
- We introduce specific mechanisms for (i) knowledge-preserving rank expansion, (ii) stability-driven gradient modulation to protect old parameters, and (iii) importance-guided projection to direct the learning of new parameters.
- We demonstrate through extensive experiments that CASE-LoRA achieves comparative performance on challenging CIL benchmarks, particularly excelling in long-sequence scenarios and outperforming existing rehearsal-

*Corresponding author: trangptq@vnu.edu.vn

free and prompt-based methods.

II. PRELIMINARIES

1) *Class-Incremental Learning*: In the Class-Incremental Learning setting, a model sequentially learns from multiple tasks $\{\mathcal{T}_1, \dots, \mathcal{T}_n\}$, with each task $\mathcal{T}_t$ presenting a non-overlapping collection of classes $\mathcal{C}_t$. Following training on task $\mathcal{T}_t$, the model must be able to classify instances from all previously encountered classes $\bigcup_{i=1}^t \mathcal{C}_i$. A key constraint is that the model does not receive task information during inference, requiring it to develop a single classification system capable of distinguishing among an increasingly large set of classes. This significantly increases the difficulty compared to Task-Incremental Learning and makes the approach vulnerable to catastrophic forgetting.

2) *Parameter-Efficient Fine-Tuning via Low-Rank Adaptation*: Low-Rank Adaptation (LoRA) [8] is a parameter-efficient fine-tuning (PEFT) technique that adapts a frozen pre-trained weight matrix W_0 by incorporating a low-rank update ΔW . Specifically, this update is decomposed into two trainable matrices, $\Delta W = W_B W_A$, where $W_A \in \mathbb{R}^{r \times k}$ and $W_B \in \mathbb{R}^{d \times r}$ have a small rank $r \ll \min(d, k)$. The forward pass is thus modified to:

$$h = W_0 x + W_B W_A x \quad (1)$$

LoRA's computational efficiency and the ability to merge updates directly into W_0 for zero-latency inference make it an attractive approach for continual learning. Nevertheless, managing these task-specific updates across a sequence of tasks while preventing catastrophic forgetting remains a significant challenge that this work aims to address.

III. METHODOLOGY

We propose CASE-LoRA, a continual learning strategy. The overall architecture is depicted in Figure 1. As illustrated, its core mechanism incrementally expands LoRA capacity to foster plasticity for new tasks while concurrently modulating updates to critical past parameters to preserve prior knowledge.

1) *Structural Adaptation via Progressive Rank Expansion*: Our method employs progressive rank expansion to incrementally augment model capacity while preserving prior knowledge. When transitioning from task $\mathcal{T}_{t-1}$ to $\mathcal{T}_t$, the LoRA rank is expanded as:

$$r_t = r_{t-1} + \Delta r_{\text{exp}}, \quad (2)$$

where Δr_{exp} is the per-task rank increment. The LoRA matrices are expanded by appending new parameters:

$$W_A^{(t)} = \begin{bmatrix} W_A^{(t-1)} \\ W_{A,\text{new}}^{(t)} \end{bmatrix}, \quad W_B^{(t)} = \begin{bmatrix} W_B^{(t-1)} & W_{B,\text{new}}^{(t)} \end{bmatrix}. \quad (3)$$

Through Kaiming initialization for $W_{A,\text{new}}^{(t)}$ and zero initialization for $W_{B,\text{new}}^{(t)}$, the LoRA update remains $\Delta W^{(t)} = \Delta W^{(t-1)}$ immediately after expansion, perfectly preserving prior knowledge and preventing catastrophic forgetting. The zero-initialization ensures new parameters remain inactive until updated through gradient learning on task $\mathcal{T}_t$.

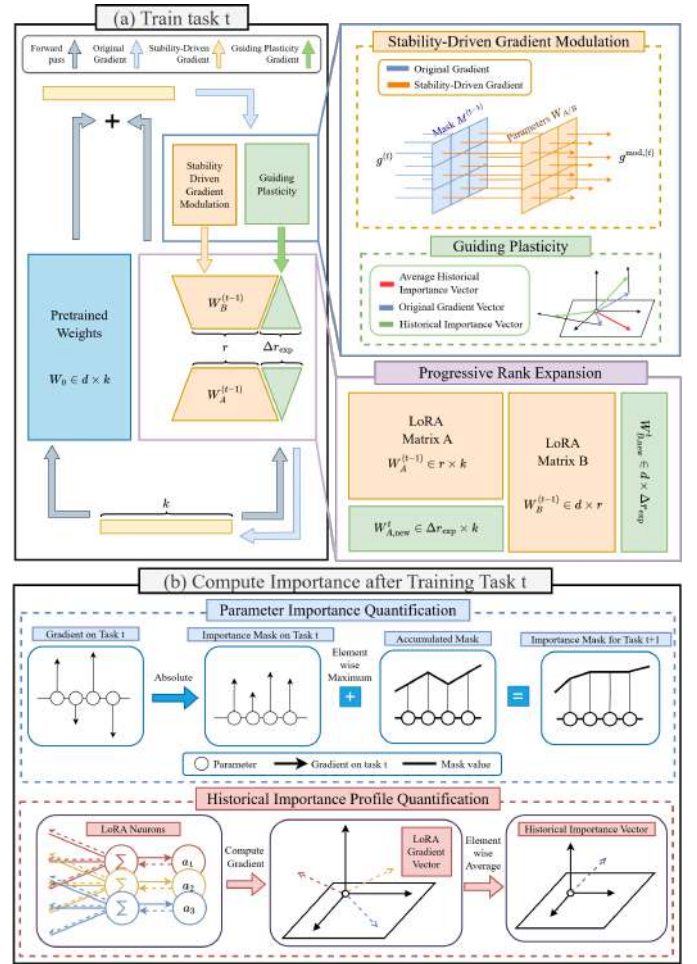


Fig. 1: The CASE-LoRA workflow for a continual learning task $\mathcal{T}_t$. (a) Training: Capacity is expanded via Progressive Rank Expansion. Learning is then managed by Stability-Driven Gradient Modulation to preserve existing knowledge and Guiding Plasticity to steer new parameter updates. (b) Post-Training: Parameter importance for $\mathcal{T}_t$ is quantified and used to calculate a Historical Importance Profile, which guides the training of the subsequent task.

2) *Stability-Driven Gradient Modulation*: CASE-LoRA employs a twofold strategy to address the stability-plasticity dilemma: parameter importance quantification and stability-driven gradient modulation.

Parameter Importance Quantification After training on task $\mathcal{T}_t$, parameter importance is estimated by gradient magnitude aggregation over the training dataset $\mathcal{D}_t$:

$$\tilde{\omega}_j^{(t)} = \left| \sum_{(x,y) \in \mathcal{D}_t} \frac{\partial \mathcal{L}_t(x,y)}{\partial \theta_j} \right| \quad (4)$$

Task-specific importance is normalized per-matrix: $\omega_j^{(t)} = \tilde{\omega}_j^{(t)} / (\max_l \tilde{\omega}_l^{(t)} + \epsilon)$. An accumulated importance mask across all tasks is maintained via:

$$\mathcal{M}_j^{(t)} = \max(\mathcal{M}_j^{(t-1)}, \omega_j^{(t)}) \quad (5)$$

Importance-Guided Gradients To prevent forgetting, gradients are modulated using both historical importance $\mathcal{M}_j^{(t-1)}$ and current instantaneous relevance $\phi_j^{(t)} = |g_j^{(t)}|/(\max_l |g_l^{(t)}| + \epsilon)$. A modulation factor combines both:

$$\rho_j^{(t)} = \mathcal{M}_j^{(t-1)} \cdot (1 - \phi_j^{(t)}) \quad (6)$$

The modified gradient is: $g_j^{\text{mod},(t)} = g_j^{(t)} \cdot (1 - \rho_j^{(t)})$. This suppresses updates to parameters that are important for past tasks but irrelevant for the current task while preserving updates to new or currently relevant parameters.

3) *Guiding Plasticity via Historical Importance Projection.*: While the previously described gradient modulation preserves knowledge in existing parameters, a distinct mechanism, illustrated as the ‘Guiding Plasticity’ block in Figure 1, is employed to steer the learning process for the newly introduced parameter segments ($W_A^{(t-1)}, W_B^{(t-1)}$) and relevant classifier weights. This mechanism, distinct from the gradient modulation, is employed to steer the learning process for the newly introduced parameter segments $W_{A,\text{new}}^{(t)}$ and $W_{B,\text{new}}^{(t)}$, resulting from the dynamic rank expansion. This guidance aims to encourage the new parameters to develop representations that are complementary to, and minimally interfere with, the knowledge encoded in the previous important LoRA components.

Let $\mathcal{M}_{A,\text{old}}^{(t-1)}$ and $\mathcal{M}_{B,\text{old}}^{(t-1)}$ denote the accumulated importance masks for the parameter segments $W_A^{(t-1)}$ and $W_B^{(t-1)}$ respectively, consolidated from tasks $\mathcal{T}_0, \dots, \mathcal{T}_{t-1}$. These masks encapsulate the historical significance of individual parameters within the established LoRA structure. For the new segment $W_{A,\text{new}}^{(t)} \in \mathbb{R}^{\Delta r_{\text{exp}} \times k}$, we compute an average historical importance profile along its feature dimension (columns) from $\mathcal{M}_{A,\text{old}}^{(t-1)} \in \mathbb{R}^{r_{t-1} \times k}$. For each column $c \in [1, k]$, this profile, denoted $\bar{m}_{A,c}^{(t-1)}$, is calculated as:

$$\bar{m}_{A,c}^{(t-1)} = \frac{1}{r_{t-1}} \sum_{j=1}^{r_{t-1}} \left(\mathcal{M}_{A,\text{old}}^{(t-1)} \right)_{j,c} \quad (7)$$

Similarly, for the new segment $W_{B,\text{new}}^{(t)} \in \mathbb{R}^{d \times \Delta r_{\text{exp}}}$, an average historical importance profile along its feature dimension (rows) is computed from $\mathcal{M}_{B,\text{old}}^{(t-1)} \in \mathbb{R}^{d \times r_{t-1}}$. For each row $r \in [1, d]$, this profile, $\bar{m}_{B,r}^{(t-1)}$, is:

$$\bar{m}_{B,r}^{(t-1)} = \frac{1}{r_{t-1}} \sum_{j=1}^{r_{t-1}} \left(\mathcal{M}_{B,\text{old}}^{(t-1)} \right)_{r,j} \quad (8)$$

During the training of task $\mathcal{T}_t$, the gradient $(g_{A,\text{new}}^{(t)})_{p,q}$ for an element $(W_{A,\text{new}}^{(t)})_{p,q}$ (where $p \in [1, \Delta r_{\text{exp}}], q \in [1, k]$) is modified as follows:

$$(g_{A,\text{new}}^{\text{mod},(t)})_{p,q} = (g_{A,\text{new}}^{(t)})_{p,q} \cdot \left(1 - \lambda_{\text{proj}}^{(t)} \cdot \bar{m}_{A,q}^{(t-1)} \right) \quad (9)$$

Likewise, the gradient $(g_{B,\text{new}}^{(t)})_{p,q}$ for an element $(W_{B,\text{new}}^{(t)})_{p,q}$ (where $p \in [1, d], q \in [1, \Delta r_{\text{exp}}]$) is modified:

$$(g_{B,\text{new}}^{\text{mod},(t)})_{p,q} = (g_{B,\text{new}}^{(t)})_{p,q} \cdot \left(1 - \lambda_{\text{proj}}^{(t)} \cdot \bar{m}_{B,p}^{(t-1)} \right) \quad (10)$$

TABLE I: Comparison on ImageNet-R. **Best** bold, second-best underlined.

Method	5 Tasks		10 Tasks		20 Tasks	
	AAA	Acc	AAA	Acc	AAA	Acc
Simple CIL	67.88	64.90	68.39	64.67	68.95	64.67
L2P	77.81	75.74	77.19	75.06	75.40	72.80
DualPrompt	75.90	74.09	77.90	72.99	75.33	71.44
CODA-Prompt	81.19	78.76	79.64	<u>76.72</u>	76.99	73.70
SD-LoRA	81.36	<u>78.43</u>	80.98	77.33	77.43	73.51
CASE-LoRA	82.41	78.30	81.26	75.23	80.65	75.29
w/o GP	82.49	78.04	81.39	75.10	<u>80.32</u>	<u>74.68</u>

TABLE II: Comparison on ImageNet-A (10 tasks of 20 classes) and CIFAR-100 (10 tasks of 10 classes) benchmarks. **Best** bold, second-best underlined.

Method	ImageNet-A (20 classes/task)		CIFAR-100 (10 classes/task)	
	AAA (%) $\uparrow$	Acc (%) $\uparrow$	AAA (%) $\uparrow$	Acc (%) $\uparrow$
Simple CIL	63.04	55.51	85.69	80.66
L2P	61.22	53.00	86.26	80.04
DualPrompt	62.45	54.56	87.75	82.76
CODA-Prompt	62.01	54.79	91.09	86.42
SD-LoRA	60.29	51.20	90.66	<u>86.17</u>
<i>Ours</i>				
CASE-LoRA	62.85	52.16	90.8	85.78
w/o GP	63.46	<u>54.92</u>	90.67	85.03

The term $\lambda_{\text{proj}}^{(t)}$ represents the guidance strength for task $\mathcal{T}_t$. It is a hyperparameter that can be constant or decay over tasks, for instance, $\lambda_{\text{proj}}^{(t)} = \lambda_{\text{proj}}^{\text{init}} \cdot (\gamma_{\text{decay}})^{t-1}$ for $t \geq 1$, where $\lambda_{\text{proj}}^{\text{init}}$ is the initial strength and γ_{decay} is a decay rate ($0 \leq \gamma_{\text{decay}} \leq 1$). This guidance mechanism steers the learning of new parameters based on the importance of previously learned ones. It reduces gradients in high-importance dimensions while allowing greater freedom in underutilized dimensions, enabling new parameters to explore less constrained directions and fostering synergy between old and new knowledge while minimizing interference with existing representations. The scaling factor $(1 - \lambda_{\text{proj}}^{(t)} \cdot \bar{m})$ modulates this process: stronger guidance occurs when the product of guidance strength $\lambda_{\text{proj}}^{(t)}$ and average past importance $\bar{m}$ is larger. Setting $\lambda_{\text{proj}}^{(t)}$ to zero deactivates this mechanism entirely.

IV. EXPERIMENTS

We evaluate CASE-LoRA in a class-incremental learning (CIL) scenario.

A. Experimental Setup

1) *Benchmarks and Evaluation Protocol.*: We evaluated CASE-LoRA using three established class-incremental learning benchmarks:

1) ImageNet-R [31]: A 200-class dataset designed to evaluate resilience to domain shifts. To test scalability, we create sequences with varying task numbers: 5 tasks (40 classes

each), 10 tasks (20 classes each), and 20 tasks (10 classes each).

2) ImageNet-A [32]: A demanding 200-class benchmark consisting of natural adversarial examples, organized into 10 tasks (20 classes per task).

3) CIFAR-100 [33]: A standard 100-class dataset commonly used for benchmarking, partitioned into 10 tasks (10 classes per task).

For evaluation, we report two metrics: Average Accuracy (Acc), which measures performance on all observed classes following the final task, and Average Anytime Accuracy (AAA) [34], which captures performance throughout the entire learning process.

2) Baselines and Implementation Details. We evaluate CASE-LoRA by comparing it against several strong rehearsal-free baselines, namely Simple CIL [30], prompt-based approaches (L2P [9], DualPrompt [10], CODA-Prompt [11]), and the LoRA-based method SD-LoRA [15]. All experiments are conducted using a ViT-B/16 model pre-trained on ImageNet-21K as the backbone architecture. Our approach integrates LoRA modules within the query and value projection components of all attention layers. The CASE-LoRA framework employs the following key hyperparameters: an initial rank of $r_0 = 10$, a per-task rank expansion of $\Delta r_{\text{exp}} = 10$, and an initial guidance strength of $\lambda_{\text{proj}}^{\text{init}} = 2.0$, which remains constant across tasks ($\gamma_{\text{decay}} = 1.0$). Model training proceeds for 15 epochs per task using the Adam optimizer, with a batch size of 16 and regularization via weight decay at 2×10^{-4} . The learning rate starts at 10^{-2} and decreases by a factor of 0.9 following each epoch.

B. Results and Discussion

1) Performance on ImageNet-R. As shown in Table I and Figure 2, CASE-LoRA consistently ranks top-two across all ImageNet-R configurations. In the challenging 20-task scenario, it achieves state-of-the-art AAA of 80.65% and final accuracy of 75.29%, outperforming all baselines. This demonstrates superior plasticity and mitigated forgetting over long task sequences through dynamic structural evolution. CASE-LoRA remains highly competitive in 5 and 10-task settings, validating its robustness across different learning granularities.

2) Performance on ImageNet-A and CIFAR-100. Our results on ImageNet-A and CIFAR-100 are presented in Table II and Figure 3. On ImageNet-A, ‘CASE-LoRA (w/o GP)’ achieves state-of-the-art AAA of 63.46%. This validates our core mechanisms: rank expansion provides capacity for challenging distributions, while gradient modulation preserves prior knowledge. The superior performance without GP suggests that, under severe domain shifts, freely adapting new parameters without guidance from historical importance maximizes plasticity.

On CIFAR-100, which features less dramatic domain shifts between tasks, prompt-based methods like CODA-Prompt hold an advantage, achieving the highest accuracy. This aligns with the understanding that the strict parameter isolation inherent in prompt tuning is a powerful strategy for scenarios

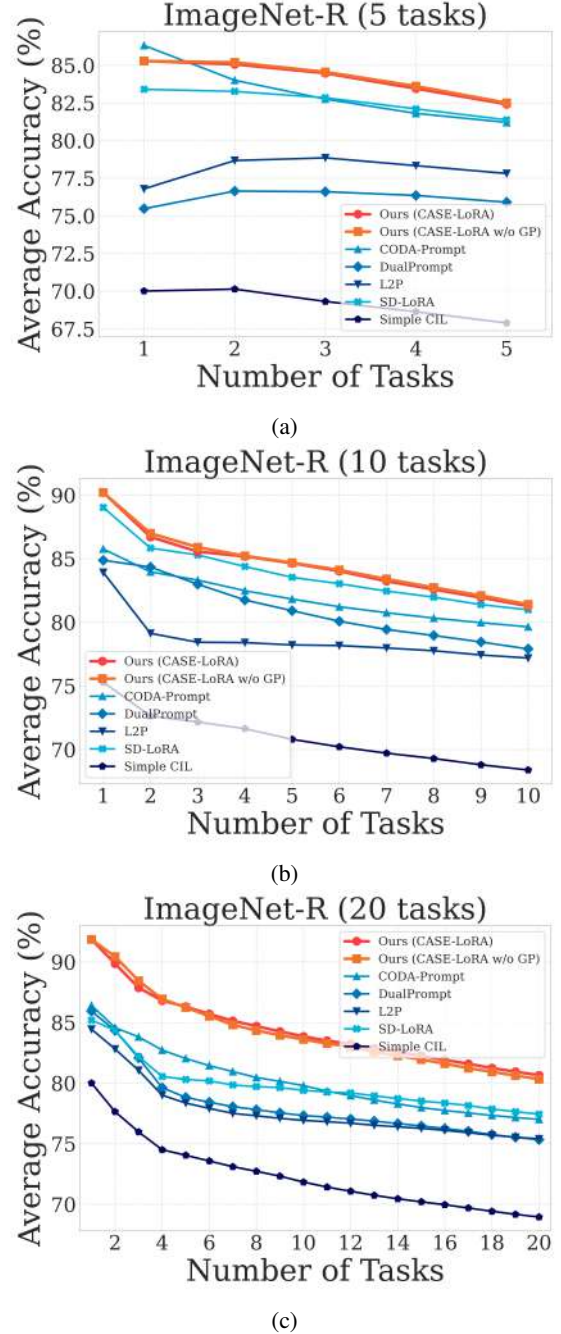


Fig. 2: Average Anytime Accuracy (%) over the learning trajectory on the ImageNet-R benchmark, partitioned into 5, 10, and 20 sequential tasks (from left to right).

with more homogeneous task sequences. Nonetheless, CASE-LoRA remains highly competitive, outperforming the other LoRA-based method, SD-LoRA, and achieving results (90.8% AAA) that are comparable to the leading prompt-based techniques. This demonstrates the robustness of our LoRA-centric adaptation strategy across different continual learning settings.

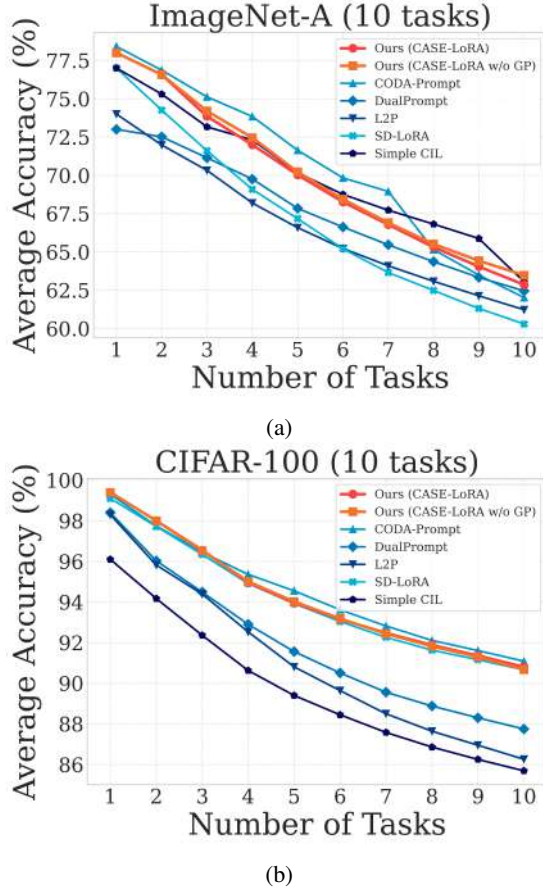


Fig. 3: Average Anytime Accuracy (%) on the ImageNet-A and CIFAR-100 benchmarks, each split into a 10-task sequence. Our method, CASE-LoRA (orange), demonstrates robust and competitive performance across both datasets.

To dissect the contribution of our proposed Guiding Plasticity via Historical Importance Projection (GP) mechanism, we evaluate an ablation variant, ‘CASE-LoRA (w/o GP)’. This variant retains the core mechanisms of dynamic rank expansion and stability-driven gradient modulation but removes the heuristic guidance applied to newly allocated parameters during training. By comparing the full model against this variant, we can precisely quantify the impact of GP across different continual learning scenarios. Tables I and II display the results.

The GP mechanism introduces a trade-off between structural coherence and plasticity under domain shifts. On ImageNet-R (Table I), while the variant without GP performs marginally better on shorter task sequences (5 and 10 tasks), the full CASE-LoRA model excels in the challenging 20-task scenario,

achieving higher AAA (80.65% vs. 80.32%) and final accuracy (75.29% vs. 74.68%). This validates that as task sequences grow, the GP mechanism effectively prevents interference between old and new task representations by directing parameter updates towards less critical feature dimensions, thereby preserving knowledge and mitigating performance degradation.

On ImageNet-A (Table II), the ‘w/o GP’ variant achieves the best performance (AAA 63.4%). Under extreme domain shifts, the GP heuristic becomes misaligned with the learning objective, as the historical importance profile provides incorrect signals about feature relevance in the adversarial domain. Removing this misleading guidance is a more effective adaptation without invalid priors.

V. CONCLUSION

In this paper, we propose CASE-LoRA, a rehearsal-free framework for Class-Incremental Learning that synergizes dynamic structural adaptation with importance-guided learning. Our method achieves state-of-the-art performance, particularly in long-sequence CIL, by effectively balancing stability and plasticity while maintaining high inference efficiency—a key advantage over prompt-based approaches. This performance is achieved by accepting key trade-offs: a moderate computational overhead during training for importance calculation and a linear growth in the parameter footprint with each new task. Future work should focus on mitigating its primary trade-offs by exploring more efficient methods for importance estimation, developing guidance mechanisms that are adaptive to inter-task distribution shifts, and investigating techniques to curb the linear growth in memory.

REFERENCES

- [1] Z. Wang et al.: Learning to Prompt for Continual Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 139-149, 2022.
- [2] L. Wang, X. Zhang, H. Su and J. Zhu: A Comprehensive Survey of Continual Learning: Theory, Method and Application. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5362-5383, 2024.
- [3] D. -W. Zhou, Q. -W. Wang, Z. -H. Qi, H. -J. Ye, D. -C. Zhan and Z. Liu: Class-Incremental Learning: A Survey. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 9851-9873, 2024.
- [4] Michael McCloskey, Neal J. Cohen: Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In: Psychology of Learning and Motivation, vol. 24, pp. 109-165, 1989.
- [5] McClelland JL, McNaughton BL, O'Reilly RC: Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. In: Psychological Review, vol. 102, no. 3, pp. 419-457, 1995.
- [6] Robert M. French: Catastrophic forgetting in connectionist networks. In: Trends in Cognitive Sciences, vol. 3, no. 4, pp. 128-135, 1999.
- [7] Kirkpatrick, James & Pascanu, Razvan & Rabinowitz, Neil & Veness, Joel & Desjardins, Guillaume & Rusu, Andrei & Milan, Kieran & Quan, John & Ramalho, Tiago & Grabska-Barwinska, Agnieszka & Hassabis, Demis & Clopath, Claudia & Kumaran, Dharmashan & Hadsell, Raia: Overcoming catastrophic forgetting in neural networks. In: Proceedings of the National Academy of Sciences, vol. 114, no. 13, pp. 3521-3526, 2017.
- [8] Edward J. Hu and Yelong Shen and Phillip Wallis and Zeyuan Allen-Zhu and Yuanzhi Li and Shean Wang and Lu Wang and Weizhu Chen: LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations, 2022.

Algorithm 1 CASE-LoRA Algorithm

Require: Task sequence $\mathcal{S} = \{\mathcal{D}_t\}$; Pre-trained model $\mathbf{W}_0$

Require: Hyperparameters: $r_0, \Delta r_{\text{exp}}, \lambda_{\text{proj}}^{\text{init}}, \gamma_{\text{decay}}$

```

1:  $\mathbf{W}_L, \mathcal{M}_L \leftarrow \text{InitLoRA}(r_0, \mathbf{0})$   $\triangleright$  Initialize LoRA weights
   and masks
2:  $\mathbf{W}_C, \mathcal{M}_C \leftarrow \text{InitClassifier}(\mathcal{C}_0, \mathbf{0})$   $\triangleright$  Initialize classifier
3:  $r_{\text{curr}} \leftarrow r_0$ 
4: for each task  $\mathcal{T}_t \in \mathcal{S}$  do
5:    $\mathbf{W}_C, \mathcal{M}_C \leftarrow \text{AdaptForNewClasses}(\mathbf{W}_C, \mathcal{M}_C, \mathcal{C}_t)$ 
6:   if  $t > 0$  then
7:      $r_{\text{old}} \leftarrow r_{\text{curr}}; r_{\text{curr}} \leftarrow r_{\text{old}} + \Delta r_{\text{exp}}$ 
8:      $\mathbf{W}_L \leftarrow \text{ExpandLoRA}(\mathbf{W}_L, r_{\text{curr}})$ 
9:      $\mathcal{M}_L \leftarrow \text{ExpandMasks}(\mathcal{M}_L, r_{\text{curr}}, \mathbf{0})$ 
10:  end if
11:  for  $E_t$  epochs do
12:    for each batch  $(\mathbf{x}, y) \in \mathcal{D}_t$  do
13:       $\mathcal{L}_t, \{g_w\} \leftarrow \text{CalculateGradients}(\mathbf{x}, \mathbf{W}_L \cup \mathbf{W}_C)$ 
14:      if  $t > 0$  then
15:         $\forall w \in \mathbf{W}_{\text{old}}: g_w \leftarrow g_w(1 - \mathcal{M}_w(1 - \phi_w))$ 
16:         $\forall w \in \mathbf{W}_{A,\text{new}}: g_w \leftarrow g_w(1 - \lambda_{\text{proj}}^{(t)} \bar{m}_A^{(t-1)})$ 
17:         $\forall w \in \mathbf{W}_{B,\text{new}}: g_w \leftarrow g_w(1 - \lambda_{\text{proj}}^{(t)} \bar{m}_B^{(t-1)})$ 
18:      end if
19:       $\mathbf{W} \leftarrow \text{Optimizer}(\mathbf{W}, \{g_w\})$ 
20:    end for
21:    end for
22:     $g_w^{\text{agg}} \leftarrow \sum_{(\mathbf{x}, y) \in \mathcal{D}_t} \nabla_w \mathcal{L}_t(\mathbf{x}, y)$ 
23:     $\omega_w^{(t)} \leftarrow \text{Normalize}(|g_w^{\text{agg}}|)$ 
24:     $\mathcal{M}_w \leftarrow \max(\mathcal{M}_w, \omega_w^{(t)})$ 
25:  end for
26: return Trained model:  $(\mathbf{W}_0, \mathbf{W}_L, \mathbf{W}_C)$ 

```

- [9] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister: Learning to prompt for continual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 139–149, 2022.
- [10] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, and Jennifer Dy: DualPrompt: Complementary prompting for rehearsal free continual learning. In: European Conference on Computer Vision, pp. 631–648, 2022.
- [11] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira: CODA-Prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11909–11919, 2023.
- [12] Wei-Cheng Huang, Chun-Fu Chen, and Hsiang Hsu: OVOR: OnePrompt with virtual outlier regularization for rehearsal-free class-incremental learning. In: International Conference on Learning Representations, 2024.
- [13] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu: Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. In: Advances in Neural Information Processing Systems, pp. 69054–69076, 2024.
- [14] Yan-Shuo Liang and Wu-Jun Li: InfLoRA: Interference-free low-rank adaptation for continual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23638–23647, 2024.
- [15] Wu, Yichen and Piao, Hongming and Huang, Long-Kai and Wang, Renzhen and Li, Wanhua and Pfister, Hanspeter and Meng, Deyu and Ma, Kede and Wei, Ying: SD-LoRA: Scalable Decoupled Low-Rank Adaptation for Class Incremental Learning. In: International Conference

- on Learning Representations, 2025.
- [16] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro: Learning to learn without forgetting by maximizing transfer and minimizing interference. In: ArXiv preprint arXiv:1810.11910, 2018.
- [17] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc Aurelio Ranzato: On tiny episodic memories in continual learning. In: ArXiv preprint arXiv:1902.10486, 2019.
- [18] Quanzhang Wang, Renzhen Wang, Yichen Wu, Xixi Jia, and Deyu Meng: CBA: Improving online continual learning via continual bias adaptor. In: IEEE/CVF International Conference on Computer Vision, pp. 19082–19092, 2023.
- [19] Yichen Wu, Hong Wang, Peilin Zhao, Yefeng Zheng, Ying Wei, and Long-Kai Huang: Mitigating catastrophic forgetting in online continual learning by modeling previous task interrelations via pareto optimization. In: International Conference on Machine Learning, pp. 53892–53908, 2024.
- [20] Zhizhong Li and Derek Hoiem: Learning without forgetting. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 12, pp. 2935–2947, 2017.
- [21] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee: Overcoming catastrophic forgetting with unlabeled data in the wild. In: IEEE/CVF International Conference on Computer Vision, pp. 312–321, 2019.
- [22] Yichen Wu, Long-Kai Huang, Renzhen Wang, Deyu Meng, and Ying Wei: Meta Continual Learning Revisited: Implicitly Enhancing Online Hessian Approximation via Variance Reduction. In: International Conference on Learning Representations, 2024.
- [23] Zenke, Friedemann and Poole, Ben and Ganguli, Surya: Continual learning through synaptic intelligence. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 3987–3995, 2017.
- [24] Arun Mallya, Dillon Davis, and Svetlana Lazebnik: Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: European Conference on Computer Vision, pp. 67–82, 2018.
- [25] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach: Adversarial continual learning. In: European Conference on Computer Vision, pp. 386–402, 2020.
- [26] Rahul Ramesh and Pratik Chaudhari: Model Zoo: A growing “brain” that learns continually. In: ArXiv preprint arXiv:2106.03027, 2021.
- [27] Andrei A. Rusu and Neil C. Rabinowitz and Guillaume Desjardins and Hubert Soyer and James Kirkpatrick and Koray Kavukcuoglu and Razvan Pascanu and Raia Hadsell: Progressive Neural Networks. In: ArXiv arXiv:1606.04671, 2016.
- [28] Yoon, Jaehong and Yang, Eunho and Lee, Jeongtae and Hwang, Sung Ju: Lifelong Learning with Dynamically Expandable Networks. In: International Conference on Learning Representations, 2018.
- [29] Konishi, Tatsuya and Kurokawa, Mori and Ono, Chihiro and Ke, Zixuan and Kim, Gyuhak and Liu, Bing: Parameter-level soft-masking for continual learning. In: Proceedings of the 40th International Conference on Machine Learning, vol. 202, no. 721, pp. 17492–17505, 2023.
- [30] Zhou, Da-Wei and Cai, Zi-Wen and Ye, Han-Jia and Zhan, De-Chuan and Liu, Ziwei: Revisiting Class-Incremental Learning with Pre-Trained Models: Generalizability and Adaptivity are All You Need. In: International Journal of Computer Vision, vol. 133, no. 3, pp. 1012–1032, 2025.
- [31] Matteo Boschini, Lorenzo Bonicelli, Angelo Porrello, Giovanni Bellitto, Matteo Pennisi, Simone Palazzo, Concetto Spampinato, and Simone Calderara: Transfer without forgetting. In: European Conference on Computer Vision, pp. 692–709, 2022.
- [32] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Christina Kwon, and Jacob Steinhardt: Natural adversarial examples. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15262–15271, 2021.
- [33] Alex Krizhevsky: Learning multiple layers of features from tiny images: Technical report, University of Toronto, 2009. URL: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- [34] Soutif-Cormerais, Albin and Carta, Antonio and Cossu, Andrea and Hurtado, Julio and Lomonaco, Vincenzo and Van De Weijer, Joost and Hemati, Hamed: A Comprehensive Empirical Evaluation on Online Continual Learning. In: IEEE/CVF International Conference on Computer Vision Workshops, pp. 3510–3520, 2023.

FDRL: Semantic-Aware Representation Learning with Feature Drift-Resistant Space Regularization

1st Ngoc-Linh Nguyen
Vietnam National University
University of Engineering and Technology
Hanoi, Vietnam
2021152@vnu.edu.vn

2st Tri-Thanh Nguyen
Vietnam National University
University of Engineering and Technology
Hanoi, Vietnam
ntthanh@vnu.edu.vn

4st Quynh-Trang Pham Thi*
Vietnam National University
University of Engineering and Technology
Hanoi, Vietnam
trangptq@vnu.edu.vn

Abstract—Continual learning in deep neural networks faces the persistent challenge of catastrophic forgetting, where learning new tasks overwrites previously acquired knowledge. While Semantic-Aware Representation Learning (SARL) addresses this through experience replay and prototype-based regularization, it remains vulnerable to feature drift—the gradual shift of internal representations as models adapt to sequential tasks. This paper introduces FDRL (Feature Drift-Resistant Space Regularization), a novel approach that adapts drift-resistant space concepts from parameter-efficient architectures to full-tuning continual learning frameworks. FDRL integrates a stability mechanism that characterizes principal feature subspaces from previous model states using Singular Value Decomposition and applies soft regularization to prevent excessive drift from these stable representations. Experimental results demonstrate the effectiveness of our proposed method when compared to some related approaches.

Index Terms—Continual learning, Catastrophic forgetting, Feature Drift-Resistant Space Regularization

I. INTRODUCTION

Continual learning (CL) aims to address the catastrophic forgetting problem of deep learning when learning on sequential tasks [3]. There are three main strategies to enable neural networks to learn sequentially presented tasks without forgetting prior knowledge: regularization-based methods that constrain parameter updates to preserve important weights [4], [5], architectural approaches that dedicate specific network components to different tasks [6], [7], and rehearsal-based methods that maintain and replay samples from previous tasks during new learning [8], [9]. Among rehearsal-based approaches, Semantic-Aware Representation Learning (SARL) [1] has demonstrated promising results by incorporating semantic relationships between objects and employing sparse coding mechanisms inspired by biological neural networks. SARL leverages experience replay combined with prototype-based regularization to maintain stability across tasks. However, despite these advances, SARL and similar methods can

still suffer from feature drift—a subtle but persistent problem where the internal representations gradually shift as the model adapts to new tasks, potentially compromising the stability of previously learned knowledge. The challenge of feature drift becomes particularly acute in full-tuning architectures where all model parameters are updated during learning, as opposed to parameter-efficient fine-tuning approaches that modify only a subset of weights [2]. While recent work has introduced the concept of Drift-Resistant Space (DRS) to address this issue, existing DRS methods are designed primarily for parameter-efficient architectures and rely on the availability of a static pre-trained backbone—assumptions that do not hold in full-tuning continual learning scenarios [2]. To address these limitations, this paper introduces SARL with Feature Drift-Resistant Space Regularization (FDRL), an approach that adapts the core principles of drift-resistant learning to full-tuning continual learning frameworks. Our method integrates a stability mechanism that characterizes principal feature subspaces from previous model states and applies soft regularization to prevent excessive drift from these stable representations. Critically, our approach operates in an exemplar-free manner during computation, requiring no access to previous task data when constructing drift-resistant subspaces. The key contributions of this work are threefold. First, we present a novel adaptation of drift-resistant space concepts to full-tuning continual learning architectures, addressing a significant gap in existing methodologies. Second, we introduce an exemplar-free regularization mechanism that constructs stable feature subspaces using only new task data, making the approach practically viable in memory-constrained scenarios. Third, through comprehensive empirical evaluation on the challenging Sequential CIFAR-100 benchmark, we provide valuable insights into the complex trade-offs inherent in combining multiple stability mechanisms, revealing how distributed regularization can create unexpected optimization dynamics that favor plasticity over stability. Our experimental results reveal a nuanced outcome that challenges conventional

*Corresponding author: trangptq@vnu.edu.vn

assumptions about stability mechanisms in continual learning. While FDRL demonstrates increased plasticity on recent tasks, it does not consistently reduce catastrophic forgetting across all prior tasks compared to the baseline. This finding highlights the non-trivial nature of integrating multiple regularization strategies and suggests that uniform constraints can unexpectedly alter the delicate stability-plasticity balance that is crucial for effective continual learning.

II. METHODOLOGY

This section details the proposed method, Feature Drift-Resistant Space Regularization (FDRL). We begin by briefly reviewing the foundational methods: Semantic-Aware Representation Learning (SARL) and the Drift-Resistant Space (DRS) concept. Subsequently, we present our novel integration of these two approaches to address the challenges of feature drift, which causes catastrophic forgetting in continual learning.

A. Semantic-Aware Representation Learning (SARL)

The proposed method is built upon the Semantic-Aware Representation Learning (SARL) framework [1]. SARL is designed to enhance continual learning by incorporating the semantic relationships between objects. To achieve this, it first induces sparse activations within the network, a technique inspired by sparse coding in the brain. For each class, an object prototype is computed from the average features of its instances and stored upon task completion.

When learning new classes, SARL computes their initial prototypes and evaluates similarities across a comprehensive set of prototypes, including both the newly computed ones from the current task and the stored ones from all previous tasks. This allows the model to establish relationships not only between new and old classes but also among the new classes themselves. A metric learning approach then guides the model to cluster the representations of similar objects while separating them from dissimilar ones. Furthermore, to ensure knowledge retention, a prototype-based regularization loss is applied, encouraging the representations of previously learned classes to remain stable and not drift from their established prototypes.

The following subsections detail the key components of the SARL framework.

1) *Sparse Coding*: To mimic the sparse coding mechanism of the brain, the method SARL [1] applies k-winners-take-all (k-WTA). This function enables only the K percent most active neurons in each layer to propagate their signal to the next layer. For convolutional layers, k-WTA is applied locally across the channels at each spatial location, rather than globally. This localized approach promotes filter specialization and encourages the extraction of more efficient and semantically meaningful features.

2) *Object Prototypes*: Following the sparse coding mechanism, each class c is represented by its object prototype, $\mathcal{O}_c$, defined as the mean of all instance representations [1]:

$$\mathcal{O}_c = \frac{1}{N_c} \sum_i \mathbb{I}(y_i = c) a'_i \quad (1)$$

where a'_i is the normalized activation of instance x_i with label y_i , and N_c is the total number of instances in class c . At the end of each task, the object prototypes for all newly introduced classes are computed and stored in a prototype memory.

3) *Semantic-Aware Metric Learning*: The similarity between any two object prototypes, $\mathcal{O}_i$ and $\mathcal{O}_j$, is measured using cosine similarity [1]:

$$\text{sim}(\mathcal{O}_i, \mathcal{O}_j) = \frac{\mathcal{O}_i \cdot \mathcal{O}_j}{\|\mathcal{O}_i\| \cdot \|\mathcal{O}_j\|} \quad (2)$$

After a warm-up stage for each new task, this similarity measure is used to identify a set of semantically similar $\mathcal{S}_c$ for each new class c [1]:

$$\mathcal{S}_c = \{i \mid \text{sim}(\mathcal{O}_i, \mathcal{O}_c) \geq \tau_s \wedge i \neq c, \forall i \in \mathbb{C}_{0:t}\} \quad (3)$$

where τ_s is the similarity threshold and $\mathbb{C}_{0:t}$ is the set of all classes observed up to the current task t .

A semantic-aware metric learning loss, $\mathcal{L}_{SM}$, is then applied to encourage the mean representation of instances of a new class c within a batch to move closer to the prototypes of its similar classes and further from dissimilar ones [1]:

$$\mathcal{L}_{SM} = \sum_{c \in \mathbb{C}^t} \frac{\sum_{i \in \mathcal{S}_c} \|\mathcal{O}_i - \bar{o}_c\|^2}{\sum_{i \notin \mathcal{S}_c} \|\mathcal{O}_i - \bar{o}_c\|^2} \quad (4)$$

where $\mathbb{C}^t$ is the set of new classes in task t .

4) *Regularization on Object Prototypes*: To enhance stability, a regularization loss is applied to samples replayed from the memory buffer. This loss, $\mathcal{L}_{OP}$, penalizes the divergence of the current representations of old classes from their stored prototypes. It is formulated as the mean squared error between the mean representation of buffered samples for a class c in the current batch, o_c^b and its corresponding stored prototype $\mathcal{O}_c$ [1]:

$$\mathcal{L}_{OP} = \sum_{c \in \mathbb{C}^{0:t-1}} \|o_c^b - \mathcal{O}_c\|^2 \quad (5)$$

5) *Functional Regularization*: Functional regularization is employed to maintain consistency in the model's input-output behavior. Before training on a new task, the output logits for the new task's samples are pre-computed using the previous model state and stored. For samples replayed from the buffer, their original output logits are also stored. During training, a consistency loss, $\mathcal{L}_{FR}$, is applied to both new and buffered samples [1]:

$$\mathcal{L}_{FR} = \alpha \|F(x^b; \theta) - z^b\|^2 + \beta \|F(x^t; \theta) - z^t\|^2 \quad (6)$$

where $F(x; \theta)$ is the model's output, (x^b, z^b) are samples and stored logits from the buffer, (x^t, z^t) are samples and pre-computed logits from the new task, and α, β are weighting hyperparameters.

6) *Overall Formulation*: The final loss function for SARL combines the standard cross-entropy loss with the various regularization terms. For each training iteration on task t (for $t > 1$), the overall loss is the following [1]:

$$\mathcal{L} = \mathcal{L}_{CE}(F(x^t; \theta), y^t) + \mathcal{L}_{CE}(F(x^b; \theta), y^b) + \lambda_{SM}\mathcal{L}_{SM} + \lambda_{OP}\mathcal{L}_{OP} + \mathcal{L}_{FR} \quad (7)$$

where λ_{SM} and λ_{OP} are hyperparameters controlling the weight of their respective loss terms.

B. Foundational Concept: Drift-Resistant Space (DRS) via Subtraction

To further improve the stability of the SARL framework against catastrophic forgetting, the core principles of the LoRA Subtraction for Drift-Resistant Space (DRS) algorithm [2] are adapted. The original method was designed for Parameter-Efficient Fine-Tuning (PEFT) models, such as Vision Transformers (ViTs), operating in an exemplar-free setting.

The fundamental insight of this algorithm is the construction of a safe drift-resistant learning subspace, achieved by temporarily removing the influence of past tasks from the model. This is done by creating a temporary model state, $\widetilde{W}$, before learning a new task t . In its original context, this state is calculated by taking the initial, pre-trained weights of the model, W_0 , and subtracting the learned low-rank adaptation (LoRA) updates from all prior tasks [2]:

$$\widetilde{W} = W_0 - \sum_{j=1}^{t-1} \Delta W_j \quad (8)$$

This temporary model, representing a “clean” version of foundational knowledge without task-specific specializations, is then probed with data from the new task t . The input features to the model’s linear layers are collected, forming a feature matrix $X \in \mathbb{R}^{N \times D}$.

The next critical step is to characterize the principal directions of this feature space using Singular Value Decomposition (SVD). The uncentered covariance matrix of the features, $\chi = \frac{1}{N}X^T X$ [2], is analyzed. The SVD of this covariance matrix, $\chi = U\Sigma U^T$ [2], yields a set of orthonormal eigenvectors in the columns of U . These eigenvectors represent the principal components of the feature space, ordered by the amount of variance they capture. The projection matrix P is then formed by selecting the top k eigenvectors of U . This matrix $P \in \mathbb{R}^{D \times k}$ defines an orthonormal basis for the Drift-Resistant Space.

In the final step of the original algorithm, this space is used to constrain the learning process. During training on the new task, the computed gradient for the learnable parameters, g , is projected onto the DRS before the weight update. The projection operation, $\Delta w = PP^T g$ [2], ensures that the parameter update Δw only occurs within the safe subspace defined by P , thus minimizing interference with the feature structures required for previous tasks.

However, this elegant approach is predicated on an architecture with a clear separation between a static, pre-trained backbone (W_0) and distinct, additive, task-specific updates (ΔW_j). This architectural assumption presents a significant challenge when translating the concept to a full-tuning framework like SARL, where the entire model’s weights are modified throughout the learning process. In such a setting, there is no static W_0 to return to, nor are there separable ΔW components for each task. This mismatch necessitates a novel adaptation of the core DRS philosophy to make it applicable and effective in SARL context, which will be detailed in the following section.

C. Proposed Method: SARL with Feature Drift-Resistant Space Regularization (FDRL)

To address the challenges of catastrophic forgetting while avoiding the pitfalls of hard gradient projection, a novel Feature Drift-Resistant Space Regularization approach is integrated into the SARL framework. This method does not modify gradients directly, but instead adds a new regularization term to the overall loss function. Critically, this mechanism is designed to be exemplar-free in its computation, as it does not require the data buffer.

The procedure for learning a new task t is as follows:

1) **Stable Subspace Construction (before a new task):**

Before training begins on task t ($t > 0$), the principal feature subspaces for key layers of the previous model, W_{t-1} , are computed. For each target layer l , data from the new task t is passed through the old model W_{t-1} . The resulting output feature maps are collected and concatenated into a single matrix $X_{\text{old}}^{(l)} \in \mathbb{R}^{N \times D}$, where N is the number of samples and D is the flattened feature dimension.

To define the stable subspace, the Singular Value Decomposition (SVD) is computed. To handle large feature dimension efficiently, the Method of Snapshots is used. SVD is applied to the smaller matrix $X_{\text{old}}^{(l)}(X_{\text{old}}^{(l)})^T = U_s \Sigma_s V_s^T$.

The basis vectors $B^{(l)}$ for the subspace are then reconstructed:

$$B^{(l)} = (\Sigma_s)^{-\frac{1}{2}} U_s^T X_{\text{old}}^{(l)} \quad (9)$$

The rows of $B^{(l)}$ form an orthonormal basis for the principal subspace of the old features. The number of basis vectors, k , is determined dynamically by selecting the minimum number of components whose cumulative explained variance, derived from the singular values in Σ_s , exceeds a threshold τ_v . These basis vectors, which define the stable subspace for each layer, are then stored.

2) **Training with Feature Space Regularization (during a new task):**

The standard SARL training loop is augmented with a new regularization loss, $\mathcal{L}_{FDRL}$. During the forward pass on a new batch of data, hooks on the current model, W_t , capture the output feature maps, $F_{\text{current}}^{(l)}$, from the same target layers.

For each layer l , its current feature map is projected onto the corresponding stable subspace defined by its basis vectors B^l . The projection matrix is:

$$P^{(l)} = \left(B^{(l)}\right)^T B^{(l)} \quad (10)$$

The component of the features that is orthogonal to the old subspace, representing the drift, is calculated as follow:

$$F_{\text{orth}}^{(l)} = F_{\text{current}}^{(l)} - F_{\text{current}}^{(l)} P^{(l)} \quad (11)$$

A normalized regularization loss is then computed for that layer, penalizing the magnitude of this drift relative to the magnitude of the current features to prevent exploding loss values:

$$\mathcal{L}_{\text{layer}}^{(l)} = \frac{\|F_{\text{orth}}^{(l)}\|_2^2}{\|F_{\text{current}}^{(l)}\|_2^2} \quad (12)$$

The total feature drift-resistant space regularization loss, $\mathcal{L}_{\text{FDRL}}$, is the average of these losses across all targeted layers.

- 3) **Overall Formulation:** The final loss function combines the original SARL loss with this new regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SARL}} + \lambda_{\text{FDRL}} \mathcal{L}_{\text{FDRL}} \quad (13)$$

Where $\mathcal{L}_{\text{SARL}}$ represents all original loss components of SARL and λ_{FDRL} is a hyperparameter controlling the strength of the new regularization.

Figure 1 illustrates the proposed FDRL method, showing how the original SARL loss components are augmented by a new feature space regularization loss $\mathcal{L}_{\text{FDRL}}$, which is computed based on pre-calculated Safe Subspaces.

III. EXPERIMENT AND RESULTS

A. Experimental Setup

1) *Dataset.*: The evaluation is conducted on the challenging class-incremental learning benchmark, Sequential CIFAR-100 (Seq-CIFAR100). Following the standard setup [1], the 100 classes are split into 5 sequential tasks, with 20 new classes introduced in each task.

2) *Baselines.*: The primary baseline for comparison is the original SARL algorithm [1], which represents a strong experience replay method with semantic-aware learning components. The proposed FDRL is compared directly against this baseline to isolate the impact of the feature space regularization module.

3) *Evaluation Metrics.*: To provide a comprehensive assessment, the performance is evaluated using three key metrics, following the protocol established in prior work [1]. The primary metric is Average Accuracy(Acc), which is the mean accuracy across all tasks measured after the final task has been trained, reflecting the overall knowledge retained. The second metric is Forgetting(Fgt), which quantifies the average performance drop on a task compared to its when it was

first learned, where a lower value indicates better knowledge preservation. Finally, the Stability-Plasticity Trade-off(S-P) is measured as the harmonic mean of Plasticity(Plt)- average accuracy on new tasks and Stability(Stb) - average accuracy on old tasks after training; to assess how well the model balances learning new information with retaining old knowledge.

4) *Implementation Details.*: All experiments are conducted using a ResNet-18 backbone with k-WTA sparse activations, consistent with the original SARL implementation. The models are trained using an SGD optimizer with a batch size of 32 for 50 epochs per task. Hyperparameters for the baseline SARL are set according to the optimal configuration provided in the original paper. For our FDRL, the regularization strength $\lambda_{\text{feat\_reg}}$ was set to 0.4 and the variance threshold for subspace for subspace construction was set to 95%, determined through preliminary experiments. All experiments were run with a buffer size of 200 samples.

B. Results and Analysis

The empirical evaluation was conducted to determine if the proposed Feature Space Regularization (FSR) mechanism could enhance the stability of the SARL baseline. For metrics with upward arrows, higher values are better, while for metrics with downward arrows, lower values are better. The best results are shown in bold, and the second-best results are shown in *italics*.

The results in Table I demonstrate that FDRL achieves superior performance compared to baseline methods on the SEQ-CIFAR10 dataset. Notably, FDRL attains the highest accuracy (69.67%) and lowest forgetting rate (14.83%), demonstrating an effective balance between stability and plasticity. The improvement over the SARL baseline (69.13% accuracy, 16.44% forgetting) validates that feature drift-resistant space regularization successfully mitigates catastrophic forgetting. The enhanced stability-plasticity trade-off score (74.60 vs. 73.45) further confirms that the novel regularization mechanism not only preserves existing knowledge but also maintains the model's capacity to learn new tasks effectively.

Table II reveals an intriguing and counter-intuitive result where FDRL (47.39%) underperforms compared to the SARL baseline (48.41%). This outcome demonstrates that integrating multiple stability mechanisms does not always yield the expected positive results. The increased forgetting (33.12% vs. 30.77%) and reduced stability (41.14% vs. 43.76%) suggest that distributed regularization can create unexpected optimization dynamics. This finding emphasizes the complexity of combining different regularization mechanisms and indicates that uniform constraints may undesirably alter the delicate stability-plasticity balance, particularly under varying experimental settings.

Table III provides crucial insights into FDRL's temporal learning patterns through task-wise analysis. The results reveal a pronounced recency bias, with improved performance on Task 0 (34.47% vs. 33.94%) and particularly Task 4 (68.87% vs. 67.0%), but significant degradation on intermediate tasks (Tasks 1-3). This pattern reflects the characteristic behavior

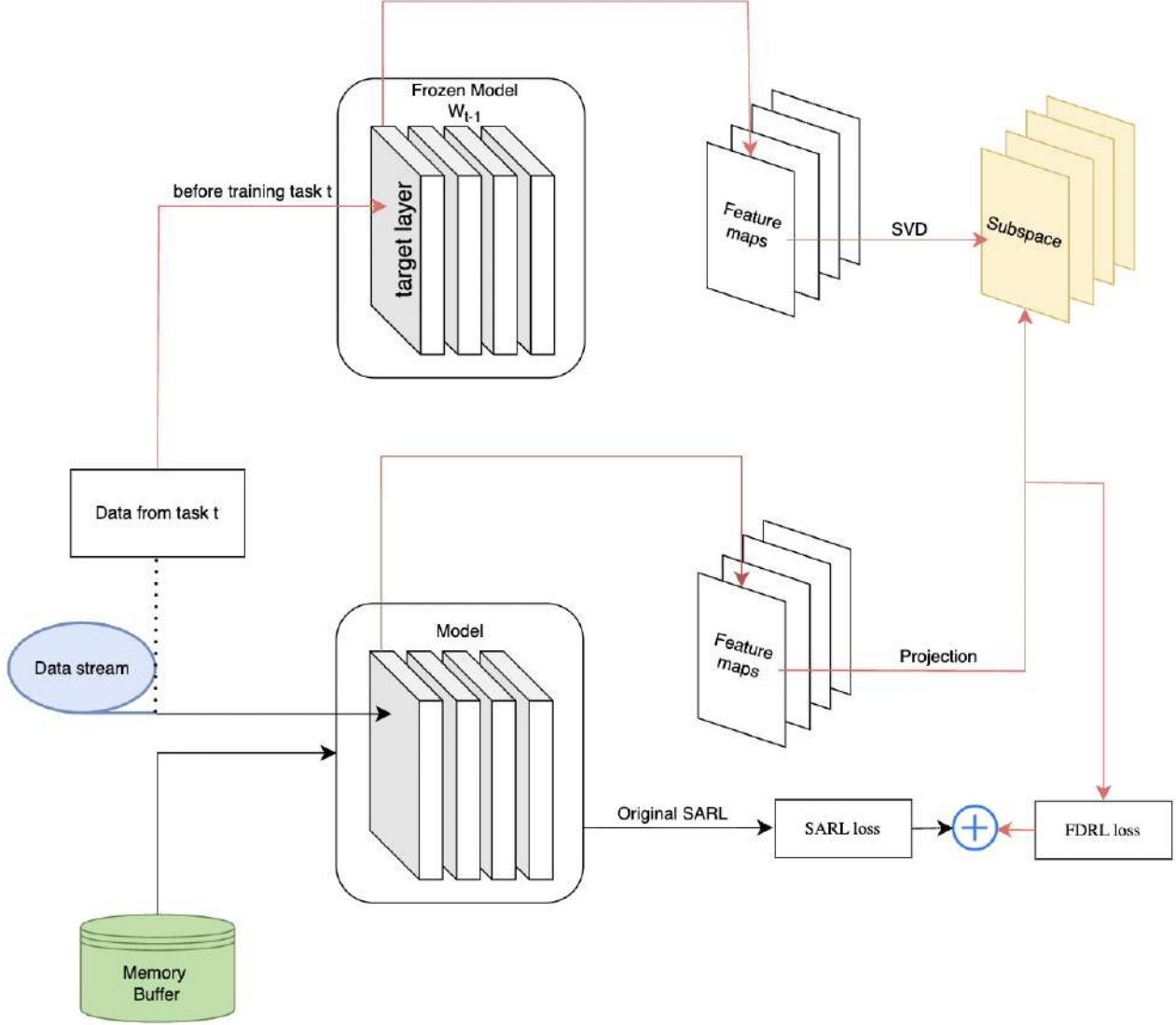


Fig. 1. The overall architecture of our proposed method(FDRL) training process, where stable feature subspaces are constructed from the frozen previous model using SVD, then used to regularize the current model’s features through projection, with the drift-resistant loss(FDRL loss) combined with original SARL loss components.

of drift-resistant mechanisms, where preserving recent feature representations may inadvertently compromise performance on middle-aged tasks. This reveals a critical trade-off: enhancing stability for recent knowledge can degrade performance on intermediate tasks, indicating that uniform regularization constraints create suboptimal dynamics in sequential learning. These findings suggest the need for adaptive regularization strategies that consider task age and importance.

IV. CONCLUSION

This paper introduced FDRL (Feature Drift-Resistant Space Regularization), a novel approach that adapts drift-resistant space concepts from parameter-efficient architectures to full-tuning continual learning frameworks. Our method success-

TABLE I
PERFORMANCE COMPARISON ON SEQ-CIFAR10 DATASET
WITH BUFFER SIZE 200

Buffer	Method	Acc(↑)	Fgt(↓)	Stb(↑)	PIs(↑)	S-P(↑)
200	ER [10]	50.36	57.60	38.40	96.44	54.88
	DER++ [11]	65.74	31.75	58.73	91.13	71.40
	CLS-ER [12]	65.92	28.74	67.70	81.17	73.83
	SCoMMER [13]	66.80	28.32	62.26	84.72	71.76
	SARL [1]	<u>69.13</u>	<u>16.44</u>	66.33	<u>82.28</u>	<u>73.45</u>
	FDRL(Ours)	69.67	14.83	<u>67.38</u>	83.55	74.60

fully addresses a significant gap in existing methodologies by integrating feature space regularization into the SARL framework through an exemplar-free mechanism that constructs

TABLE II
PERFORMANCE COMPARISON ON SEQ-CIFAR100 DATASET WITH BUFFER
SIZE 200

Buffer	Method	Acc($\uparrow$)	Fgt($\downarrow$)	Stb($\uparrow$)	Pls($\uparrow$)	S-P($\uparrow$)
200	ER	21.73	76.13	5.29	82.63	9.94
	DER++	30.68	65.10	17.19	82.77	28.44
	CLSER	43.80	36.33	45.75	50.47	47.99
	SCoMMER	40.74	42.20	35.98	63.30	45.80
	SARL	48.41	30.77	43.76	<u>73.03</u>	54.73
	FDRL(Ours)	47.39	<u>33.12</u>	<u>41.14</u>	73.88	<u>52.85</u>

TABLE III
SEQ-CIFAR100 TASK-WISE FINAL ACCURACY COMPARISON. FDRL
SHOWS SIGNIFICANT DEGRADATION ON INTERMEDIATE TASKS BUT
IMPROVED PERFORMANCE ON THE MOST RECENT TASKS.

Task	SARL (%)	FDRL (%)
Task 0	33.94	34.47
Task 1	39.65	37.6
Task 2	49.6	42.7
Task 3	53.3	51.88
Task 4	67.0	68.87

stable subspaces using only new task data.

REFERENCES

- [1] Sarfraz, F., Arani, E., Zonooz, B.: Semantic-Aware Representation Learning for Lifelong Learning. In: International Conference on Learning Representations (ICLR), (2025).
- [2] Xuan, L., Xiaobin, C.: LoRASubtraction for Drift-Resistant Space in Exemplar-Free Continual Learning. 10.48550/arXiv.2503.18985
- [3] Chen, Z., & Liu, B. (2018). Continual learning and catastrophic forgetting. In Lifelong Machine Learning (pp. 55-75). Cham: Springer International Publishing.
- [4] Lu, Y., Qian, B., Yuan, C., Jiang, H., & Wang, X. (2024). Controlled low-rank adaptation with subspace regularization for continued training on large language models. arXiv preprint arXiv:2410.16801.
- [5] Zhou, K., Wang, L., Zhang, X., Shum, H. P., Li, F. W., Li, J., & Liang, X. (2024, September). Magr: Manifold-aligned graph regularization for continual action quality assessment. In European Conference on Computer Vision (pp. 375-392). Cham: Springer Nature Switzerland.
- [6] Huang, L., Cao, X., Lu, H., & Liu, X. (2024, September). Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In European Conference on Computer Vision (pp. 214-231). Cham: Springer Nature Switzerland.huang2024class
- [7] Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., He, Y. (2024). Boosting continual learning of vision-language models via mixture-of-experts adapters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 23219-23230).
- [8] Tong, R., Liu, Y., Shi, J. Q., Gong, D. (2025, December). Coreset selection via reducible loss in continual learning. In The Thirteenth International Conference on Learning Representations.
- [9] Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. Trends in cognitive sciences, 24(12), 1028-1040.
- [10] Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., & Wayne, G. (2019). Experience replay for continual learning. Advances in neural information processing systems, 32.
- [11] Buzzega, P., Boschini, M., Porrello, A., Abati, D., & Calderara, S. (2020). Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems, 33, 15920-15930.
- [12] Arani, E., Sarfraz, F., & Zonooz, B. (2022). Learning fast, learning slow: A general continual learning method based on complementary learning system. arXiv preprint arXiv:2201.12604.
- [13] Sarfraz, F., Arani, E., & Zonooz, B. (2023, June). Sparse coding in a dual memory system for lifelong learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 8, pp. 9714-9722).

A Comparative Analysis of Deep Learning Models for Interpretable Multi-omics Cancer Subtyping using Integrated Gradients

Ha Tang Vinh, Hien Nguyen Minh, Diep Thi Hoang*

Faculty of Information and Technology

VNU University of Engineering and Technology, Vietnam National University of Hanoi

Hanoi, Vietnam

{22028129, 22021106, diepht}@vnu.edu.vn

*Corresponding author

Abstract—The integration of multi-omics data has advanced cancer subtype classification, yet the deep learning models employed often lack interpretability. Integrated Gradients (IG), an explainable AI (XAI) method, addresses this by attributing model predictions to input features, thereby facilitating biomarker discovery. While prior studies typically apply IG to a single architecture, this research systematically benchmarks IG across three distinct deep learning frameworks: MOGONET, MMDynamics, and HTML, using multi-omics data from The Cancer Genome Atlas (TCGA). We leverage IG to identify and compare top-ranked biomarkers, assessing their biological relevance and reproducibility across models. Our experiments, conducted with stratified K-fold cross-validation. This work provides practical guidance on selecting appropriate deep learning architectures and IG baselines for multi-omics research, aiming to foster more interpretable and clinically meaningful applications of AI in oncology.

Index Terms—integrated gradients, omics data, MOGONET, MMDynamics, HTML, cancer subtyping, biomarker discovery

I. INTRODUCTION

High-throughput next-generation sequencing revolutionized omics analysis, enabling large-scale molecular studies, particularly in cancer research [7], [8]. This progress fueled precision medicine, which personalizes treatments based on individual genetic, environmental, and lifestyle profiles. Pioneering initiatives like the Human Genome Project and The Cancer Genome Atlas (TCGA) laid the groundwork for this field [9]. By integrating multi-omics data with clinical records, precision medicine offers benefits for disease screening, detection, and treatment [10]. However, challenges like high dimensionality and noise have driven a shift from traditional methods to deep learning, which excels at learning non-linear relationships [11]. These "black box" models are difficult to interpret clinically, a limitation that has spurred the development of explainable artificial intelligence (XAI) for more transparent and reliable models [12].

Among XAI techniques, Integrated Gradients (IG) is a theoretically grounded and widely used method for interpreting deep learning models [13]. IG attributes prediction outcomes back to input features, helping researchers identify which genes or molecular markers most strongly influence model outputs [2],

[14]. However, a systematic evaluation of IG's performance across diverse deep learning architectures on the same multi-omics dataset is currently lacking.

In this study, we aim to fill this gap by conducting a comprehensive survey and experimental benchmark of applying Integrated Gradients to various deep learning architectures for multi-omics cancer subtype classification. Specifically, we:

- Apply IG to 3 deep learning architectures, including: MOGONET [3], MMDynamics (MMD) [5], HTML [15] to evaluate and compare how effectively IG highlights biologically meaningful biomarkers across these models.
- Conduct more rigorous experiments with the dataset split by the stratified K-fold method and ensure the transparency and reproducibility of the experiments.
- Provide practical insights and recommendations for selecting deep learning architectures and IG baselines for biomarker discovery in multi-omics research.

By integrating explainability with robust modeling, our study seeks to advance the interpretability and practical applicability of deep learning in cancer multi-omics. Ultimately, our goal is to support researchers and clinicians in identifying reliable biomarkers that can guide diagnosis and personalized treatment.

II. RELATED WORK

Various machine learning and deep learning models now leverage multi-omics data. MOGONET [3] integrates multi-omics data using graph-based networks and a cross-omics fusion module for patient classification and biomarker discovery. Another graph-based architecture, GREMI [4], is a two-phase explainable framework that first enhances disease prediction using a graph attention network on co-functional omics networks. It then employs a multi-view interpretation strategy with Monte Carlo Tree Search to identify biomolecular interaction modules that contribute most to disease outcomes.

HTML [15] is a multi-phase framework using dynamic architectures to customize feature and modality weights for each patient, enabling personalized cancer diagnosis and prognosis. Through extensive evaluation on numerous TCGA cohorts,

HTML demonstrated superior performance over static models and identified patient-specific, interpretable biomarkers. DP-NET [6] uses a dynamic attention mechanism to adaptively fuse multi-modal inputs by learning their representations, enhancing reliability under varying data quality. This architecture integrates confidence-aware fusion strategies to attenuate noisy or unreliable modalities, thereby improving overall classification performance.

Other studies use autoencoder-based architectures [2], [14]. Le et al. [2] uses denoising autoencoders to extract compact representations from each omics layer, which are then classified by a feed-forward neural network. In its second phase, biomarkers are identified by analyzing the classifier’s feature importance and validated on external cohorts, showing improved performance. Triphasic DeepBRCA [14] uses a three-stage pipeline for breast cancer stratification: (1) an autoencoder extracts low-dimensional embeddings, (2) a neural network classifies subtypes, and (3) a custom algorithm uses relevance propagation and filtering to identify robust biomarkers. Another approach is considering multi-omics data as multimodal data and use some dynamic mechanisms. Han et al. [5] propose a dynamic fusion approach for trustworthy multimodal classification that evaluates both feature-level and modality-level informativeness on a per-sample basis.

III. DATASET

We obtained all data from the UCSC Xena Hub and used four TCGA cancer cohorts. We chose breast cancer (BRCA) and lung cancer (Lung) because they have the largest number of samples in TCGA. BRCA also represents a dataset with clear class imbalance. To complement them, we included two smaller cohorts: glioblastoma (GBM) and colorectal adenocarcinoma (CRC). GBM represents a more balanced subtype distribution, while CRC illustrates an unbalanced one. This allows us to study both large- and small-sample settings under different class distributions. Table I contains the datasets’ information. For the GBM cohort, we utilized three omics data

TABLE I: Overview of the TCGA datasets after preprocessing.

Cancer Cohort	Subtypes (Count)	Sample Size	Selected Features
GBM	Classical: 71, Neural: 46, Proneural: 72, Mesenchymal: 81	270	GE: 2000, CNV: 2000, Meth: 2000
BRCA	Luminal A: 409, Luminal B: 187, Basal-like: 134, HER2-enriched: 67, Normal-like: 22	819	GE: 2000, CNV: 2000
Lung	Adenocarcinoma: 512, Squamous cell carcinoma: 498	1010	GE: 2000, CNV: 2000
CRC	CMS1: 37, CMS2: 110, CMS3: 38, CMS4: 79	264	GE: 2000, CNV: 2000

types: gene expression (GE), copy number variation (CNV), and DNA methylation (Meth). For the BRCA, Lung, and CRC cohorts, we used two omics types (GE and CNV), sourcing the pre-matched samples from Le et al. [2] to ensure larger cohort sizes, given the limited availability of methylation data

for these cancers. Gene expression data were obtained as log2-transformed RSEM-normalized counts from the TCGA Illumina HiSeq platform, which were already normalized for library size by the TCGA pipeline. CNV data were processed by the TCGA FIREHOSE pipeline using the GISTIC2 method to generate gene-level copy number estimates, and DNA methylation profiles were provided as beta values (0–1) from the Illumina HumanMethylation27 array, representing the proportion of methylation at each CpG site.

We preprocessed the data following a procedure similar to MOGONET [3]. Each omics type was preprocessed and filtered independently to reduce noise and redundancy. The data was split into training–validation (80%) and test (20%) sets using stratified sampling to maintain subtype distribution. For BRCA (the largest cohort), we used 5-fold cross-validation, while GBM and CRC (smaller cohorts) used 4-fold. The Lung dataset was also evaluated with 4-fold cross-validation, mainly because it has only two balanced classes, so fewer folds were sufficient while reducing computation time. After that, we ranked features in the training–validation set using ANOVA F-values with FDR correction, and retained the top 2,000 features per omics. To ensure feature diversity, we required that the first principal component explain less than 50% of variance. The same features were then selected in the test set. To train and evaluate the models, we employed a stratified k-fold cross-validation strategy on the training-validation set. Except for CNV data, features were scaled to [0,1] using min-max normalization based on training statistics, with the same parameters applied to validation and test sets to prevent data leakage.

IV. METHODS

We chose three models from distinct and widely recognized families of deep learning approaches for multi-omics analysis: graph-based integration - MOGONET, dynamic feature- and modality-level adaptation - MMD, and personalized multi-omics modeling with contrastive learning - HTML. We further note that these methods have publicly available implementations, which makes benchmarking and reproducibility feasible

A. MOGONET architecture for subtyping problem

MOGONET processes each omics modality with a separate graph convolutional network (GCN) that leverages both molecular features and sample–sample similarity. The outputs of these modality-specific GCNs are combined through a cross-omics discovery tensor, which is then passed to the VCDN for joint classification. This design enables modality-specific learning while capturing cross-omics interactions, and all components are optimized together in an end-to-end framework.

B. MMDynamics architecture for subtyping problems

Figure 1 is the illustration for MMDynamics deep learning architecture. MMDynamics introduces a sample-adaptive mechanism that dynamically adjusts the importance of individual features and entire modalities for each patient. At the feature level, it suppresses uninformative signals while emphasizing relevant ones, effectively performing feature selection per sample. At the modality level, it evaluates the reliability of

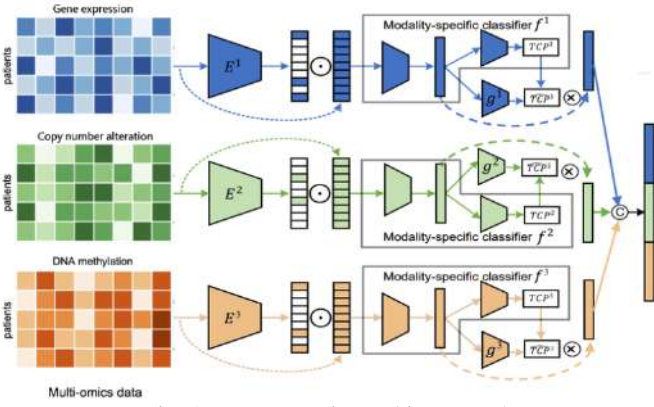


Fig. 1: MMDynamics architecture [5]

each omics type and assigns weights accordingly, ensuring that more informative modalities contribute more strongly to the final classification.

C. HTML architecture for subtyping problems

We also adopted the HTML (Highly trustworthy multi-omics learning) framework as described by Lu et al. [15]. Omics modalities were processed by modality-specific encoders to produce sample-wise embeddings. HTML leverages self-adaptive dynamic learning, applying adaptive weights at both the feature and modality level to tailor the model to each patient. A methylation-guided attention module further integrates regulatory relationships across modalities. To enhance cross-modal alignment, modality embeddings from the same sample are trained via a contrastive learning setup, while those from different samples serve as negatives. Final classification is carried out using a Dirichlet uncertainty estimator, allowing the model to output predictions with associated confidence.

D. Framework for discovering biomarkers

We utilize the same framework in our previous study [16]. More specifically, our framework includes three key phases. Phase 1 is utilizing a deep learning model [8] for molecular subtyping (after preprocessing TCGA cohort data) which was MOGONET in our previous study. We then adopted and replaced the model in phase 1 of our original framework with two other deep learning architectures mentioned above: MMDynamics and HTML to experiment and observe the changes of results in phase 1 and phase 3.

Phase 2 is to apply the integrated gradients (IG) algorithm for biomarker identification. IG is an explainable AI technique that attributes the contribution of each input feature to a model’s prediction by integrating gradients along a linear path from a baseline input x' to the actual input x . For a model F and input $x = (x_1, \dots, x_n)$, the attribution for feature i is defined as:

$$\text{IG}(x)_i := (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (5)$$

We applied IG to quantify the importance of each gene (from GE and CNV) in predicting cancer subtypes. Using a zero vector as baseline, we computed IG scores for each feature in models trained per subtype, then averaged the scores

TABLE III: Phase 1 results averaged over stratified k-fold splits on the four TCGA datasets. The results are reported on test datasets.

Dataset	Metric	MOGONET	MMDynamics	HTML
GBM (4-fold)	Accuracy	72.7	77.3	72.2
	F1 macro	69.9	76.3	68.7
	F1 weighted	72.2	77.4	70.9
BRCA (5-fold)	Accuracy	74.3	85.7	78.8
	F1 macro	53.5	86.3	63.0
	F1 weighted	71.9	85.8	77.0
Lung (4-fold)	Accuracy	95.8	93.0	96.9
	F1 macro	95.8	92.9	96.9
	F1 weighted	95.8	92.9	96.9
CRC (4-fold)	Accuracy	81.6	81.1	95.3
	F1 macro	74.0	68.6	95.2
	F1 weighted	79.4	77.4	95.2

across models and subtypes. Genes with the highest average attributions were considered potential biomarkers. We also experimented with the baselines we proposed in our previous paper [16] and the list of baselines can be found in TABLE II.

TABLE II: Baselines and their mathematical definitions

Baseline	Mathematical representation
Zeros	$a_j = 0$
Micro means	$a_j = \frac{1}{m} \sum_{i=1}^m A_{ij}$
Macro means	$a_j = \frac{1}{c} \sum_{g=1}^c \frac{\sum_{\text{class}(i)=g} A_{ij}}{\sum_{\text{class}(i)=g} 1}$
Default micro means (for class g)	$a_j = \frac{\sum_{\text{class}(i)=g} A_{ij}}{\sum_{\text{class}(i)=g} 1}$
Default macro means (for class g)	$a_j = \frac{1}{c-1} \sum_{g' \neq g} \frac{\sum_{\text{class}(i)=g'} A_{ij}}{\sum_{\text{class}(i)=g'} 1}$

The baselines were designed to address potential bias from class imbalance. Attribution scores from standard baselines may not reliably capture feature importance, and models often treat the majority class as a “default” reference. To mitigate this, we employed micro- and macro-mean baselines to balance contributions across samples or classes, as well as “default” variants that exclude the target class, allowing interpretation relative to the dominant class.

Phase 3 validates the identified biomarkers using two strategies: (i) comparison with curated gene sets from the Comparative Toxicogenomics Database (CTD) [17] and the COSMIC Cancer Gene Census [18], and (ii) evaluation of predictive performance with classical machine learning models. Our preliminary analysis showed limited overlap between computationally derived biomarkers and those reported in biological databases. For clarity and focus, we omit detailed database comparison results and instead emphasize the second strategy, which directly assesses the predictive value of the identified biomarkers.

V. EXPERIMENTS AND RESULTS

We conducted experiments with the same setting in our previous study [16].¹ Phase 1 results are presented in TABLE III. Since similar trends were observed in the phase 3 results across the four datasets, and the findings for the GBM dataset were already discussed in our previous study, we focus here on presenting the results of phase 3 for the remaining three

¹The implementation for this paper is available at: <https://www.kaggle.com/code/kaitonmh/xai-public>, <https://www.kaggle.com/code/kaitonmh/xai-mogonet>, <https://www.kaggle.com/code/kaitonmh/xai-html>, and <https://www.kaggle.com/code/tangha2005/xai-mmd>.

TABLE V: Phase 1 results averaged over stratified k-fold splits on the four TCGA datasets. The results are reported on train datasets.

Dataset	Metric	MOGONET	MMDynamics	HTML
GBM (4-fold)	Accuracy	93.7	99.2	100
	F1 macro	92.9	99.3	100
	F1 weighted	93.4	99.2	100
BRCA (5-fold)	Accuracy	92.4	99.9	94
	F1 macro	72.3	99.5	83.5
	F1 weighted	90.9	99.9	93.3
Lung (4-fold)	Accuracy	98.9	94.1	97.9
	F1 macro	98.9	94.1	97.9
	F1 weighted	98.9	94.1	97.9
CRC (4-fold)	Accuracy	83.4	79.7	99.7
	F1 macro	75.8	68.7	99.8
	F1 weighted	80.8	76	99.7

TABLE VI: Phase 1 results averaged over stratified k-fold splits on the four TCGA datasets. The results are reported on validation datasets.

Dataset	Metric	MOGONET	MMDynamics	HTML
GBM (4-fold)	Accuracy	82.9	85.2	87
	F1 macro	82.2	85.4	86.3
	F1 weighted	82.6	85.3	86.8
BRCA (5-fold)	Accuracy	85.2	94	90.9
	F1 macro	65.1	89.4	73
	F1 weighted	83.5	83.9	89.5
Lung (4-fold)	Accuracy	94.9	91.5	95.3
	F1 macro	94.9	91.4	95.3
	F1 weighted	94.9	91.4	95.3
CRC (4-fold)	Accuracy	85.4	69.7	94.8
	F1 macro	79.2	54.7	93.9
	F1 weighted	83.4	64.2	94.6

datasets: BRCA, lung, and CRC cancer, summarized together in the 3×3 plot (see Figure 2). This figure shows the F1 scores of the softmax classifier when using Q top biomarkers sorted by attribution scores across different values of Q .

The short comparison of using selected and non-selected features on the same classifier, softmax classifier can be found in TABLE IV:

TABLE IV: Comparison of softmax F1-score on the test set using original features versus different IG-selected features

Dataset	Original features	IG-selected MOGONET	IG-selected MMDynamics	IG-selected HTML
GBM	76.5	79.9	78.1	82
BRCA	83.8	86.5	84.7	83.1
Lung	97.1	98.47	97.5	98
CRC	97.4	97.5	100	97.5

VI. DISCUSSION AND CONCLUSION

A. Discussion

1) Phase 1 results

As shown in TABLE III, MMDynamics performs best on the GBM, BRCA, and Lung datasets, while HTML performs best on the CRC dataset. We hypothesized MMDynamics would outperform other models due to its sample-adaptive design, which is ideal for high-dimensional, low-sample-size genomic data. Additionally, MMDynamics has a relatively compact architecture compared to MOGONET and HTML, reducing overfitting risk—evidenced by stable performance. The overfitting trends of MOGONET and HTML can be seen when evaluating trained deep learning models on training set (TABLE V) and validation set (TABLE VI).

From TABLE V and TABLE VI, we observed that the performance gap between the training and validation sets was smaller for MMDynamics compared to MOGONET and HTML. It suggests that MMDynamics is less affected by overfitting. However, HTML’s superior performance on CRC was unexpected. A possible explanation is that HTML also incorporates sample-adaptive strategies, with a contrastive learning module that aligns multi-omics embeddings per sample. This likely leads to more coherent latent representations. In the CRC dataset, which may have simpler feature interactions, such alignment could be more effective than MMDynamics’ dynamic modeling. MOGONET consistently ranks second-best, though it also shows signs of overfitting. HTML shows the most overfitting (excluding CRC), likely due to its complex architecture.

2) Potential role of CTC-related miRNAs in CRC subtyping

A circulating tumor cell (CTC) is a cancer cell shed from a tumor into the bloodstream. CTCs are key drivers of metastasis and serve as promising biomarkers for monitoring tumor burden, treatment response, and prognosis in CRC. A recent study [19] shows that specific circulating microRNAs—like miR-199a-5p—are significantly downregulated in CTC-positive metastatic CRC patients and can accurately distinguish CTC status (AUC > 0.7). These findings suggest circulating miRNAs carry information on metastatic potential. Therefore, our framework could be extended by incorporating plasma-derived miRNA data for the CRC subtype to improve discrimination and prognosis prediction. Downregulated miRNAs from metastatic CRC could be prioritized as features for validation as non-invasive biomarkers. We expect these features would improve classification accuracy, especially for distinguishing CRC subtypes with different metastatic risks.

3) Phase 3 results

In Phase 3, we validated biomarkers by assessing their predictive power with traditional machine learning models. On the BRCA and GBM datasets, Phase 3 performance positively correlated with Phase 1 results. As Phase 1 classification performance improved from MOGONET to MMDynamics, a similar trend emerged in Phase 3, indicating IG effectively captured feature contributions. This underscores that a strong Phase 1 architecture is crucial, as poor initial results yield weak biomarker sets. Evidently, a good biomarker set—like that from MMDynamics—allows simple classifiers like SVM to match or even outperform complex deep learning models using the original 2000 genes.

Second, MMDynamics’ baselines consistently underperformed MOGONET’s zero baseline. This is likely due to MMDynamics’ dynamic nature, where feature weights vary per sample. Consequently, our approach of averaging IG scores may be suboptimal; summing the absolute IG values could be a better approach. We hypothesize that IG score variance is greater in dynamic models, making averaging less effective at capturing a feature’s contribution. Our findings suggest that simpler, fully connected architectures like MMDynamics tend to outperform graph-based ones like MOGONET.

The Lung and CRC datasets showed different patterns. As a

simple binary task, most models performed well on the Lung dataset in both phases. In contrast, CRC showed poor Phase 1 results for MOGONET and MMDynamics but surprisingly good performance in Phase 3. This may be due to hidden biological patterns in the CRC data, like the previously discussed CTC biomarkers, which require further study.

Using the same softmax classifier, we observed across all datasets that the best scores obtained with IG-selected features were slightly higher than those achieved with the original features in TABLE IV.

B. Conclusion

This study demonstrates the effectiveness of using IG for identifying informative biomarkers from deep learning-based classification models in multi-omics cancer subtyping. Among the evaluated models, MMDynamics consistently achieves strong performance due to its sample-adaptive mechanism and moderate architectural complexity, making it particularly well-suited for high-dimensional omics datasets. HTML, while more complex, performs best on the CRC dataset—possibly due to latent biological signals such as circulating miRNAs associated with metastatic behavior.

In Phase 3 analysis, we observed that the effectiveness of IG depends not only on the model but also on how IG scores are aggregated. Specifically, for models with fixed feature weights (e.g., MOGONET), averaging IG scores across samples yields reliable results. However, for sample-adaptive models like MMDynamics—where feature weights vary across samples—averaging may underrepresent the contribution of key features. In these cases, summing the absolute IG values may better capture a feature’s overall importance. This suggests that the choice of IG aggregation strategy should align with the model’s design to ensure more accurate biomarker identification.

Based on our findings, we recommend prioritizing applying IG on deep learning models like MMDynamics that balance adaptability and simplicity for most genomic classification tasks. More complex architectures such as HTML can be explored when initial results are unsatisfactory. Additionally, applying regularization techniques like early stopping and class-weighted loss remains essential to mitigate overfitting and address class imbalance.

Our results highlight the practical value of explainable AI in multi-omics research. By applying the Integrated Gradients algorithm, we were able to trace model predictions back to individual genes and rank them as potential biomarkers. This attribution step made it possible to reduce thousands of features to smaller, interpretable sets while still maintaining strong predictive performance in Phase 3. For example, biomarkers identified from MMDynamics allowed simple models such as SVM to reach performance comparable to or even exceeding the original deep learning models trained on the full feature set. This shows that explainable AI not only supports biological interpretation but also provides a way to validate and reuse informative features in downstream tasks. In addition, our analysis indicates that the choice of attribution aggregation strategy should match the model design, which

further demonstrates that explainability methods can guide researchers in selecting both models and biomarkers more effectively

In future work, we plan to incorporate somatic mutation data, as many driver mutations at the DNA level play a critical role in cancer development and could further improve the accuracy and clinical relevance of cancer subtyping. In conclusion, our framework and observations contribute practical guidance for integrating deep learning and explainability in biomarker discovery, with implications for both performance and interpretability in cancer subtype analysis.

REFERENCES

- [1] Franco, Edian F., et al. "Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data." *Cancers* 13.9 (2021): 2013.
- [2] Le, Hoang, et al. "A biphasic Deep Semi-supervised framework for Suptype Classification and biomarker discovery." *bioRxiv* (2022): 2022-01.
- [3] Wang, Tongxin, et al. "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification." *Nature communications* 12.1 (2021): 3445.
- [4] Liang, Hong, et al. "GREMI: an explainable multi-omics integration framework for enhanced disease prediction and module identification." *IEEE Journal of Biomedical and Health Informatics* (2024).
- [5] Han, Zongbo, et al. "Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
- [6] Zou, Xin, et al. "Dpnet: Dynamic poly-attention network for trustworthy multi-modal classification." *Proceedings of the 31st ACM international conference on multimedia*. 2023.
- [7] Elaine R Mardis. "A decade's perspective on DNA sequencing technology". In: *Nature* 470.7333 (2011), pp. 198–203
- [8] Ritchie, Marylyn D., et al. "Methods of integrating data to uncover genotype–phenotype interactions." *Nature Reviews Genetics* 16.2 (2015): 85–97.
- [9] Eric DGreen, James DWatson, and Francis S Collins. "Human Genome Project: Twenty-five years of big biology". In: *Nature* 526.7571 (2015), pp. 29–31
- [10] Beckmann, Jacques S., and Daniel Lew. "Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities." *Genome medicine* 8.1 (2016): 134.
- [11] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics". In: *Briefings in bioinformatics* 18.5 (2017), pp. 851–869.
- [12] Wojciech Samek et al. "Explaining deep neural networks and beyond: A review of methods and applications". In: *Proceedings of the IEEE* 109.3 (2021), pp. 247–278.
- [13] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." *International conference on machine learning*. PMLR, 2017.
- [14] S. Rajpal, M. Agarwal, V. Kumar, A. Gupta, N. Kumar (2021). *Triphasic DeepBRCA-A Deep Learning-Based Framework for Identification of Biomarkers for Breast Cancer Stratification*. IEEE Access. Volume 9.
- [15] Lu, Yuxing, et al. "Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes." *Briefings in bioinformatics* 24.6 (2023): bbad378.
- [16] H. Le, H. N. Minh, H. T. Vinh and D. T. Hoang, "Multi-Omics Biomarker Discovery with Explainable Artificial Intelligence: A Case Study in Glioblastoma," 2024 16th International Conference on Knowledge and System Engineering (KSE), Kuala Lumpur, Malaysia, 2024.
- [17] Davis, Allan Peter, et al. "Comparative toxicogenomics database (CTD): update 2023." *Nucleic acids research* 51.D1 (2023): D1257–D1262.
- [18] Tate, John G., et al. "COSMIC: the catalogue of somatic mutations in cancer." *Nucleic acids research* 47.D1 (2019): D941–D947.
- [19] Doğan, Berkcan, et al. "Candidate Biomarkers Associated With Circulating Tumor Cell Status in Metastatic Colorectal Cancer." *Journal of clinical laboratory analysis* 39.7 (2025): e70013.

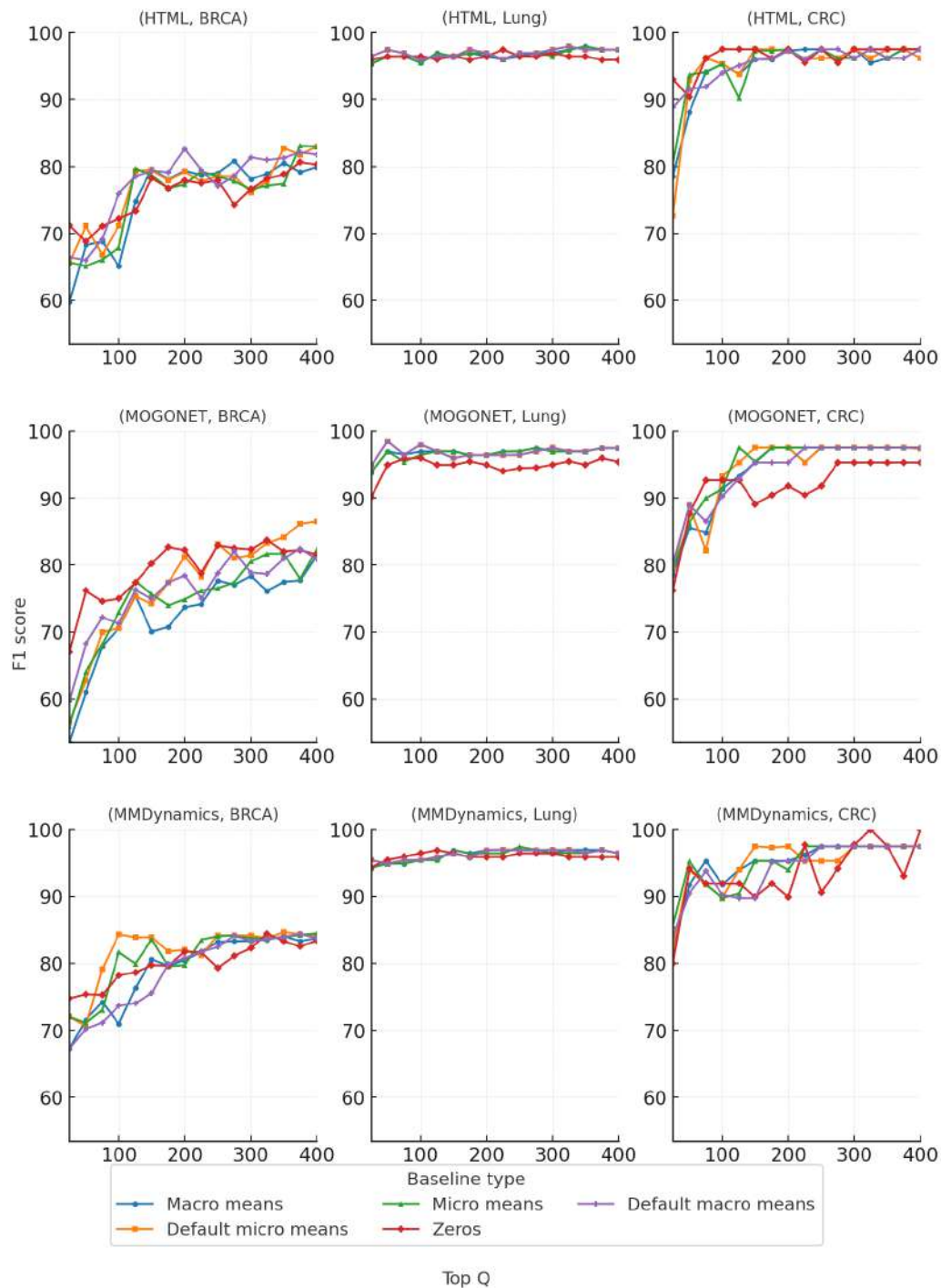


Fig. 2: F1 scores of the softmax classifier on three cancer datasets (BRCA, Lung, and CRC) across three architectures (HTML, MOGONET, and MMDynamics). Each panel shows the performance trend as a function of the number of top Q biomarkers ranked by attribution scores, with five curves corresponding to different baseline types.

AI-Based Eutrophication Monitoring to Support Aquaculture from Remote Sensing and Field Data: A Case Study of Hoa Binh Reservoir

1st Ngoc C. Lê

*Faculty of Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
lechingoc@yahoo.com*

2nd Trung Duong Dinh

*Department for Roads of Vietnam
Ministry of Construction
Hanoi, Vietnam
duongdinhtrung@gmail.com*

3rd Cuong Do-Chi

*ITSol
Hanoi, Vietnam
cuongdc@itsol.vn*

4th Pham Ngoc Linh

*Faculty of Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
phamngoclinh96th@gmail.com*

5th Nguyen Hong Son

*Faculty of Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
nguyenhongson.kstn.hust@gmail.com*

6th Minh Nhat Dang

*Data Science Institute
Columbia University
New York, NY, USA
nd2802@columbia.edu*

7th Anh Son Ta

*Faculty of Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
son.taanh@hust.edu.vn*

8th Thi Thu Ha Nguyen

*VNU University of Science
Vietnam National University
Hanoi, Vietnam
hantt_kdc@vnu.edu.vn*

9th Thu-Huong Thi Hoang

*School of Chemistry and Life Science
Hanoi University of Science and Technology
Hanoi, Vietnam
huong.hoangthithu@hust.edu.vn*

Abstract—Eutrophication of freshwater reservoirs is a critical environmental issue that can lead to harmful algal blooms, fish kills, and degraded water quality. In this study, a data-driven approach combining remote sensing data with machine learning (ML) models, ranging from baseline regressions to advanced ensemble and deep learning methods, is proposed to estimate key water quality indices from spectral data in Hoa Binh lake – one of Vietnam’s largest hydropower and aquaculture reservoirs [1]. Experimental results demonstrate that modern methods can achieve high prediction accuracy, significantly outperforming traditional approaches. It highlights the potential of integrating satellite remote sensing with ML for near-real-time water quality monitoring and early eutrophication warning, reducing the reliance on laborious field sampling.

Index Terms—Eutrophication, Water Quality, Remote Sensing, Machine/Deep/Ensemble Learning, Reservoir Aquaculture.

I. INTRODUCTION

Water quality degradation due to eutrophication is a pressing concern for lakes and reservoirs worldwide. Excess nutrients, particularly nitrogen and phosphorus from alluvium, agricultural runoff, aquaculture, and domestic waste, can trigger explosive growth of algae and cyanobacteria. This process can deplete dissolved oxygen levels, leading to dead zones that harm aquatic life, cause mass fish die-offs and public health risks. Effective monitoring and early warning of eutrophication are therefore crucial.

Hoa Binh is a prime example of a water body facing eutrophication pressures. Located on the Da (Black) river about 70 km west of Hanoi, it is Vietnam’s largest hydropower reservoir and also supports extensive cage aquaculture intensifying recently. Moreover, it also provides drinking water to Hanoi’s population. Hence the upstream nutrient inputs have raised concerns about water quality and algal bloom incidents and regular monitoring is needed [1]. Traditional methods involving field sampling and lab analysis of Chl-a, nutrients, etc., are labor-intensive, expensive, and inefficient for timely management. The reservoir differs markedly from other locations previously studied in Vietnam. Unlike shallow urban lakes such as West lake or Hoan Kiem lake in Hanoi, where eutrophication is mainly driven by domestic wastewater inputs, Hoa Binh is deep, subject to large seasonal inflows of nutrient-rich sediments from the Da river. Compared with other large reservoirs such as Tri An or Dau Tieng, primarily used for irrigation and water supply, Hoa Binh plays a dual role in both hydropower generation and intensive cage aquaculture. These factors create unique nutrient dynamics, including water level fluctuations from dam operations and concentrated nutrient loads near aquaculture cages. Consequently, this research not only provides the first comprehensive dataset for this major reservoir, but also contributes novel insights relevant for other mountainous hydropower reservoirs with aquaculture in Southeast Asia and highlights the significance of the case study and differentiates it from prior works.

Satellite sensors (e.g. Landsat, Sentinel-2) offer a promising alternative for synoptic water quality monitoring. They provide frequent, spatial observations for detecting changes in water surface related to algal pigments and suspended matter. Past research has shown that satellite observations can be used to estimate chlorophyll concentrations, turbidity, and trophic state indices (TSI) in lakes. In China, Sentinel-2's high-resolution multispectral data have been leveraged to monitor cyanobacterial blooms in Taihu lake [2]. It demonstrates that remote sensing, when calibrated with ground measurements, can complement or even replace some routine sampling.

Meanwhile, ML and AI techniques have increasingly been applied to improve water quality estimates from remote sensing. ML models (e.g. decision trees, random forests, neural networks) can learn intricate patterns from multispectral or hyperspectral data. Previous studies have reported that combining in-situ data with satellite imagery and applying ML significantly enhances prediction accuracy for parameters like Chl-a and TSI. For instance, Guo [2] successfully employed an artificial neural network (ANN) to forecast eutrophication levels in Lake Taihu using satellite data. Deng et al. [3] provide a comprehensive review of remote sensing and ML methods for lake water quality management, highlighting the effectiveness of advanced algorithms and the integration of multi-source data. In Vietnam, remote sensing has been used to assess surface water pollution and trophic state in some systems (e.g. Mekong River [4], Tri An [5], and Dau Tieng [6], West lake or Hoan Kiem lake [6]). However, to date there has been little work focusing on Hoa Binh using AI in combination with remote sensing.

Given this gap, this research aims to develop an AI-based framework for water quality prediction in Hoa Binh using remote sensing data, toward a practical early warning tool for eutrophication. Field measurements of water quality were leveraged to train and evaluate a variety of predictive models, including ensemble and deep learning (DL) methods. By comparing these approaches, the most suitable model for capturing the relationships between spectral data and water quality indicators is identified. The ultimate goal is to enable future operational monitoring where spectral data alone can reliably estimate water quality parameters (such as chlorophyll concentrations) without requiring frequent laboratory analyses, to help local authorities and stakeholders in fisheries to obtain timely information on water conditions and respond proactively to emerging water quality issues.

II. RELATED WORK

Satellite remote sensing has long been used to observe water quality in lakes and reservoirs. Instruments like Landsat (30m) and Sentinel-2 MSI (10–20m resolution) have multi-spectral bands that are sensitive to water constituents, including Chl-a, suspended sediments, and colored dissolved organic matter. For example, Landsat-8 imagery has been utilized to map chlorophyll concentrations and trophic states in large lakes like the North American Great Lakes. Ho et al. [8] analyzed decades of satellite data and found a global increase in

intense lake phytoplankton blooms, underscoring the value of remote sensing for detecting long-term eutrophication trends. Sentinel-2, higher spectral and spatial resolution, has shown greater potential in monitoring harmful algal blooms. A study on Lake Taihu demonstrated that Sentinel-2 could effectively track cyanobacterial bloom dynamics and drivers [10].

On the other hand, ML models are increasingly employed to enhance the estimation of water quality parameters from remote sensing. In Europe, researchers at Lake Geneva successfully combined satellite-derived indices with ML to improve predictions of eutrophication levels [9]. The methods can accommodate nonlinear relationships and multiple input features. Ensemble learning techniques such as random forests (RF) and gradient boosting (e.g. XGBoost) have been popular due to their high accuracy and robustness. Prior studies have reported the models successfully retrieving chlorophyll and turbidity from satellite reflectance with better accuracy than traditional regression [11]. DL has also emerged in this domain. Deep neural networks can learn complex representations and have been used to analyze spectral imagery for water quality. ANNs were introduced in eutrophication studies at least as early as 2007 for Taihu lake [2]. Convolutional neural networks (CNN) are particularly well-suited for spatial (hyperspectral) image analysis and have also been applied to spectral data and achieved high accuracy in water quality classification. Likewise, recurrent neural networks (RNN) and especially long-short term memory (LSTM) are useful when temporal sequences are involved, such as forecasting algal bloom trends from time series of satellite data. Recent comparative studies indicate that advanced models like CNN, LSTM, and tree ensembles tend to outperform simpler models (linear or single-tree) for predicting indices like TSI from remote sensing data. For example, Deng et al. [3] noted that combining in-situ measurements with satellite data and using sophisticated ML algorithms can substantially improve water quality predictions. In Vietnam, there has been growing interest in applying these techniques to local water bodies. Vinh et al. [7] used Landsat-8 imagery and regression models to monitor the trophic state of several shallow lakes in Hanoi. Do et al. [12] explored ML algorithms (support vector machines (SVM) and RF) to assess surface water pollution in the Hanoi area with satellite data. These studies proved the feasibility of remote sensing-based water quality monitoring in Vietnam's context.

III. DATASET AND MODELS

A. Study Area and Field Data Collection

Hoa Binh is about 70 km in length and 1–2 km in average width. It has multiple branching sub-basins formed by flooded river valleys. For this study, field sampling campaigns were conducted to collect water quality data at various locations covering the upstream, midstream, downstream of the reservoir and near-dam, near-cage-aquaculture sections. In total, six campaigns with 177 survey points were conducted to capture spatial variations. Sites were selected to ensure representation of areas with expected differences in nutrient dynamics, including zones affected by river inflows (nutrient

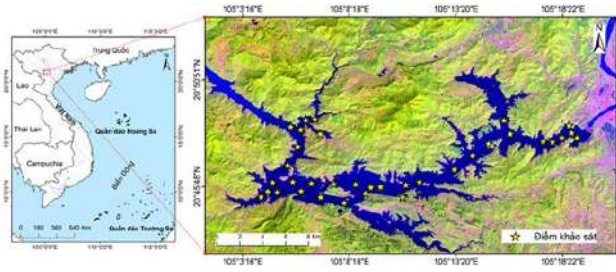


Fig. 1. Map of Hoa Binh Reservoir with water sampling sites (yellow stars) used in this study. The reservoir is located in northern Vietnam (inset) and spans approximately 70 km in length. Sampling points cover upstream (west/southwest) to downstream near the dam (east).

loading), aquaculture activities (localized enrichment), and hydropower regulation (fluctuating water levels). Conducted between 2023–2024, the campaigns’ temporal coverage included both dry and rainy seasons, July 2023 (high) and May 2024 (low inflow period), reflecting contrasting hydrological conditions. This spatially stratified and seasonally repeated design increases the representativeness of the dataset and strengthens the robustness of AI models.

At each site, surface water samples were taken just below the surface (at 0.3–0.5m depth) using a Van Dorn sampler. The sampling depth was chosen to correspond to the penetration depth of light (within the Secchi disk visibility depth) so that water quality measurements represent the water column seen by the satellite. Water clarity was measured in the field with a Secchi disk depth (SDD) recorded). In-situ measurements of basic parameters (water temperature, pH, dissolved oxygen (DO)) were also noted, but the primary analysis focused on laboratory-derived water quality indicators. The collected water samples were kept on ice and transported to the lab on the same day for analysis.

In the laboratory, for each sample, the following water parameters were measured: Chl-a ($\mu\text{g/l}$), which is a proxy for algal biomass and a key indicator of trophic state; Total Phosphorus (TP, $\mu\text{g/l}$), the main nutrients driving eutrophication; and Total Suspended Solids (TSS, mg/l) which relate to turbidity. Chl-a was determined by filtering a known volume of water and extracting pigments with acetone, then measuring optical absorbance using spectrophotometry (following the Vietnamese standard TCVN 6662:2000). TP were analyzed using persulfate digestion and colorimetric methods (TCVN 6625), and TSS was measured gravimetrically by filtering water through pre-weighed GF/F filters and drying them. Then, TSI of [14] was calculated using formulas for SDD, Chl-a, and TP. The TSI is a dimensionless index commonly used to classify waters as oligotrophic, mesotrophic, or eutrophic.

Table I presents descriptive statistics of the measured water quality parameters (Chl-a, TP, and TSS) from the sampling campaigns. Chl-a varied widely, from (very oligotrophic) to $80.73 \mu\text{g/L}$ (hypereutrophic), with a mean of 6.92 and high variability ($\text{SD} \approx 15.67$), indicating that most samples were low but a few exhibited elevated concentrations. TP ranged from 0.0 to $70 \mu\text{g/L}$ (mean 30), reflecting generally low

phosphorus availability, characteristic of phosphorus-limited reservoirs. TSS varied between 0.87 and 3.80 mg/L (mean 2.26), with relatively low dispersion ($\text{SD} \approx 0.78 \text{ mg/L}$). SDD ranged from about 0.5m (very turbid) to greater than 2m in the clearest sites. These statistics confirm that the dataset captures both oligotrophic and mesotrophic conditions, with occasional eutrophic cases, providing a representative basis for training and evaluating the AI models.

TABLE I
DESCRIPTIVE STATISTICS OF WATER QUALITY PARAMETERS IN HOA BINH RESERVOIR (N = 177)

Parameters	Chl-a	TP	TSS
Min	0.63	0.0	0.87
Max	80.87	70	3.80
Mean	6.92	30	2.26
Median	3.36	30	2.47
SD	15.67	17	0.78

B. Remote Sensing Data

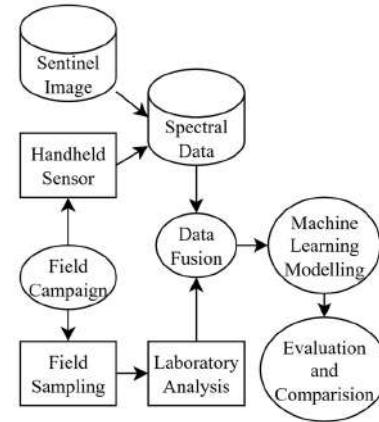


Fig. 2. Research Workflow

In parallel with field sampling, remote sensing data characterizing the water’s optical properties were obtained. Because satellite overpass timing was considered, many of our surveys were coordinated with satellite observation dates.

Before using, standard preprocessing were applied to satellite data: atmospheric correction was performed to convert top-of-atmosphere reflectance to surface reflectance using the ESA Sen2Cor processor for Sentinel-2. This step corrects for atmospheric scattering and absorption, ensuring that spectral values more accurately represent water-leaving reflectance. Then the reflectance values for the pixels corresponding to each sampling site were extracted. Since the reservoir water is optically deep, the reflectance at the sampling coordinates can be directly related to water quality without bottom reflectance interference. A 3×3 pixel window around each GPS location was averaged to reduce any geolocation mismatch and noise. The features taken from Sentinel-2 for each site include reflectance in the visible and near-infrared bands: specifically the Blue (Band 2, 490 nm), Green (Band 3, 560 nm), Red (Band 4, 665 nm), Red-edge/NIR bands

(Band 5–8a, 705–865 nm), and Shortwave IR bands (Bands 11,12 at 1610 and 2190 nm). These wavelengths capture the spectral signatures of chlorophyll absorption (red band, red-edge) and turbidity (red and NIR reflectance) among others. Several spectral indices reported in literature as related to water quality, such as the Normalized Difference Chlorophyll Index (NDCI) and turbidity indices, were computed to use as potential input features.

Besides satellites, field spectroradiometer data were also utilized for a subset of points. During the field campaigns, a handheld spectrometer (Analytical Spectral Devices FieldSpec or GER1500) was used at each site to measure the water's spectral reflectance in-situ across 350–1050 nm wavelengths. These high-resolution spectra (1 nm sampling) provide detailed information on absorption and backscatter features of the water. The reflectance in bands corresponding to the Sentinel-2 bandwidths were averaged to validate and supplement the satellite reflectance. The in-situ spectra were particularly useful on a few dates where satellite data were not available due to cloud cover. In essence, independent variables (features) for modeling consist of the multi-band reflectance of the water (from satellite and/or field spectral measurements) and derived indices. All data were compiled into a single dataset, with each sample having a set of spectral features and corresponding measured water quality values.

C. Data Preprocessing

The dataset was carefully preprocessed before modeling. All numeric features (reflectance values and indices) were normalized to a consistent scale (using min-max scaling to [0,1]). Outlier detection—any samples were also performed where field notes indicated potential errors (e.g., sample bottle contamination) or where reflectance data were affected by adjacency to land (mixed pixels near shoreline) were examined. Two samples were excluded due to anomalous readings (one with abnormally high turbidity likely due to a local disturbance during measurement). The remaining dataset had no missing values for the key parameters.

D. Machine Learning Models

A broad range of ML models to map the relationship between remote sensing inputs (spectral features) and water quality outputs was implemented. The models can be grouped into three categories: (1) Baseline, (2) Ensemble tree-based, and (3) DL models. All were developed using Python (scikit-learn and TensorFlow libraries) and trained/tested on the same compiled dataset. Hyperparameters were determined through a combination of grid search and cross-validation. Given the limited sample size, we restricted model complexity to compact architectures (parameter counts $\approx 8\,000$ – $33\,000$) to avoid overfitting. For each model, candidate hyperparameters (hidden units, filter sizes, dropout rates, learning rates) were first screened using 5-fold cross-validation on the training data. Early stopping with a patience of 10 epochs was applied to automatically terminate training when validation error stopped improving. The final hyperparameters were selected

as those that minimized cross-validated RMSE (for regression tasks) or maximized macro-F1 score (for classification tasks). This systematic approach ensured reproducibility and prevented bias from manual tuning.

Linear Models (LM). As simple baselines, linear regression was implemented to predict water quality indices from the spectral features and logistic regression for classification task.

Decision Tree (CART). More complex baseline, the single decision tree model was used for both classification and regression tasks.

RF. The model can automatically leverage the most informative spectral bands for each target (for example, it might find the red-edge band most useful for chlorophyll). The number of trees and maximum depth were tuned via cross-validation, finding that 100 trees of depth up to 5 was optimal.

XGBoost. The model was configured with 100 boosting rounds, a learning rate of 0.1, max depth 3, and early stopping based on validation error.

SVM and SVR. Support Vector Regressor and SVM were used for regression and classification (RBF kernel) respectively. In trials, SVM achieved reasonable accuracy but was slower and did not significantly outperform the tree ensembles.

ANN. Feed-forward neural networks were developed using soft max for classification (or ReLU for regression task) output layers. The network architectures had an input layer with nodes for each feature (reflectance bands), two hidden layers with 32 and 16 neurons respectively (ReLU activation), and an output linear layer for the target variable. The network using the Adam optimizer (learning rate = 0.001) with a mean squared error loss was trained over 500 epochs with early stopping. Total parameter count $\approx 33\,000$. Careful regularization (L2 weight decay and dropout of 20% in hidden layers) was applied to prevent overfitting.

ANFIS. The ANFIS model in this research uses Gaussian for input and fuzzification layer to give out three membership functions ("low", "medium", "high") per input. Normalization layer fires strength of rules using soft-max function. Output layer combines output from all rules by averaging.

CNN. To exploit the contiguous spectral nature of the data, a simple 1-D CNN that treats the reflectance spectrum as a one-dimensional image was deployed. The model had one convolutional layer, 16 filters (size 3), followed by a second convolutional layer, 8 filters (size 3), then a max pooling layer to downsample, and finally a fully-connected layer, 32 neurons, and a linear output. Total parameter count $\approx 10\,000$. Dropout (rate = 0.3) was applied after the convolutional and dense layers. The model was trained with Adam (lr = 0.001), batch size = 8, for 150 epochs with early stopping.

LSTM. The model extracting feature spectral data analysis consists of (1) Input layer: accepts features derived from spectral bands; (2) One LSTM layers, 64 memory cells and tanh activation; (3) Dropout layer (e.g., rate = 0.2) for regularization; and (4) Fully connected dense output layer, linear activation function, to predict continuous target

variables (e.g., Chl-a, TP, etc.). The model is optimized using Adam ($\text{lr} = 0.001$) and Mean Squared Error (MSE) as a loss function. Total parameter count ≈ 8500 . The LSTM was trained with batch size = 10, for 150 epochs with early stopping.

Ensemble Stacking (SE). In addition to individual models, a stacked ensemble one were created to combine the strengths of multiple approaches. Simple stacking consists the predictions of several top models (specifically RF, XGBoost, ANN, LSTM, and CNN) on the training set were used as inputs to meta-learners (which were Ridge Linear Regression for the regression and logistic regression with L2 regulation for classification task). The stacking model was evaluated via cross-validation to ensure it generalizes.

IV. EXPERIMENTS AND RESULTS

A. Regression Task.

Table II summarizes the performance of various models in predicting Chl-a, TP, and TSS concentration from the spectral data. The average RMSE (in $\mu\text{g/l}$ for Chl-a and TP and mg/l for TSS) and R^2 (in percentage) from 5-fold cross-validation for each model were reported.

As shown, the ensemble and deep learning models achieved the best accuracy in our experiments. The LR baseline reached an $R^2 \approx 60$ for Chl-a, meaning it could explain 60% of the variance - a reasonable result, but with room for improvement. Corresponding results for TSS, 50%, and TP, 35% are not so good. Its RMSE (12 $\mu\text{g/l}$) is quite high relative to the range of chlorophyll values, indicating significant errors especially at high concentrations (the linear model tended to under-predict the highest chlorophyll readings, as it cannot model saturation effects). Similarly for TSS and TP.

CART did not improve much over linear regression ($R^2 = 65\%$). This suggests that while there is some nonlinearity that a tree can capture, the model likely overfit some noise and was not very generalizable. Indeed, the performance varied more between folds. A little bit better than CART is SVR with $R^2 = 70\%$.

Both RF and XGBoost performed substantially better. RF achieved $R^2 = 73\%$, roughly halving the RMSE much better compared to LR. This indicates that tree ensembling

was able to capture the complex relationships in the data. XGBoost slightly outperformed RF (mean $R^2 = 78\%$, RMSE = 5.5 $\mu\text{g/l}$). The improvement of XGBoost may be due to its more nuanced fitting.

The DL models also yielded strong results. The ANN reached $R^2 = 70\%$, better than LR but slightly below RF. ANFIS is slightly better but still weaker compared to XGBoost. The CNN and LSTM, as expected are better exploit the spectral shape, achieved R^2 at 77% and 80%, respectively. This indicates the deep models are indeed effective, although not dramatically better than the best tree-based model in this case. One reason could be the limited dataset, the neural networks had to be kept fairly small, and even then, some signs of overfitting are observed if training too long. Early stopping and drop out are employed to mitigate this. The CNN had a slight edge over the basic ANN and ANFIS, implying that the convolutional filters were beneficial in extracting features. LSTM are slightly better since it is suitable for sequential data like spectra. These spectral features correlate with chlorophyll and were likely picked up by the CNN and LSTM automatically, whereas the ANN and ANFIS might have needed more neurons to capture such patterns explicitly. Finally, the best distribution in this research is that a combination of models could yielding a further slight improvement. The SE, which took best models, say RF, XGBoost, ANN, LSTM, and CNN outputs as inputs to a linear blender, achieved the best overall R^2 of 82% and the lowest RMSE (5 $\mu\text{g/l}$). It indicates that single models had somewhat complementary strengths, for example, in a few fold cases the CNN predicted extreme chlorophyll values better, whereas LSTM or XGBoost were more accurate in moderate ranges. The ensemble can balance that, and for small datasets, it reduces reliance on a single model's inductive bias.

The performance rankings for TSS (turbidity) prediction were broadly similar: LR $R^2 = 50\%$, improving to 78% – 87% with RF, XGBoost, and CNN, LSTM. Turbidity correlates with reflectance in the red and NIR bands, which our models captured. For TP, the prediction was more challenging (the spectral signals of nutrients are indirect, mainly via their influence on algae or water clarity) and the best models achieved R^2 of about 75%.

B. Trophic State Classification.

For classification of trophic state, accuracy, precision, recall, and F1-score were evaluated (Table III). Each sample's trophic state was categorized based on the Carlson TSI (oligotrophic/mesotrophic/eutrophic) [14]. All models achieved moderate accuracy (around 71–88%), which is in line with previous studies (e.g. appx. 74% accuracy reported using ANN/RF/SVM on Sentinel-2 data [15]).

Simpler models like LM struggled to identify the few eutrophic cases (yielding lowest F1 = 70%). CART had accuracy (74%). Ensemble trees (RF, XGBoost) and SVM (similarly for ANN) achieved higher (76–79%). Similar to the regression task, among single models, the CNN attained the highest accuracy (82%). It appears that the CNN could better recog-

TABLE II
MODEL PERFORMANCE ON REGRESSION TASK

Machine Models	Chl-a		TSS		TP	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
LM	12 $\pm$ 2	60	0.56 $\pm$ 0.1	50	13 $\pm$ 3	35
CART	9.0 $\pm$ 1.5	65	0.43 $\pm$ 0.08	60	11 $\pm$ 2	40
SVR	7.5 $\pm$ 1.2	70	0.39 $\pm$ 0.07	65	10 $\pm$ 2	40
RF	6.5 $\pm$ 1	73	0.37 $\pm$ 0.06	68	9.5 $\pm$ 1.5	55
XGBoost	5.5 $\pm$ 1	78	0.33 $\pm$ 0.05	72	8.5 $\pm$ 1	60
ANN	7 $\pm$ 1.3	70	0.39 $\pm$ 0.06	65	10 $\pm$ 2	50
ANFIS	5.5 $\pm$ 0.9	75	0.35 $\pm$ 0.05	70	8.5 $\pm$ 1	60
CNN	6 $\pm$ 1.1	77	0.30 $\pm$ 0.05	75	8 $\pm$ 1	65
LSTM	5 $\pm$ 0.8	80	0.28 $\pm$ 0.04	77	7.5 $\pm$ 0.8	68
SE	4.5 $\pm$ 0.8	82	0.25 $\pm$ 0.04	80	6.5 $\pm$ 0.7	75

TABLE III
MODEL PERFORMANCE ON CLASSIFICATION TASK

Models	Accuracy	Precision	Recall	F1-Score
LM	0.71	0.69	0.72	0.70
CART	0.74	0.73	0.75	0.74
SVM	0.76	0.75	0.77	0.76
RF	0.77	0.76	0.78	0.77
XGBoost	0.79	0.77	0.80	0.78
ANN	0.75	0.74	0.75	0.74
CNN	0.82	0.81	0.83	0.82
SE	0.88	0.87	0.89	0.88

nize spectral signatures of high-nutrient (eutrophic) samples – likely by detecting pigment absorption features – leading to slightly improved recall of that class. The SE adapted for classification gave 88% accuracy. Given the high regression performance, these classification results are expected and demonstrate that the models can be used not only to predict exact values but also to issue category-based alerts.

V. CONCLUSION

In this paper, a comprehensive approach to predict water quality parameters in Hoa Binh reservoir using remote sensing data integrated with ensemble and DL models is represented. It is demonstrated that the models can accurately estimate key indicators of eutrophication such as Chl-a concentration from multi-spectral reflectance inputs. The best models achieved high predictive performance (R^2 over 80%) for chlorophyll and related measures, significantly exceeding the accuracy of simple approaches. These results affirm the potential of AI methods, especially for the SE model, to reduce reliance on labor-intensive field sampling.

The findings are especially pertinent for aquaculture management and environmental monitoring. An operational system based on our models could continuously ingest satellite imagery (e.g., Sentinel-2's 5-day revisits) and output near-real-time maps of chlorophyll, turbidity, and inferred trophic state. This would allow managers to identify emerging algal blooms or areas of deteriorating water quality early and take preventive actions (such as adjusting fish feeding practices or aeration in fish cages) to mitigate risks. The framework is general and can be transferred to other lakes and reservoirs facing eutrophication, with appropriate re-calibration using local data. There are several avenues to extend this research. Firstly, accumulating a larger dataset covering more seasons would allow us to further improve and validate the DL models. More temporal data would also enable incorporating time series to explicitly model seasonal trends in water quality. Secondly, additional remote sensing sources, such as satellite-derived temperature and precipitation data, as proxy drivers for algal growth, are also considered. Another direction is to refine the trophic state index forecasting – using the developed models to not just map current conditions but predict short-term changes (e.g., forecasting chlorophyll a week in advance) by combining meteorological forecasts with ML models.

ACKNOWLEDGMENT

This research was supported by NAFOSTED Project, code: 12/2022/TN.

REFERENCES

- [1] D. T. N. Phan, B. D. Nguyen, P. Liew, H. T. T. Pham, T. T. Nguyen, H. T. Nguyen "Evaluating the potential of aquaculture in the Hoa Binh reservoir with carrying capacity and water quality indices", Vietnam Journal of Science, Technology and Engineering, Vol. 63(4), pp. 66–71, May 2021.
- [2] L. Guo, "Doing battle with the green monster of Taihu Lake", Science, Vol. 317, No. 5842, p. 1166, 2007.
- [3] Y. Deng, Y. Zhang, D. Pan, S. X. Yang, and B. Gharabaghi, "Review of recent advances in remote sensing and machine learning methods for lake water quality management", Remote Sensing, Vol. 16, No. 22, p. 4196, 2024.
- [4] N. T. B. Phuong, V. P. D. Tri, N. B. Duy, and N. C. Nghiem, "Remote Sensing for Monitoring Surface Water Quality in the Vietnamese Mekong Delta: The Application for Estimating Chemical Oxygen Demand in River Reaches in Binh Dai, Ben Tre", Vietnam Journal of Earth Science, Vol. 39, No. 13, pp. 256–268.
- [5] B. P. Duc, "Comparison of Synthetic Aperture Radar Sentinel-1 and ALOS-2 observations for lake monitoring", Vietnam Journal of Earth Sciences, Vol. 46, No. 3, pp. 322–338, 2024.
- [6] T. N. Hoang, and V. Tran Thi, "Spatial Estimation of Trophic State for Reservoir Using Ground Monitoring and Remote Sensing Data", Applied Environmental Research, Vol. 47, No. 1, 2025.
- [7] P. Q. Vinh, N. T. T. Ha, N. T. P. Thao, N. T. Linh, L. T. Oanh, L. T. Phuong, and N. T. T. Huyen, "Monitoring the trophic state of shallow urban lakes using Landsat 8/OLI data: a case study of lakes in Hanoi (Vietnam)", Frontiers of Earth Science, Vol. 19, pp. 25–40, 2025.
- [8] J. C. Ho, A. M. Michalak, and N. Pahlevan, "Widespread global increase in intense lake phytoplankton blooms since the 1980s" Nature, Vol. 574, pp. 667–670, 2019.
- [9] A. I. Rahaghi, D. Odermatt, O. Anneville, O. S. Steiner, R. S. Reiss, M. Amadori, M. Toffolon, S. Jacquet, T. Harmel, M. Werther, F. Soullignac, E. Dambrine, D. Jezequel, C. Hatte, V. Tran-Khac, S. Rasconi, F. Rimet, and D. Bouffard, "Combined Earth observations reveal the sequence of conditions leading to a large algal bloom in Lake Geneva", Communications Earth & Environment, Vol. 5, p. 229, 2024.
- [10] T. Song, H. Zhang, Y. Xu, X. Dai, F. Fan, Y. Wang, G. Liu, "Cyanobacterial blooms in Lake Taihu: Temporal trends and potential drivers", Science of The Total Environment, Vol. 942, 2024.
- [11] M. Allam, L. Zhang, X. Sun, O. Kisi, N. L. Kushwaha, L. Yongxin, M. Menenti, and S. Heddiam. "Enhancing chlorophyll-a predictions using optimal machine learning models and field spectral reflectance", Earth Science Information Vol. 18, p. 384, 2025.
- [12] T.-N. Do, D.-M. T. Nguyen, J. Ghimire, K.-C. Vu, L.-P. D. Dang, S.-L. Pham, and V.-M. Pham, "Assessing surface water pollution in Hanoi, Vietnam, using remote sensing and machine learning algorithms", Environmental Science and Pollution Research, Vol. 30, pp. 82230–82247, 2023.
- [13] N. T. P. Thao, D. H. Toan, N. T. T. Ha, P. Q. Vinh, H. C. Tien, T. M. Tung, T. T. Hien, "Modeling Spatial Distribution of Suspended Sediments in Hoa Binh Reservoir Using Sentinel-2 Image and Co-Kriging", VNU Journal of Science: Earth and Environmental Science, Vol. 40, No. 4, pp. 56–65, 2024.
- [14] R. E. Carlson "A trophic state index for lakes", Limnology and Oceanography, Vol. 22(2), pp. 361–369, 1977.
- [15] A. C. Cruz, C. Portela, C. C. F. Barbosa, D. A. Maciel, R. L. do N. Wanderley, R. Paulino, L. G. da Silva, V. S. Martins, "Machine Learning Classification of Trophic State Index from Sentinel-2 MSI Image", Proceedings of the XX Brazilian Symposium on Remote Sensing, INPE 2023.

ELLA: Enhancing Long-Tailed Online Continual Learning with Layer Augmentation

Dinh-Nhat Loi, Tri-Thanh Nguyen, Thanh-Hai Dang, Quynh-Trang Pham Thi*

Vietnam National University, University of Engineering and Technology

Hanoi, Vietnam

{2021152@vnu.edu.vn, ntthanh@vnu.edu.vn, hai.dang@vnu.edu.vn, trangptq@vnu.edu.vn}

Abstract—Online continual learning (OCL) systems must adapt to sequential data streams while mitigating catastrophic forgetting. However, existing OCL methods assume balanced class distributions, limiting applicability to real-world long-tailed scenarios. We formalize Long-Tailed Online Continual Learning (LT-OCL), wherein models learn from imbalanced streams while being evaluated on balanced test sets. We propose ELLA (Enhancing Long-tailed Learning with Layer Augmentation), a decoupled framework with three innovations: (1) dual-stage training combining contrastive learning with equalization loss, then fine-tuning a balanced classifier; (2) adaptive equalization dynamically rebalancing gradients based on observed class frequencies; and (3) rehearsal feature augmentation generating perturbations at intermediate layers to improve generalization on exemplars. Experiments on CIFAR-100-LT and VFN-LT demonstrate that ELLA substantially outperforms state-of-the-art OCL methods across various memory budgets and task configurations.

Index Terms—online continual learning, long-tailed classification, contrastive learning, catastrophic forgetting, rehearsal feature augmentation

I. INTRODUCTION

Humans continuously acquire and assimilate knowledge from sequential, predominantly unlabeled, and rarely recurrent information streams throughout their lifespans [4]. In contrast, Artificial Intelligence (AI) systems typically operate within narrower task-specific domains and employ multi-stage learning paradigms, wherein knowledge acquisition is conducted on static datasets in batch processing modes. Online continual learning (OCL) [9], [10], [14] provides a mechanism that enables AI models to emulate human-like learning processes by facilitating continuous knowledge acquisition from perpetual data streams. However, continual learning is confronted with a fundamental challenge: catastrophic forgetting, whereby models experience significant degradation of previously acquired knowledge. Contemporary research has increasingly directed attention toward continual learning within online scenarios. In the online configuration, each data sample undergoes a single pass through the training process, precluding subsequent reuse. This online continual learning paradigm demonstrates substantial alignment with real-world constraints of data accessibility and computational resource limitations.

However, prevailing OCL methodologies typically presuppose balanced class distributions [10], [14], [17], thereby

circumscribing their practical applicability in real-world deployments. In practical scenarios, data distributions frequently exhibit long-tailed characteristics, manifesting as pronounced class imbalances where certain classes are represented by substantially more instances than others—exemplified in domains such as animal species classification and medical image analysis. This work specifically addresses the problem of *Long-Tailed Online Continual Learning* (LT-OCL), a particularly demanding paradigm wherein inherently imbalanced data with long-tailed distributions arrives sequentially. Although prior investigations have identified the presence of class imbalance within continual learning frameworks [6], [9], the preponderance of existing literature has concentrated on relatively modest imbalances or offline learning paradigms [11]. Class imbalance refers to the phenomenon wherein certain classes are substantially overrepresented relative to others within the training corpus. However, long-tailed distributions present considerably greater complexity. In such distributions, a small cohort of classes exhibits exceedingly high sample density, while the majority of remaining classes are characterized by severely limited representations. A critical distinction between offline and online continual learning lies in their respective data access mechanisms: offline learning permits multiple epochs over the complete dataset, enabling a comprehensive understanding of overall class distributions, whereas online scenarios restrict access to sequential task-specific batches, resulting in limited observability of subsequent task distributions. Consequently, severe class imbalance substantially exacerbates catastrophic forgetting phenomena and significantly degrades model performance within OCL settings.

In this paper, we investigate LT-OCL for image classification, specifically in class-incremental learning, where task identity is unknown during both training and testing. We propose a novel dual-stage learning strategy that integrates Contrastive Learning and Equalization Loss to recalibrate the feature space and classifier weights. By augmenting rehearsal sample features, we maximize the utility of rehearsal data to mitigate catastrophic forgetting. Our framework consistently outperforms state-of-the-art OCL methods in long-tailed settings, demonstrating superior performance across diverse configurations.

*Corresponding author: trangptq@vnu.edu.vn

II. RELATED WORK

A. Long-Tailed Classification

Long-tailed classification has been extensively studied in the offline learning setting. *Decoupled Training*, proposed by Kang et al. [8], demonstrates that decoupling feature learning and classifier training substantially improves generalization on long-tailed datasets. This approach identifies that the classification component is the primary bottleneck when training on imbalanced distributions. Similarly, *Balanced Softmax* [18] mitigates the distributional discrepancy between training and testing by re-weighting classifier logits according to inverse class frequencies. However, both methods inherently assume access to the complete data distribution during training, a prerequisite that is fundamentally incompatible with the OCL paradigm. In online continual learning scenarios, data arrive sequentially over time, making comprehensive distributional characterization infeasible and rendering these offline long-tailed methods inapplicable.

B. Contrastive Learning

Contrastive learning represents a systematic methodology for obtaining robust feature representations by simultaneously maximizing agreement among semantically related samples while minimizing agreement among unrelated samples [12]. By leveraging this dual-objective optimization framework, models can effectively uncover and capitalize upon the latent structural and semantic characteristics inherent in the underlying data distribution, thereby enabling the derivation of feature representations with considerable applicability to numerous downstream tasks [2], [16]. Under the supervised contrastive learning paradigm, intra-class samples serve as positive example pairs, while inter-class samples constitute negative example pairs.

C. Rehearsal Feature Augmentation

Rehearsal-based methods [9], [10], [14] try to prevent catastrophic forgetting by storing a small replay memory of previously learned tasks. However, recent studies show that Rehearsal-based methods are prone to overfitting on rehearsal samples, resulting in poor generalization on previous tasks, meaning low performance on actual new data from previous tasks. Since the generalization error is bounded by the margin on the training dataset, *Rehearsal Feature Augmentation* [20] are supposed to acquire larger margin thus better generalization on rehearsal samples.

III. METHODOLOGY

We investigate long-tailed online continual learning (LT-OCL) for supervised image classification. Data distributions exhibit exponential decay in class sample sizes, parameterized by ρ as the ratio between the least and most frequent classes.

A. Problem Formulation

Consider a sequence of T tasks. At task t , the model receives k^t new classes with training samples $\mathcal{X}_t = \{x_i^t, y_i^t\}_{i=1}^{n^t}$, where n_j^t denotes the number of samples for class j , and $\sum_{j=1}^{k^t} n_j^t = n^t$. Training data are non-i.i.d. samples drawn from a task-specific distribution D_t , which shifts to D_{t+1} between consecutive tasks. The model can only access a single batch of data per task due to sequential data flow, rendering the model unable to observe future distributions. Let $k^{1:t}$ represent the cumulative number of classes seen up to task t .

B. Proposed Framework

We propose a dual-stage learning framework specifically designed to address the pronounced class imbalance inherent in long-tailed data streams. The architectural overview is illustrated in Figure 1.

Stage 1: Representation learning

Drawing upon prior method [15], this component incorporates principles from SimCLR [2], which leverage both positive and negative samples to enhance representation learning. In long-tailed distributions, wherein numerous classes exhibit severe sample scarcity, the abundance of negative samples from well-represented classes facilitates the acquisition of more discriminative feature representations for underrepresented classes. Furthermore, SimCLR employs aggressive data augmentation strategies to generate contrastive positive pairs. This methodological approach effectively mitigates the effects of class imbalance by ensuring that the model acquires robust and generalizable feature representations across all classes, independent of their frequency within the dataset.

The contrastive learning pipeline comprises three principal components:

- **Data Augmentation:** A transformation function is applied to each input sample to generate augmented variants:

$$\hat{x} = \text{Aug}(x). \quad (1)$$

- **Encoder Network:** The encoder $\text{Encoder}(\cdot)$ extracts feature representations from image samples, projecting them into a high-dimensional embedding space:

$$e = \text{Encoder}(x) \in \mathbb{R}^{D_N}, \quad (2)$$

where the resulting embeddings are normalized onto the unit hypersphere.

- **Projection Head:** The projection network $\text{Projection}(\cdot)$ further transforms encoder embeddings into a lower-dimensional projected space:

$$v = \text{Projection}(e) \in \mathbb{R}^{D_P}, \quad (3)$$

followed by L2 normalization.

The contrastive loss is defined as follows:

$$L_{\text{contrastive}}(i) = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \left(\frac{\exp(v_i \cdot v_p / \tau)}{\sum_{k \in A(i)} \exp(v_i \cdot v_k / \tau)} \right) \quad (4)$$

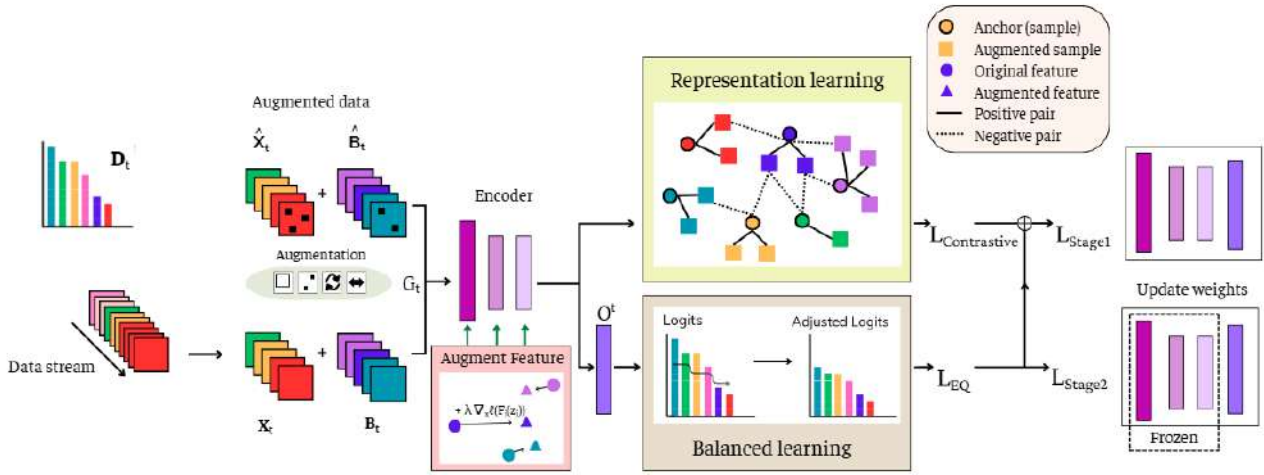


Fig. 1. Overall architecture of ELLA. At task t , current batch ($\mathcal{X}_t$) and buffer samples (B_t) are augmented and merged into G_t . Stage 1: contrastive learning on pairs ($\mathcal{X}_t, B_t$). Stage 2: decoupled classification with weight equalization loss (L_{EQ}) on layer (O^t) for balanced classification.

Let $G_t = \mathcal{X}_t \cup \text{Aug}(\mathcal{X}_t) \cup B_t \cup \text{Aug}(B_t)$ denote the combined dataset comprising the current batch $\mathcal{X}_t$, its augmented counterpart $\text{Aug}(\mathcal{X}_t)$, the replay buffer B_t , and augmented buffer samples $\text{Aug}(B_t)$. For each anchor sample $i \in G_t$, we define:

- $P(i)$: the set of positive samples in G_t sharing the same class label as sample i , excluding i itself, i.e., $P(i) = \{k \in G_t \mid k \neq i, y_k = y_i\}$.
- $A(i)$: the set of all other samples in G_t excluding i itself, i.e., $A(i) = \{k \in G_t \mid k \neq i\}$.

The parameter $\tau \in \mathbb{R}^+$ serves as a temperature factor to modulate class separation, and the symbol “ $\cdot$ ” denotes the dot product operation between two feature vectors.

While Contrastive Loss is effective in learning discriminative features, it primarily focuses on generating homogeneous representations for similar samples and separating dissimilar ones, largely based on information extracted by the encoder. This loss, in its original nature, does not incorporate information from the classifier.

Cross-entropy loss is susceptible to producing biased gradient updates when applied to highly imbalanced class distributions. To counteract gradient bias, we incorporate an Equalization Loss (L_{EQ}), defined in Equation 7.

However, in LT-OCL settings, the complete training distribution is unavailable a priori. To adapt L_{EQ} to the online continual learning paradigm, we introduce a task-specific probability distribution vector $P(k^t)$, which is dynamically updated upon encountering each data stream within task t , as specified in Equation 6. The equalization loss dynamically modulates the classifier’s output logits based on the observed class distribution within the current batch, thereby mitigating the gradient bias induced by severe class imbalance.

The combined training objective is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{EQ}, \quad (5)$$

where λ is a hyperparameter balancing the two loss components.

Let $O^t(I_x)$ denote the classifier logits for input image I_x . For each training batch, we compute a task-specific probability distribution $P(k^t)$ that captures the class distribution observed across the current batch and replay buffer:

$$P(k^t) = \left[\frac{n_1}{\sum_{i=1}^{k^{1:t}} n_i}, \frac{n_2}{\sum_{i=1}^{k^{1:t}} n_i}, \dots, \frac{n_{k^{1:t}}}{\sum_{i=1}^{k^{1:t}} n_i} \right], \quad (6)$$

where n_h is the sample count for class $h \in \{1, 2, \dots, k^{1:t}\}$ aggregated from the current training batch and exemplars in the memory buffer.

The Equalization Loss (L_{EQ}) is then defined as:

$$L_{EQ}(O^t(I_x)) = \sum_{i=1}^{k^{1:t}} -l_{y(i)} \sigma([\log(P(k^t)) + O^t(I_x)]) \quad (7)$$

where I_y is the corresponding label for input I_x and σ represents the softmax function. By adding $\log(P(k^t))$ to the logits $O^t(I_x)$, this loss function effectively reduces the gradient contribution from majority classes.

The overall loss for Stage 1 is the combination of these two components:

$$L_{\text{stage1}}(G_t) = L_{\text{contrastive}}(G_t) + L_{EQ}(O^t(G_t)) \quad (8)$$

Stage 2: Balanced classifier learning

The primary objective of our second stage is to create a balanced training environment for the classifier. This is achieved by decoupling the learning of effective feature representations from the actual classification task itself. By separating these concerns, the second stage can specifically address the classification task with debiasing mechanisms, ensuring that the final decision boundaries are fair and balanced across classes.

Loss function for stage 2:

$$L_{\text{stage2}}(G_t) = L_{\text{EQ}}(O^t(G_t)) \quad (9)$$

We propose augmenting the representations of rehearsal samples at each encoder layer to generate enriched feature variants for training. By pushing rehearsal samples away from the decision boundary, these augmented features provide robustness against boundary shrinkage during incremental learning, thereby enlarging the classification margin and improving generalization on previously learned tasks (see Figure 2).

Augmentation Strategy. At each training batch comprising samples from both the current task and the rehearsal buffer, we augment rehearsal samples using gradient-based perturbations. Following [20], we compute gradients with respect to each layer’s input to determine the augmentation direction. For each rehearsal sample x_i in the batch, we randomly select one layer’s input (immediate feature representation) and apply the augmentation as follows:

$$\hat{z}_i = z_i + \lambda \nabla_x \ell(F_i(z_i)) \quad (10)$$

where ℓ is uniformly sampled from the set of encoder layers $\{1, 2, \dots, L\}$, $z_i = f_0 \circ f_1 \circ \dots \circ f_\ell(x_i)$ represents the intermediate feature representation obtained by applying successive transformations from layer 0 to layer ℓ to sample x_i , and λ controls the magnitude of the feature perturbation.

The augmentation magnitude λ is defined as the product of a layer-specific scaling factor and the L2 norm of the intermediate feature representation:

$$\lambda = \hat{\beta} \|z_{i,l}\|_2, \quad (11)$$

where $\hat{\beta} \sim \mathcal{U}(0, \beta)$ is a randomly sampled scaling coefficient and β is a hyperparameter controlling the maximum augmentation strength. This formulation serves two purposes: first, normalizing by the L2 norm of $z_{i,l}$ adaptively scales perturbations relative to feature magnitude, thereby ensuring consistent augmentation across layers with heterogeneous scales; second, randomly sampling $\hat{\beta}$ from a uniform distribution introduces stochastic variability in augmentation intensity, creating samples of varying difficulty levels—a property particularly beneficial during the late stages of model convergence. Equation 10 implements a gradient ascent step on the selected layer l , augmenting the intermediate feature representation $z_{i,l}$ by the magnitude λ .

IV. EXPERIMENTS AND RESULTS

A. Datasets

We evaluate our method on two datasets:

- **CIFAR100-LT:** Long-tailed version of Split CIFAR-100 with imbalance factor $\rho = 0.01$ [11]. Training: 10K images (5–500 per class). Test: balanced with 100 images per class.
- **VFN-LT:** Real-world food distribution dataset [7]. Training: 2.5K images across 74 classes (1–288 per class). Test: balanced with 25 images per class.

B. Implementation Details

Model Architecture and Training Setup. We employ ResNet18 as the encoder with a fully connected projection layer mapping to a 128-dimensional latent space [3]. All models are trained from scratch using a fixed random seed (1993) to ensure reproducibility. Input image dimensions are 32×32 for CIFAR-100-LT and 224×224 for VFN-LT.

Dataset Configurations. For CIFAR-100-LT, we evaluate two task configurations: (i) 20 tasks with 5 classes per task, and (ii) 10 tasks with 10 classes per task. For VFN-LT, we employ: (i) 15 tasks (4 classes in the initial task, 5 classes in subsequent tasks), and (ii) 7 tasks (14 classes initially, 10 classes per subsequent task). The contrastive learning temperature parameter is set to $\tau = 0.09$.

Experimental Setup. Each experiment is executed 40 times on a single NVIDIA T4 GPU to ensure statistical robustness. **For evaluation,** We use Average Anytime Accuracy (AAA) [1], [17], defined as $\text{AAA}_T = \frac{1}{T} \sum_{i=1}^T \frac{1}{i} \sum_{j=1}^i a_{i,j}$, where $a_{i,j}$ is the model’s performance on task j ’s test set after training up to task i , to evaluate the model’s ability to be deployed at any point during the learning process. **Baseline Methods and Comparison.** For comparative evaluation, we implement existing OCL methods from publicly available repositories [5], [9], [11], [13], [19]. We adapt the offline long-tailed learning method [11] to the online setting by performing a single training epoch in a task-agnostic manner for fair comparison. Additionally, we include DELTA [15], the state-of-the-art method in the LT-OCL setting, for comparative analysis.

C. Results

Tables I and II demonstrate that existing OCL methods achieve comparatively low accuracy on long-tailed datasets. Among these methods, SCR and OnPRO perform best through embedding clustering with NCM classifiers and discriminative feature learning via contrastive loss, respectively. However, performance degrades inconsistently with increased buffer size (0.5K–2K), indicating instability.

Our method, ELLA, surpasses DELTA, the current state-of-the-art in LT-OCL. The dual-stage training framework combined with feature augmentation substantially improves model performance and effectively mitigates catastrophic forgetting across tasks (see Figure ??).

D. Hyperparameters

The feature augmentation process requires a hyperparameter β to control the scale of the augmentations. Our findings indicate that $\beta=0.1$ is optimal (see Figure 3).

E. Computational Cost and GPU Memory

Figure 4 reveals that ELLA incurs increased GPU memory consumption and training time compared to DELTA across all evaluated configurations. The rehearsal feature augmentation mechanism necessitates an additional forward pass and backward pass per training iteration to compute gradients for generating layer-wise perturbations. Furthermore, storing

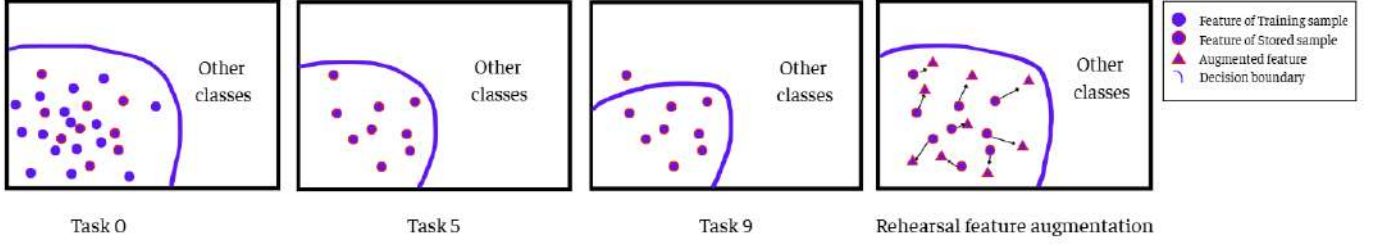


Fig. 2. Decision boundary shrinkage during incremental learning. Task 0 shows the original decision boundary (blue line) for a class. As learning progresses to Tasks 5 and 9, the boundary shrinks around stored samples. Our method generates augmented features (final panel) that push outward against the shrinking boundary to improve generalization on previous tasks.

TABLE I
AVERAGE ACCURACY ON CIFAR100-LT

Methods	CIFAR100-LT					
	20 tasks			10 tasks		
	M = 0.5K	1K	2K	0.5K	1K	2K
DELTA	17.87 $\pm$ 0.30	19.33 $\pm$ 0.19	20.84 $\pm$ 0.22	17.33 $\pm$ 0.41	18.54 $\pm$ 0.21	19.98 $\pm$ 0.40
SCR[CVPRW '21]	18.55 $\pm$ 0.56	18.73 $\pm$ 0.30	19.65 $\pm$ 0.57	17.28 $\pm$ 0.40	17.80 $\pm$ 0.38	18.56 $\pm$ 0.66
OnPRO[ICCV '23]	14.02 $\pm$ 0.44	16.28 $\pm$ 0.81	18.01 $\pm$ 0.22	16.53 $\pm$ 0.55	16.92 $\pm$ 0.08	18.85 $\pm$ 0.32
ASER[AAAI '21]	8.86 $\pm$ 0.30	7.86 $\pm$ 0.61	8.18 $\pm$ 0.31	12.68 $\pm$ 0.70	13.76 $\pm$ 0.01	15.90 $\pm$ 0.91
PRS[ECCV '20]	7.61 $\pm$ 0.09	7.54 $\pm$ 0.21	7.03 $\pm$ 0.13	7.34 $\pm$ 0.92	8.95 $\pm$ 0.33	9.01 $\pm$ 0.39
CBRS[ICML '20]	8.51 $\pm$ 0.19	8.66 $\pm$ 0.61	8.91 $\pm$ 0.33	9.50 $\pm$ 0.48	7.22 $\pm$ 0.43	7.31 $\pm$ 0.08
GSS[NeurIPS '19]	5.16 $\pm$ 0.10	5.22 $\pm$ 0.22	5.09 $\pm$ 0.21	8.97 $\pm$ 0.65	10.12 $\pm$ 0.02	9.96 $\pm$ 0.47
LT-CIL (offline)	3.01 $\pm$ 0.77	2.67 $\pm$ 0.04	2.43 $\pm$ 0.02	1.76 $\pm$ 0.11	2.36 $\pm$ 0.25	3.76 $\pm$ 0.22
ELLA(Ours)	19.80 $\pm$ 0.21	21.12 $\pm$ 0.17	23.19 $\pm$ 0.35	19.71 $\pm$ 0.22	21.18 $\pm$ 0.26	22.46 $\pm$ 0.13

TABLE II
AVERAGE ACCURACY ON VFN-LT

	VFN-LT					
	15 tasks			7 tasks		
	M = 0.5K	1K	2K	0.5K	1K	2K
DELTA	14.90 $\pm$ 0.38	14.98 $\pm$ 0.29	15.07 $\pm$ 0.22	12.18 $\pm$ 0.18	12.69 $\pm$ 0.19	12.86 $\pm$ 0.30
SCR[CVPRW '21]	13.11 $\pm$ 0.57	13.68 $\pm$ 0.66	14.23 $\pm$ 0.33	10.21 $\pm$ 0.79	10.82 $\pm$ 0.46	11.55 $\pm$ 0.47
OnPRO[ICCV '23]	11.93 $\pm$ 0.04	12.77 $\pm$ 0.07	13.50 $\pm$ 0.05	8.02 $\pm$ 0.60	9.38 $\pm$ 0.21	11.84 $\pm$ 0.49
ASER[AAAI '21]	6.85 $\pm$ 0.34	7.61 $\pm$ 0.38	7.22 $\pm$ 0.36	7.46 $\pm$ 1.18	7.52 $\pm$ 1.09	6.35 $\pm$ 0.19
PRS[ECCV '20]	7.17 $\pm$ 0.83	8.72 $\pm$ 0.15	8.39 $\pm$ 0.19	7.85 $\pm$ 0.50	8.66 $\pm$ 0.22	9.21 $\pm$ 0.30
CBRS[ICML '20]	8.12 $\pm$ 0.94	8.35 $\pm$ 0.33	8.18 $\pm$ 0.44	7.52 $\pm$ 0.11	7.64 $\pm$ 0.08	7.92 $\pm$ 0.34
GSS[NeurIPS '19]	5.86 $\pm$ 0.30	6.01 $\pm$ 0.91	5.86 $\pm$ 0.06	5.92 $\pm$ 0.54	4.30 $\pm$ 0.22	4.66 $\pm$ 0.60
LT-CIL (offline)	1.82 $\pm$ 0.45	2.02 $\pm$ 0.44	2.38 $\pm$ 0.08	3.08 $\pm$ 0.71	2.92 $\pm$ 0.04	1.99 $\pm$ 0.31
ELLA(Ours)	15.75 $\pm$ 0.33	15.98 $\pm$ 0.32	16.39 $\pm$ 0.23	13.03 $\pm$ 0.27	14.09 $\pm$ 0.28	14.11 $\pm$ 0.21

both perturbation vectors and intermediate feature representations across multiple encoder layers substantially amplifies memory requirements. However, this computational overhead represents a justifiable trade-off. While the method increases training cost, the resulting improvements in continual learning accuracy substantially outweigh these expenses. Importantly, the augmentation process only occurs during training; the final model's inference speed remains unaffected, making the approach practical for deployment scenarios.

V. CONCLUSION

This work addresses the challenging problem of online continual learning under long-tailed class distributions. Our proposed framework strategically leverages contrastive learning, which demonstrates exceptional efficacy in acquiring robust and discriminative feature representations, particularly when confronted with pronounced class imbalance. The incorporation of an equalization loss mechanism directly within the representation learning phase provides essential feedback signals from the classifier to the feature encoder. This methodological integration, coupled with the deliberate decoupling of feature learning and classification training stages, enables the

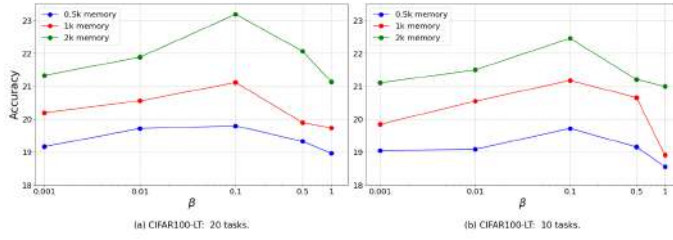


Fig. 3. Impact of β on average accuracy for CIFAR100-LT (20 and 10 Tasks, varying memory sizes).

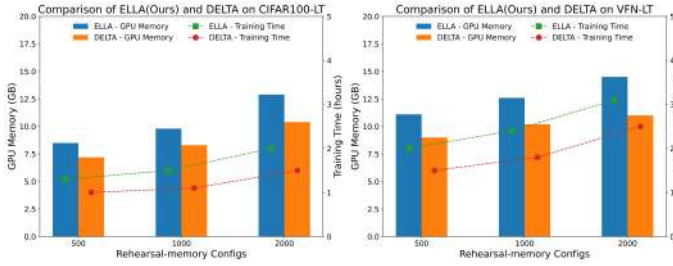


Fig. 4. GPU memory and training time of ELLA(Ours) compared to DELTA on CIFAR100-LT and VFN-LT datasets, each experiment is run 40 times.

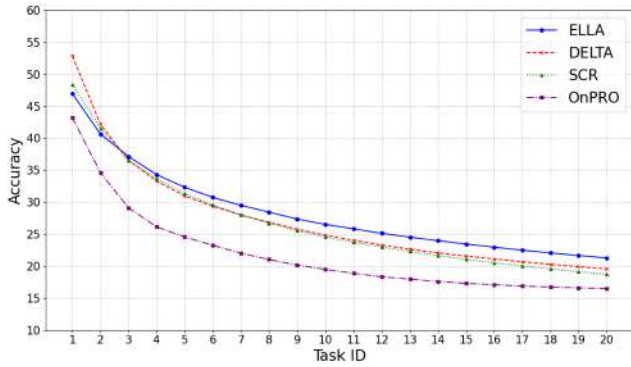


Fig. 5. Average accuracy across tasks with CIFAR100-LT, 1k memory, 20 tasks.

model to effectively mitigate gradient bias induced by severe class imbalance. Additionally, the deployment of rehearsal sample augmentation mechanisms amplifies the informational content extracted from exemplar samples, thereby substantially attenuating catastrophic forgetting on previously learned tasks and consequently enhancing overall model performance across the continual learning trajectory.

REFERENCES

- [1] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. 2021.
- [2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020.
- [4] Zhiyuan Chen and Bing Liu. *Lifelong machine learning*. Morgan & Claypool Publishers, 2018.
- [5] Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. pages 1952–1961, 2020.
- [6] Tyler L Hayes and Christopher Kanan. Online continual learning for embedded devices. *arXiv preprint arXiv:2203.10681*, 2022.
- [7] Jiangpeng He, Luotao Lin, Heather Eicher-Miller, and Fengqing Zhu. Long-tailed food classification. 2022.
- [8] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *CoRR*, abs/1910.09217, 2019.
- [9] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 411–428. Springer, 2020.
- [10] Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24246–24255, 2023.
- [11] Xiaolei Liu, Yu-Song Hu, Xu-Sheng Cao, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Long-tailed class incremental learning. In *European Conference on Computer Vision*, pages 495–512. Springer, 2022.
- [12] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.
- [13] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.
- [14] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3589–3599, 2021.
- [15] Siddeshwar Raghavan, Jiangpeng He, and Fengqing Zhu. Delta: Decoupling long-tailed online continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4054–4064, June 2024.
- [16] Nishant Rai, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Cocon: Cooperative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3384–3393, 2021.
- [17] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [18] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.
- [19] Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, and Hongming Shan. Online prototype learning for online continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18764–18774, 2023.
- [20] Bowen Zheng, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Multi-layer rehearsal feature augmentation for class-incremental learning. In *Forty-first International Conference on Machine Learning*, 2024.

A Reliability-Enhanced Approach for Neighborhood-Aware Possibilistic Semi-Supervised Fuzzy Clustering

Hiep Huynh*, Hung C. Nguyen*, and Quang-Thinh Bui^{†‡}

*Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam
hiephuynh.hut@gmail.com, hungcn.ds@gmail.com

[†]Office of Scientific Research, Technology Management and International Cooperation,
Tien Giang University, Dong Thap, Vietnam
buiquangthinh@tgu.edu.vn
[‡]Corresponding Author

Abstract—Fuzzy clustering is very important for data mining and deep learning, with applications in medical diagnosis, image processing, and text analysis. Traditional fuzzy clustering algorithms, such as Fuzzy C -Means and Possibilistic C -Means, face challenges when dealing with noisy, imbalanced, or large-scale data due to uncertain membership values at cluster boundaries. Recently, the RPSSFC-NAFW algorithm, a robust possibilistic semi-supervised fuzzy clustering approach with neighborhood-aware feature weighting, has improved performance through feature weighting, neighborhood information, and kernel metrics, yet still suffers from uncertainty, cluster overlap, and imbalance. This paper proposes RePSSFC-NAFW, an extended version of RPSSFC-NAFW that integrates a reliability control mechanism via five regularization terms: entropy, margin, balance, intra-cluster and membership regularization. These components help reduce uncertainty, enhance cluster separability, and ensure balanced distributions. Experiments conducted on 15 UCI benchmark datasets demonstrate that RePSSFC-NAFW consistently outperforms its state-of-the-art predecessors. The average ACC increases from 0.8542 to 0.9038, with significant gains on Vowel (+14.5%) and SpectfHeart (+13.2%). Improvements are also observed in NMI and F1-score, demonstrating the robustness and effectiveness of the proposed technique in complicated, noisy, and imbalanced data scenarios.

Index Terms—Fuzzy Clustering, Reliability Control, Feature Weighting, Possibilistic C -Means, Semi-Supervised

I. INTRODUCTION

Fuzzy clustering is one of the essential techniques in data mining and deep learning, allowing the detection of patterns and hidden structures in large-scale datasets [1], [2]. Classical algorithms such as Fuzzy C -Means and Possibilistic C -Means have been widely applied in fields such as medical diagnosis [3], image processing [4], and text analysis. However, they often encounter difficulties when processing noisy data, class imbalance, or large-scale data, mainly due to membership values at cluster boundaries reducing data robustness and cluster separability [5], [6].

Algorithms such as FCM [7] and PCM [8] have been designed to manage uncertainty in cluster assignments, particularly at cluster boundaries where data points may exhibit nearly equal membership values across multiple groups [9], [10]. This ambiguity can reduce model confidence, thereby

diminishing robustness against noise and outliers. Additional models like Fuzzy C -Means with Weighted Membership (FCM-WM) [11], Weight Possibilistic Fuzzy C -Means (WPFCM) [12], and Semi-Supervised Possibilistic Fuzzy C -Means (SSPFCM) [13] further address these issues by incorporating weighted memberships and labeled data to improve clustering accuracy on noisy datasets [14], [15]. Semi-supervised approaches, such as Feature Weighted Safe-Semi-Supervised Kernel-Based Intuitionistic Fuzzy C -Means (FW-S3KIFCM) [16], provide robust alternatives to handle uncertainty in unlabeled data, noise, and outliers through kernel metrics and feature weights [17]. In response, recent advancements in semi-supervised clustering, the RPSSFC-NAFW algorithm [18] was introduced, which has integrated mechanisms like feature weights, neighborhood information [19], and kernel metrics to improve the handling of complex data structures [20]. Despite such advancements, current methods still exhibit certain limitations as previously mentioned. Specifically, membership values may not favor any particular cluster, resulting in ambiguous cluster assignments, especially at the cluster boundaries.

The primary contributions of this paper are threefold: (1) Introducing RePSSFC-NAFW, an enhanced version of RPSSFC-NAFW incorporating a novel reliability control mechanism with three regularization terms entropy, margin, and balance to mitigate uncertainty, improve cluster separability, and promote balanced sample distribution. (2) Developing a generalized objective function that seamlessly integrates these reliability components, enabling adaptable clustering on noisy, imbalanced, and complex datasets. (3) Demonstrating, through extensive experiments on benchmark datasets, that the proposed method surpasses the original in metrics such as ACC, NMI, and F1-score, while exhibiting stable performance in challenging scenarios.

The remainder of this paper is organized as follows: Section II provides an overview of related works and the essential theoretical background. Section III describes the proposed methodology and details its principal innovations. Section IV presents experimental results and an in-depth analysis of the

algorithm's effectiveness on various standard datasets. Finally, Section V discusses the findings, draws conclusions, and suggests directions for future research.

II. BACKGROUND

This section presents Semi-supervised possibilistic Fuzzy C-Means (SSPFCM) and Neighborhood-Aware Feature Weighting (NAFW), showing how labels, typicality, and neighborhood-weighted features improve robust clustering.

A. Semi-supervised possibilistic Fuzzy C-Means

SSPFCM is an extension of PCM that integrates semi-supervised learning. The algorithm leverages a small amount of labeled data to guide the clustering process, aiming to improve noise robustness and address key limitations of prior methods, such as cluster bias due to the original membership constraints u_{nk} and the inability to utilize available label information fully. The objective function of SSPFCM is explained in detail by the following formula:

$$J(U, T, H) = \sum_{k=1}^K \sum_{n=1}^N (a \cdot u_{nk}^n + b \cdot t_{nk}^n \cdot \|x_n - c_k\|^2) + \sum_{k=1}^K \sum_{n=1}^N \delta_k (1 - t_{nk})^\eta. \quad (1)$$

In the above formula, a and b are the coefficients that control the importance of the membership and possibilistic terms, respectively, where $a > 0$, $b > 0$, $m > 0$, and $\eta > 1$. In addition, δ_k is a constant whose value is greater than zero and is defined by the user. Three main components are used to guide the clustering process, including: the matrix U , which represents the fuzzy membership level of each data sample to the clusters; the matrix T , which indicates the typicality of the sample to each cluster and helps reduce the influence of noise and outliers; and the set $H = h_1, h_2, \dots, h_K$, which includes the cluster center vectors representing each cluster in the feature space.

B. Neighborhood-Aware Feature Weighting

NAFW [21] is a weighting mechanism that considers the influence of neighboring data points during clustering. It combines feature weighting and neighborhood awareness:

$$NAFW = \text{Neighborhood Aware} + \text{Feature Weighting}.$$

This methodology enables adaptive cluster assignment by leveraging local data structure and implementing cluster-dependent feature importance, thereby improving the algorithm's ability to capture complex data patterns. Drawing from these core principles, the mathematical formulation of NAFW is given as follows:

$$w_{km} = \frac{1}{\left[\frac{\sum_{n=1}^N (au_{nk}^2 + bt_{nk}^\eta + a_2(u_{nk} - b_n f_{nk})^2) (1 - k(x_{nm}, c_{nm}))}{\sum_{n=1}^N (au_{nk}^2 + bt_{nk}^\eta + a_2(u_{nk} - b_n f_{nk})^2) (1 - k(x_{ns}, c_{ks}))} \right]^{\frac{1}{q-1}}} \quad (2)$$

The above equation encapsulates the synergy between neighborhood awareness and feature weighting, optimizing the clustering of data.

III. METHODOLOGY

This section elaborates on the proposed RePSSFC-NAFW framework, which is developed upon the RPSSFC-NAFW algorithm by incorporating a reliability control mechanism and refining the overall clustering architecture.

A. Reliability Control

The RPSSFC-NAFW algorithm has effectively addressed numerous limitations of previous SSPFCM methods by integrating feature weighting, neighborhood awareness, and kernel metrics to enhance its capability for recognizing complex data structures. However, experimental evaluations show that this algorithm still has certain weaknesses, including membership values that are still soft at cluster boundaries leading to high uncertainty, the possibility of unbalanced clustering on non-uniform datasets, and insufficient control over inter-cluster distances, which can lead to cluster overlap.

To overcome the aforementioned limitations, this study proposes an improved version named **RePSSFC-NAFW**, developed based on the RPSSFC-NAFW framework by integrating a *reliability control mechanism* into the objective function. This extension introduces five regularization terms: L_{entropy} sharpens the membership matrix by penalizing distributions with similar values, encouraging each sample to have a clearly dominant membership in a single cluster. This mechanism reduces ambiguity at cluster boundaries and enhances clustering reliability, particularly for noisy datasets or overlapping cluster regions.

L_{margin} penalizes samples with a small difference in membership values between the two highest-ranked clusters, forcing decision boundaries to lie in low-density regions, thereby increasing the distance between cluster centers and reducing the risk of overlap.

L_{balance} adjusts the distribution of samples among clusters toward a balanced state, preventing cluster dominance, maintaining complete data coverage, and ensuring fairness across groups, even when the initial data distribution is imbalanced.

$L_{\text{intra-cluster}}$ reduces the dispersion of features within each cluster after applying feature weighting and kernel mapping, ensuring feature homogeneity, improving cluster compactness, and stabilizing cluster centers during optimization.

Finally, L_{regu} regulates both the mass and concentration of membership values, preventing clusters from having too few or too many samples, thereby mitigating overfitting and improving the model's generalization capability on unseen data.

These regularization terms are weighted by the parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 respectively. The modified objective function can be formulated as follows:

$$F_{\text{total}} = F_{\text{RPSSFC-NAFW}} + \lambda_1 L_{\text{entropy}} + \lambda_2 L_{\text{margin}} + \lambda_3 L_{\text{balance}} + \lambda_4 L_{\text{intra-cluster}} + \lambda_5 L_{\text{regu}} \quad (3)$$

where $F_{\text{RPSSFC-NAFW}}$ is the original objective function, and F_{total} is the new objective function of the proposed RePSSFC-NAFW method.

$$\begin{aligned}
F(U, C, T, W, S) = & \sum_{n=1}^N \sum_{k=1}^K \sum_{m=1}^M (a u_{nk}^2 + b t_{nk}^2) w_{km}^q \|\varphi(x_{nm}) - \varphi(c_{km})\|^2 \\
& + \frac{\alpha_1}{N_R} \sum_{k=1}^K \sum_{n=1}^N \sum_{m=1}^M u_{nk}^2 w_{km}^q \\
& \times \sum_{r \in N_n} S_{nr}^p (1 - u_{rk})^2 \|\varphi(x_{rm}) - \varphi(c_{km})\|^2 \\
& + \alpha_2 \sum_{n=1}^N \sum_{k=1}^K \sum_{m=1}^M (u_{nk} - b_n f_{nk})^2 w_{km}^q \|\varphi(x_{nm}) - \varphi(c_{km})\|^2 \\
& + \sum_{k=1}^K \delta_k \sum_{n=1}^N (1 - t_{nk})^\eta
\end{aligned} \tag{4}$$

After incorporating the reliability control regularization terms, the complete objective function can be explicitly written as:

$$\begin{aligned}
F(U, C, T, W, S) = & F(U, C, T, W, S) + \lambda_1 \sum_{n=1}^N \sum_{k=1}^K (-u_{nk} \log(u_{nk} + \epsilon)) \\
& + \lambda_2 \sum_{n=1}^N (u_{n,[1]} - u_{n,[2]})^2 \\
& + \lambda_3 \sum_{k=1}^K \left(\frac{1}{N} \sum_{n=1}^N u_{nk} - \frac{1}{K} \right)^2 \\
& + \lambda_4 \sum_{k=1}^K \sum_{m=1}^M \left(\frac{1}{|C_k|} \sum_{i \in C_k} w_{km}^q |\varphi(x_{im}) - \varphi(c_{km})| \right)^2 \\
& + \lambda_5 \sum_{k=1}^K \left(\frac{1}{\sum_{n=1}^N u_{nk} + \epsilon} \sum_{n=1}^N u_{nk}^2 \right)
\end{aligned} \tag{5}$$

where λ_1 , λ_2 , λ_3 , λ_4 , and λ_5 are regularization parameters that control the effects of the entropy, margin, balance, intra-cluster, and regu terms respectively. Specifically, λ_1 sharpens the membership matrix and reduces uncertainty at cluster boundaries, λ_2 increases the separation between cluster centers to prevent cluster overlap, λ_3 promotes balanced cluster sizes and reduces unbalanced clustering, λ_4 controls the feature consistency among clusters, and λ_5 adjusts the membership distribution among clusters. The simultaneous integration of these five regularization components not only helps the model mitigate uncertainty at cluster boundaries and enhance cluster separability, but also maintains balanced data distribution and ensures intra-cluster feature consistency. Consequently, the extended objective function becomes more flexible and stable compared with the original RPSSFC-NAFW, which mainly relies on feature weighting and neighborhood information. While RPSSFC-NAFW already improves clustering robustness, it still suffers from soft memberships, potential cluster overlap, and imbalance. By contrast, RePSSFC-NAFW with the five regularization terms effectively addresses these limitations, thereby delivering clearer decision boundaries, more balanced clusters, and higher reliability of the membership matrix.

B. Proposed Architecture

In this section, we present the overall architecture of the RPSSFC-NAFW algorithm with the addition of the reliability

control component. The main objective of this design is to enhance the stability and effectiveness of the semi-supervised fuzzy clustering approach by incorporating advanced mechanisms. Specifically, the framework integrates feature weighting, neighbor's trade-off weighting, and a kernel metric to optimize the clustering process and data representation. The key novelty lies in the incorporation of the reliability control mechanism, which directly addresses issues such as uncertainty, cluster overlap, and unbalanced clustering. The overall architecture of the proposed method is illustrated in Fig. 1.

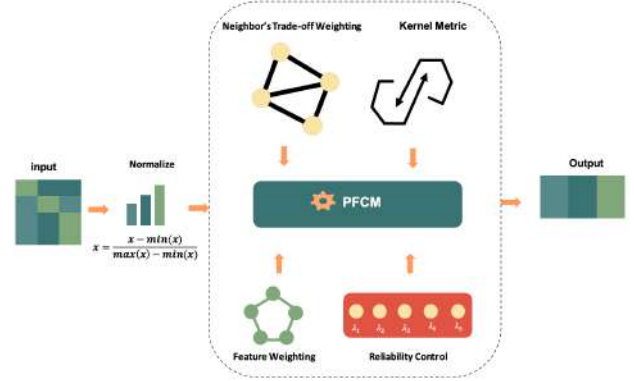


Fig. 1: Overview of the proposed clustering framework highlighting the addition of the reliability control component.

IV. EXPERIMENTS

This section presents the experimental design conducted to evaluate the effectiveness and robustness of the proposed RePSSFC-NAFW method across diverse benchmark datasets and parameter settings.

A. Datasets

To evaluate the effectiveness of the proposed methodology, this study employs 15 benchmark datasets sourced from the UCI Machine Learning Repository. The dataset selection was strategically designed to encompass diverse dimensionality, sample size, and class distribution characteristics, providing a comprehensive evaluation framework that captures the performance challenges inherent in different data configurations. A detailed summary of the selected datasets and their key characteristics is presented in Table I.

B. Configuration

To evaluate the effectiveness of the proposed algorithm, this study conducts experiments with a comprehensive set of parameters optimized through a grid search. The regularization parameters λ_1 , λ_2 , λ_3 , λ_4 and λ_5 are investigated in different value ranges: $\lambda_1 \in \{0, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, $\lambda_2 \in \{0, 0.05, 0.1, 0.2, 0.5\}$, $\lambda_3 \in \{0, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$, $\lambda_4 \in \{0, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$ and $\lambda_5 \in \{0, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$. These ranges were selected to control the regularization level and ensure an appropriate balance among the different components of the objective function.

TABLE I: Statistics of datasets used in the experiment.

Dataset	Samples	Features	Classes	Status
Australian Credit Approval	690	14	2	Balanced
Balance	625	4	3	Imbalanced
Bupa (Liver Disorders)	345	7	2	Balanced
CMC	1473	9	3	Semi-Imbalanced
Dermatology	366	33	6	Imbalanced
Ionosphere	351	34	2	Semi-Imbalanced
Iris	150	4	3	Balanced
Letter Recognition	2291	16	3	Balanced
Pima Indians Diabetes	768	8	2	Semi-Imbalanced
Seed	210	7	3	Balanced
Spectf heart	267	44	2	Imbalanced
Thyroid	215	6	3	Imbalanced
Vowel	871	3	3	Imbalanced
Waveform	5000	21	3	Balanced
Wine	178	13	3	Balanced

The labeled data ratio is configured at 10 %, 20%, and 30% to assess the semi-supervised learning capacity of the algorithm under varying information conditions. The maximum iteration parameter $t_{\max}$ is examined with values $\{100, 150, 200\}$ to ensure convergence, while the weighting parameters $\alpha_1 \in \{2, 4, 6\}$ and $\alpha_2 \in \{0.5, 1, 1.5\}$ are adjusted to optimize the balance between different algorithmic components. The entire data set is partitioned into an 80:20 ratio for training and test sets, ensuring objectivity in the performance evaluation process.

The experimental performance is assessed using three complementary evaluation metrics that capture different aspects of cluster quality.

Accuracy (ACC). Measures the proportion of correctly classified instances, providing a straightforward assessment of overall classification performance:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

F1-Score. Combines precision and recall into a single harmonic mean metric, offering a balanced evaluation that accounts for both false positives and false negatives:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

Normalized Mutual Information (NMI). Quantifies the amount of information shared between the predicted clusters and ground truth labels:

$$NMI(Y, C) = \frac{2 \times I(Y; C)}{H(Y) + H(C)} \quad (8)$$

Here, X represents the predicted clustering result, Y represents the ground truth labels, N is the total number of data points, and the NMI value ranges from 0 to 1, where higher values indicate better clustering structure preservation.

C. Results

Table II presents the performance of the proposed RePSSFC-NAFW algorithm when compared with the best baseline model on each dataset. Three main metrics are used for evaluation, including accuracy (ACC), F1-score (F1), and normalized mutual information (NMI). The columns com_acc , com_f1 , and com_nmi represent the absolute improvement of RePSSFC-NAFW over the best remaining model. Although the proposed RePSSFC-NAFW method achieves superior performance on most datasets, a few cases, such as Iris and SpectfHeart in terms of F1-score, yield slightly lower results compared to the best baseline model. This decline may be due to the intrinsic characteristics of these datasets: Iris already has high separability with very low noise, making the additional regularization components bring limited improvement. Similarly, for SpectfHeart, the class distribution and the structure of the feature space may cause the model's reliability control components to prioritize cluster separability and balance, leading to a slight decrease in F1.

As illustrated in Fig. 2, the RePSSFC-NAFW method demonstrates significant improvements across most datasets in all three metrics: ACC, F1-score, and NMI, compared to RPSSFC-NAFW and traditional approaches. In particular, for challenging datasets such as CMC, Bupa, and SpectfHeart, the increases in ACC and F1-score highlight better adaptability to complex or imbalanced data, whereas methods like FWCW-FCM, FWCW-SSFCM, FW-R-PCM, VK-WFCM, and FWCW-NISSCM often perform well only on easy datasets and degrade on more difficult ones. The inclusion of entropy, margin, and balance components has contributed to higher efficiency, improved control, enhanced cluster separability, and ensured balanced sample distribution.

This illustration Fig. (d) demonstrates the effectiveness of the RePSSFC-NAFW clustering algorithm on the waveform dataset after clustering. The data are displayed with three distinct clusters, color-coded (blue, orange, yellow). The clear separation among the clusters in the two-dimensional PC1-PC2 space indicates successful pattern recognition and confirms the algorithm's effectiveness in waveform data classification.

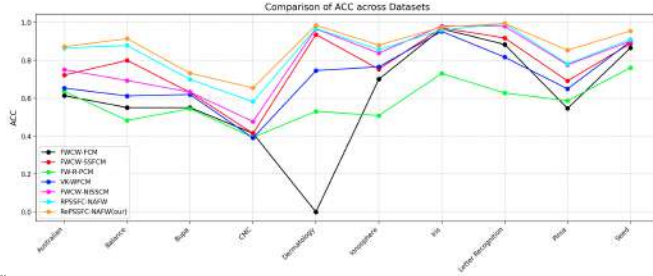
These improvements are especially important in the context of clustering imbalanced datasets, where the F1-score reflects both accuracy and coverage, enabling the algorithm to effectively identify both large and small clusters while maintaining fairness across groups. Furthermore, the increase in NMI indicates that the algorithm not only partitions data effectively but also preserves strong alignment with the underlying data distribution, thereby affirming the stability and reliability of the clustering results across multiple benchmark datasets.

D. Discussion

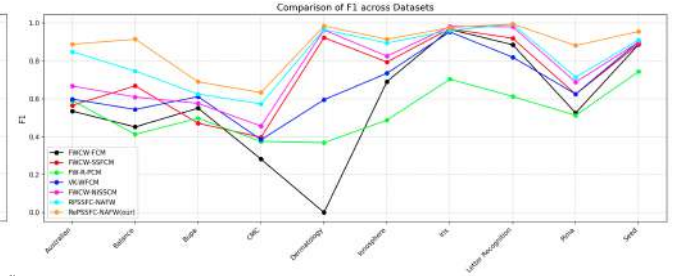
Based on the above analyses, the experimental results show that integrating the reliability control mechanism into the ReSSFC-NAFW algorithm has brought significant improvements in clustering performance on most benchmark datasets with medium to high complexity, while enhancing the values

TABLE II: Performance Comparison: RePSSFC-NAFW vs. Best Baseline Method

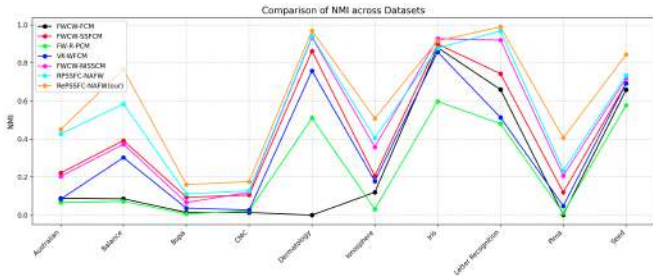
Dataset	ACC			F1			NMI		
	Best_Baseline	Ours	↑	Best_Baseline	Ours	↑	Best_Baseline	Ours	↑
Australian	0.8652	0.8728	0.8%	0.8482	0.8879	4.0%	0.4266	0.4502	2.4%
Balance	0.8787	0.9147	3.6%	0.7466	0.9148	16.8%	0.5839	0.7650	18.1%
Letter	0.9940	0.9949	0.1%	0.9940	0.9949	0.1%	0.9674	0.9890	2.2%
Pima	0.7817	0.8536	7.2%	0.7140	0.8809	16.7%	0.2336	0.4055	17.2%
Spectfheart	0.7932	0.9251	13.2%	0.8644	0.8537	-1.1%	0.1596	0.6040	44.4%
Bupa	0.7000	0.7328	3.3%	0.6250	0.6895	6.5%	0.1112	0.1604	4.9%
CMC	0.5820	0.6548	7.3%	0.5729	0.6332	6.0%	0.1275	0.1760	4.8%
Ionosphere	0.8564	0.8803	2.4%	0.8946	0.9140	1.9%	0.4059	0.5078	10.2%
Iris	0.9820	0.9760	-0.6%	0.9819	0.9760	-0.6%	0.9272	0.9154	-1.2%
Dermatology	0.9688	0.9857	1.7%	0.9656	0.9842	1.9%	0.9392	0.9842	4.5%
Seed	0.9119	0.9557	4.4%	0.9104	0.9555	4.5%	0.7337	0.8436	11.0%
Thyroid	0.9300	0.9566	2.7%	0.8995	0.9235	2.4%	0.6925	0.7719	7.9%
Vowel	0.8412	0.9862	14.5%	0.8265	0.9769	15.0%	0.6973	0.9226	22.5%
Waveform	0.8400	0.9840	14.4%	0.8343	0.9840	15.0%	0.4930	0.9166	42.4%
Wine	0.9640	0.9702	0.6%	0.9462	0.9691	2.3%	0.8201	0.8834	6.3%



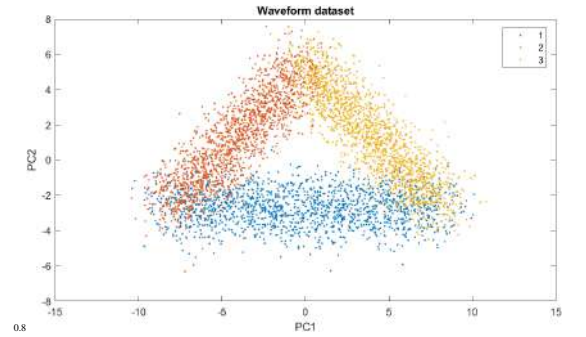
(a) Accuracy



(b) F1-score



(c) NMI



(d) Waveform clustering

Fig. 2: Performance comparison (Accuracy, F1-score, NMI) and Waveform clustering results of RePSSFC-NAFW against Baseline Models.

of all evaluation metrics. Specifically, the proposed method achieves ACC, NMI, and F1-score values that are superior to those of the RPSSFC-NAFW algorithm, particularly showing a clear advantage on imbalanced datasets. To illustrate, on the SpectfHeart dataset, ACC improves from 0.7932 to 0.9251, and NMI increases from 0.1596 to 0.6040. Similarly, on the CMC dataset, which is known for its class imbalance, ACC increases from 0.5820 to 0.6548, and F1-score rises

from 0.5729 to 0.6332. These improvements highlight the effectiveness of the added regularization terms in enhancing clustering reliability.

The above statements indicate that the integrated components in ReSSFC-NAFW effectively reduce instability, limit cluster overlap, and improve class balance. Overall, these improvements help the proposed algorithm achieve superior accuracy and stability, while demonstrating strong general-

ization capability on both balanced and imbalanced datasets, thereby emphasizing its practical application value in solving complex clustering problems.

V. CONCLUSION

This study proposes RePSSFC-NAFW, an improved version of RPSSFC-NAFW incorporating five regularization terms: entropy, margin, balance, intra-cluster consistency, and membership regularization, aimed at reducing uncertainty, enhancing cluster separation, and promoting balanced data distribution. Results on 15 UCI benchmark datasets demonstrate that the proposed method outperforms the original across all ACC, NMI, and F1-score metrics, with particularly notable improvements on the CMC(7,3%) and SpectfHeart(13,2%) datasets. This work contributes a practical and interpretable clustering framework that improves accuracy, reliability, and stability, while exhibiting strong generalization capability on both balanced and imbalanced datasets.

In future research, we will focus on developing an automatic mechanism for tuning the regularization parameters to minimize dependence on manual configuration and enhance the algorithm's flexibility when applied to different datasets and scenarios. In addition, the proposed method will be integrated into large-scale clustering systems, optimized for distributed computing environments, and extended to efficiently process unstructured data, big data, and real-time streaming data. Furthermore, we will evaluate the effectiveness of the method in application domains beyond healthcare, such as anomaly detection in IoT networks, user behavior analysis, and customer segmentation in e-commerce, in order to demonstrate the versatility and scalability of the proposed solution. Moreover, it would be stronger with explicit comparisons to modern deep clustering methods (e.g., deep fuzzy clustering or hybrid deep learning approaches), which could further highlight the competitiveness of the proposed model. In addition, although parameter sensitivity is briefly discussed, the method still requires manual tuning of multiple λ -parameters, which may limit usability in practice. To address this, population-based optimization techniques such as Particle Swarm Optimization (PSO) will be explored for automatic parameter tuning. Ultimately, integrating these improvements will not only enhance robustness and generalization but also strengthen the method's applicability across diverse real-world domains.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

- [1] P. Bhattacharjee and P. Mitra, "A survey of density based clustering algorithms," *Frontiers of Computer Science*, vol. 15, p. 151308, Sep 2020.
- [2] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artificial Intelligence Review*, vol. 56, pp. 6439–6475, Jul 2023.
- [3] Y. Zheng, Z. Xu, T. Wu, and Z. Yi, "A systematic survey of fuzzy deep learning for uncertain medical data," *Artificial Intelligence Review*, vol. 57, p. 230, Aug 2024.
- [4] M. Hu, Y. Zhong, S. Xie, H. Lv, and Z. Lv, "Fuzzy system based medical image processing for brain disease prediction," *Frontiers in Neuroscience*, vol. Volume 15 - 2021, 2021.
- [5] S. Zhou, H. Xu, Z. Zheng, J. Chen, Z. Li, J. Bu, J. Wu, X. Wang, W. Zhu, and M. Ester, "A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions," *ACM Comput. Surv.*, vol. 57, Nov. 2024.
- [6] K. Taha, "Semi-supervised and un-supervised clustering: A review and experimental evaluation," *Information Systems*, vol. 114, p. 102178, 2023.
- [7] S. E. Hashemi, F. Gholian-Jouybari, and M. Hajiaghahi-Keshteli, "A fuzzy c-means algorithm for optimizing data clustering," *Expert Systems with Applications*, vol. 227, p. 120377, 2023.
- [8] J. B. M. Benjamin and M.-S. Yang, "Tensor-based possibilistic c-means clustering," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 10, pp. 5939–5950, 2024.
- [9] M. A. B. Siddique, R. B. Arif, M. M. R. Khan, and Z. Ashrafi, "Implementation of fuzzy c-means and possibilistic c-means clustering algorithms, cluster tendency analysis and cluster validation," *CoRR*, vol. abs/1809.08417, 2018.
- [10] Y. Pu, W. Yao, and X. Li, "Em-icfm: Fuzzy c-means clustering algorithm based on edge modification for imbalanced data," *Information Sciences*, vol. 659, p. 120029, 2024.
- [11] B. A. Pimentel, R. de Amorim Silva, and J. C. S. Costa, "Fuzzy c-means clustering algorithms with weighted membership and distance," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 30, no. 04, pp. 567–594, 2022.
- [12] J. Chen, H. Zhang, D. Pi, M. Kantardzic, Q. Yin, and X. Liu, "A weight possibilistic fuzzy c-means clustering algorithm," *Scientific Programming*, vol. 2021, no. 1, p. 9965813, 2021.
- [13] D. S. Mai, L. T. Ngo, L. H. Trinh, and H. Hagrais, "A hybrid interval type-2 semi-supervised possibilistic fuzzy c-means clustering and particle swarm optimization for satellite image analysis," *Information Sciences*, vol. 548, pp. 398–422, 2021.
- [14] N. Samadi, J. Tanha, and M. Jalili, "A weighted semi-supervised possibilistic fuzzy c-means clustering for data stream classification and emerging class detection," *Knowledge-Based Systems*, vol. 309, p. 112831, 2025.
- [15] H. Yu, X. Xu, H. Li, Y. Wu, and B. Lei, "Semi-supervised possibilistic c-means clustering algorithm based on feature weights for imbalanced data," *Knowledge-Based Systems*, vol. 286, p. 111388, 2024.
- [16] S. Khezri, N. Aghazadeh, M. Hashemzadeh, and A. G. Oskouei, "Fw-s3kicfm: Feature weighted safe-semi-supervised kernel-based intuitionistic fuzzy c-means clustering method," *Fuzzy Information and Engineering*, vol. 17, no. 2, pp. 227–260, 2025.
- [17] A. Golzari Oskouei, N. Samadi, S. Khezri, A. Najafi Moghaddam, H. Babaei, K. Hamini, S. Fath Nojavan, A. Bouyer, and B. Arasteh, "Feature-weighted fuzzy clustering methods: An experimental review," *Neurocomputing*, vol. 619, p. 129176, 2025.
- [18] A. N. Moghaddam, N. Aghazadeh, M. Hashemzadeh, and A. G. Oskouei, "A robust possibilistic semi-supervised fuzzy clustering algorithm with neighborhood-aware feature weighting," *International Journal of Machine Learning and Cybernetics*, Aug 2025.
- [19] A. K. Jasim, J. Tanha, and M. A. Balafar, "Neighborhood information based semi-supervised fuzzy c-means employing feature-weight and cluster-weight learning," *Chaos, Solitons & Fractals*, vol. 181, p. 114670, 2024.
- [20] A. Golzari Oskouei, N. Samadi, and J. Tanha, "Feature-weight and cluster-weight learning in fuzzy c-means method for semi-supervised clustering," *Applied Soft Computing*, vol. 161, p. 111712, 2024.
- [21] K. Yuan, D. Miao, H. Zhang, and W. Pedrycz, "An efficient and robust feature selection approach based on zentropy measure and neighborhood-aware model," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2025.

An Enhanced Semi-Supervised Gaussian Fuzzy C-Means Clustering Algorithm for Remote Sensing Image Data Analysis

Le Tuan Anh

¹Graduate Uni. of Sciences and Technology, VAST

²Thai Nguyen University of Information and
Communication Technology, Thai Nguyen, Vietnam
ltanh@ictu.edu.vn

Dinh Sinh Mai*

Le Quy Don Technical University

Hanoi, Vietnam
maidinhsinh@lqdtu.edu.vn

Ha Thi Lien Doanh

SAMCOM Co.Ltd

Hanoi, Vietnam
htld5988@gmail.com

Abstract—Satellite image data has many advantages for mapping and assessing land cover changes, environmental monitoring, and urban planning. However, remote sensing image data often contains noise and outliers. Traditional clustering algorithms often encounter challenges in processing this data. This paper proposes an enhanced semi-supervised Gaussian fuzzy c-Means clustering algorithm for remote sensing image analysis (ESS-GFCM). The proposed algorithm integrates Gaussian membership functions with semi-supervised learning constraints and a boosting mechanism to improve land cover classification performance from remote sensing imagery. The proposed method can use some labeled or known data to aid the clustering process, incorporate Gaussian kernel weights to better model class distributions, and apply boosting updates to stabilize convergence. Experiments performed on Sentinel-2 and Landsat-8 datasets demonstrate that ESS-GFCM outperforms FCM, Gaussian FCM, and Semi-supervised FCM algorithms in terms of classification accuracy, robustness to noise, and computational speed.

Index Terms—Semi-supervised learning, remote sensing image, fuzzy c-means, landcover classification, Gaussian membership function

I. INTRODUCTION

In recent years, remote sensing technology has made great strides as the number of satellites observing the Earth has increased. One of the important applications of satellites is to provide multi-temporal remote sensing data sources [1]. Remote sensing data offers the advantage of wide coverage and can be captured both during the day and at night. However, this data source also has disadvantages, such as large amounts of data that often contain noise and outliers as they are affected by climatic conditions [2]. One of the important applications of remote sensing images is to use in land cover classification, map making, and object recognition, in which pixels in the image are assigned to specific classes (ex, vegetation, water, urban areas) [3].

Clustering methods such as fuzzy clustering (FCM) have been widely used in satellite image clustering due to their ability to model uncertainty and handle overlapping pixels [4]. However, conventional FCM suffers from several drawbacks: it is sensitive to noise and outliers, lacks prioritization between data points in membership function calculations, and is an unsupervised algorithm that cannot utilize labeled data to guide the clustering process.

Land cover classification studies are still limited because most fuzzy clustering algorithms and their variants are unsupervised methods [5]. The clustering process only uses information on the spectral values of the objects. Objects also contain additional information, such as spatial relationship information, shape information, and labeled data, among others. Therefore, spectral information does not fully describe the characteristics of objects, and the clustering results may not achieve the desired accuracy. On the other hand, the goal of the clustering process is to optimize the objective function. Therefore, in addition to using additional information to propose a new and more suitable objective function, incorporating reinforcement learning solutions to select optimal parameters will bring more stability and efficiency to the algorithm.

To enhance the robustness of semi-supervised clustering under uncertain label conditions, several advanced approaches have been introduced. Building upon the concept of safety-aware learning, Gan et al. proposed the Adaptive Safety-Aware Semi-Supervised Fuzzy C-Means algorithm (AS3FCM), which adaptively adjusts the weighting of mislabeled samples to mitigate the adverse effects of annotation errors [7]. Earlier, Santos et al. developed the KNN-based Graph-Based Semi-Supervised Fuzzy C-Means algorithm (K-GBS3FCM), which exploits the K-Nearest Neighbors (KNN) strategy to dynamically capture neighborhood relationships between labeled and unlabeled data [6]. This method introduces a set of regularization and mean-safety parameters that enable flexible control over the influence of labeled samples, thereby optimizing the contribution of reliable supervision while minimizing the impact of incorrect labels.

More recently, Sun et al. proposed a contrastive learning framework based on a mixture of Gaussian mixture distributions [8]. Their model employs a probabilistic prototype predictor to regulate intra-class prototype consistency within pixel-level representations and maps the features into a multivariate Gaussian space, effectively alleviating pseudo-label uncertainty [9]. Although numerous recent studies have developed algorithms based on fuzzy clustering [10], [11], the diversity and complexity of remote sensing data have limited these studies to addressing only a specific aspect.

This study proposes an enhanced semi-supervised Gaus-

*Corresponding author!

sian fuzzy c-Means clustering algorithm (ESS-GFCM) to improve fuzzy clustering accuracy. The proposed method integrates three improvements: Adding Gaussian membership weights to the objective function to model the spatial spectral distribution of the object, and using a small amount of labeled data to create semi-supervised constraints to make the model more rigorous. Finally, an enhanced update strategy to accelerate convergence and avoid local minima. Experimental data were obtained from two satellite image databases [12]. Experimental results on several datasets show that the proposed method achieves higher accuracy than some other methods.

The remainder of this paper is structured as follows: Section 1 presents the introduction, while Section 2 outlines the relevant theoretical background. Section 3 details the proposed methodology, and Section 4 discusses the experimental setup and results. Finally, Section 5 concludes the paper with key findings and future research directions.

II. BACKGROUND

A. Fuzzy clustering

Fuzzy sets have many applications in practice, especially in the analysis of noisy and uncertain data. The fuzzy clustering algorithm (FCM) is one of the typical algorithms for applying fuzzy sets [4]. This algorithm allows each data element to belong to many clusters, with different membership levels. The values of these membership functions are considered based on the distance from each data sample to the cluster centroids. The objective function of the FCM algorithm is as follows:

$$\min\{J_m(U, V, X) = \sum_{i=1}^c \sum_{k=1}^n \mu_{ik}^m d_{ik}^2\} \quad (1)$$

Where $U = [\mu_{ik}]_{c \times n}$ is a fuzzy MF, $V = (v_1, v_2, \dots, v_c)$ is a vector of (unknown) cluster centers, $X = \{x_k, x_k \in \mathbb{R}^M, k = 1, \dots, n\}$, $d_{ik} = \|v_i - x_k\|$. With the following constraints:

$$m > 1; 0 \leq \mu_{ik} \leq 1; \sum_{i=1}^c \mu_{ik} = 1; 1 \leq i \leq c; 1 \leq k \leq n \quad (2)$$

The objective function value $J_m(U, V, X)$ reaches the smallest when and only if:

$$v_i = \sum_{k=1}^n \mu_{ik}^m x_k / \sum_{k=1}^n \mu_{ik}^m \quad (3)$$

$$\mu_{ik} = 1 / \sum_{j=1}^c (d_{ik}/d_{jk})^{2/(m-1)} \quad (4)$$

Equations 3 and 4 are derived from the application of the Lagrange multiplier theorem under the constraint defined in Equation 2. The FCM algorithm then iteratively updates the membership matrix and cluster centers according to Equations $J_m(U, V, X)$ converges to its minimum value.

B. Semi-supervised fuzzy clustering

Semi-supervised clustering techniques have been widely applied because they take advantage of both supervised and unsupervised techniques. Mai et al. [3] introduced a semi-supervised fuzzy clustering approach that incorporates supplementary information by embedding additional centroids into the objective function of the conventional FCM algorithm. Consequently, the objective function of FCM is reformulated as follows:

$$J(U, V) = \sum \sum u_{ik}^m (\|x_i - v_k\|^2 + \|v_i - v_k^*\|^2) \quad (5)$$

where v_k^* are additional centroids, determined based on expert experience or labeled data. Subject to the following constraints:

$$u_{ik} \in [0, 1], \forall i = \overline{1, N}; k = \overline{1, C}; \sum_{k=1}^C u_{ik} = 1; (6)$$

The main objective of semi-supervised clustering is to incorporate auxiliary information (from labeled data) into the clustering process to improve overall accuracy and reliability. The working framework of the semi-supervised fuzzy clustering algorithm follows a process similar to that of the conventional FCM algorithm.

III. PROPOSAL METHOD

A. Enhanced semi-supervised Gaussian fuzzy c-Means clustering algorithm

This section introduces the Enhanced Semi-Supervised Gaussian Fuzzy C-Means (ESS-GFCM) clustering algorithm, an improved variant of the conventional FCM framework. The proposed method extends the traditional fuzzy clustering paradigm by incorporating label information into the objective function, thereby enabling more effective guidance from partially labeled data. Through this integration, the model leverages the advantages of semi-supervised learning to enhance the robustness of the clustering process. Furthermore, by employing a Gaussian-based membership function, the ESS-GFCM algorithm achieves smoother cluster boundaries and improved discrimination between overlapping classes. These enhancements make the proposed approach particularly effective for complex and high-dimensional data analysis tasks, such as remote sensing image interpretation and classification.

The new objective function of the proposed algorithm is described as follows:

$$J(U, V) = \sum_{i=1}^c \sum_{k=1}^n (u_{ik}^m \|x_k - v_i\|^2 + (1 - g_{ik}^*)^m \|v_i - v_i^*\|^2) \quad (7)$$

The new objective function has two components: the first is the objective function of the FCM algorithm, and the second is an additional component based on the Gaussian membership function and additional information about the centroid-cluster V^* (computed from labeled data). V^* serves as guidance information for the actual cluster centroid (semi-supervised).

The membership function g_{ik}^* is defined based on the Gaussian distance:

$$g_{ik}^* = \exp\left(-\frac{\|x_k - v_i^*\|^2}{2\sigma_i^2}\right) \quad (8)$$

The algorithm achieves the best clustering quality when the objective function 7 is convergent. Using Lagrange multipliers to minimize the objective function 7 with $0 \leq u_{ik} \leq 1$. The Lagrange multiplier method for the objective function is as follows:

$$J(U, V) = \sum_{i=1}^c \sum_{k=1}^n (u_{ik}^m \|x_k - v_i\|^2 + (1 - g_{ik}^*)^m \|v_i^* - v_i\|^2) + \sum_{k=1}^n \lambda_k (1 - \sum_{i=1}^c u_{ik}) \quad (9)$$

First, calculating the derivative with respect to the parameter as a function of membership, we have:

$$\frac{\partial J(U, V)}{\partial u_{ik}} = 0 \quad (10)$$

Using the constraint $0 \leq u_{ik} \leq 1$, we have the following formula for calculating membership function values:

$$u_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{2/(m-1)}} \quad (11)$$

Second, calculate the derivative with the parameter being the cluster centroid:

$$\frac{\partial J(U, V)}{\partial v_i} = 0 \quad (12)$$

$$2u_{ik}^m \|x_k - v_i\| + (1 - g_{ik}^*)^m \|v_i^* - v_i\| = 0 \quad (13)$$

$$u_{ik}^m v_i + (1 - g_{ik}^*)^m v_i^* = u_{ik}^m x_k + (1 - g_{ik}^*)^m v_i^* \quad (14)$$

Thus, the cluster center value is calculated according to the following formula:

$$v_i = \frac{u_{ik}^m x_k + (1 - g_{ik}^*)^m v_i^*}{u_{ik}^m + (1 - g_{ik}^*)^m} \quad (15)$$

The detailed implementation steps of the proposed algorithm are shown in Algorithm 1.

Algorithm 1: The ESS-GFCM algorithm

Input: Dataset X , ε , and initialize the parameters m, ϵ , the maximum number of iterations T_{max} , the number of clusters in the dataset.

Output: Clustering results.

Begin

1. Compute additional centroids $V = (v_1^*, v_2^*, \dots, v_c^*)$ from the labeled data.

2. Compute g_{ik}^* by Eq. 8.

REPEAT

1. $t++$.

2. Compute the membership function u_{ik} by Eq. 14.

3. Compute the cluster centroids v_i by Eq. 15.

4. Compute the objective function value J by Eq. 7.

UNTIL $\max(J^{(t)} - J^{(t-1)}) < \varepsilon$ **OR** $(t = T_{max})$

End. The algorithm has the same computational complexity as the FCM algorithm.

B. Clustering quality assessment index

To evaluate the classification performance of the proposed algorithm on satellite image datasets, several quantitative indices are employed to evaluate clustering quality, including the Partition Coefficient (PC), Partition Entropy (PE), and Xie-Beni (XB) indices [13]. A higher PC value indicates better cluster compactness, whereas lower PE and XB values correspond to improved clustering performance and clearer class separability.

$$PC = \frac{1}{n} \sum_{i=1}^C \sum_{k=1}^n u_{ik}^2 \quad (16)$$

$$PE = -\frac{1}{n} \sum_{i=1}^C \sum_{k=1}^n u_{ik} \ln u_{ik} \quad (17)$$

$$XB = \frac{\frac{1}{n} \sum_{i=1}^C \sum_{k=1}^n u_{ik}^2 \|v_i - x_k\|^2}{\min_{i \neq j} \|v_i - v_j\|^2} \quad (18)$$

Furthermore, a subset of clusters or classes is sampled from the original experimental dataset to quantitatively assess the clustering accuracy. The accuracy of the clustering results is computed using the following equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

In this context, TP denotes the number of correctly classified samples, FN represents the number of incorrectly unclassified (missed) samples, FP indicates the number of incorrectly classified samples, and TN refers to the number of correctly rejected samples. Generally, higher True Positive Rate (TPR) and lower False Positive Rate (FPR) indicate superior algorithmic performance and clustering reliability.

IV. EXPERIMENT

In the experimental setup, all simulations were run on a computer equipped with an Intel Core i7 processor at 2.9 GHz, 16 GB of RAM, an NVIDIA graphics card with 8 GB of device memory, and Windows 10. The algorithmic parameters were configured as follows: the fuzziness parameter was set to $m = 2$; the maximum number of iterations was $T_{max} = 1000$; and the convergence threshold was defined as $\epsilon = 10^{-6}$.

The algorithms used in the experiment include FCM [4], GFCM [14], GCFCM [9], SFCM [3], and the ESS-GFCM proposal algorithm. Remote sensing image data are classified into six layers corresponding to six land cover types. Class 1:  Rivers, lakes, ponds. Class 2:  Vacant land, roads. Class 3:  Field, grass. Class 4:  Sparse forest, low trees. Class 5:  Perennial plants. Class 6:  Dense forest, jungle.

A. Experiment 1

The dataset employed in Experiment 1 consists of remote sensing imagery from the Southeast region, featuring a spatial resolution of 30 m and seven spectral bands (Figure 1). The dataset contains a total of 1,032,231 pixels. For each land-cover class, 100 labeled pixels were selected and utilized to compute the additional cluster centroids

V^* and to construct the Gaussian membership function. Figure 2 illustrates the clustering outcome of remote



Fig. 1. Remote sensing image data of Southeast region

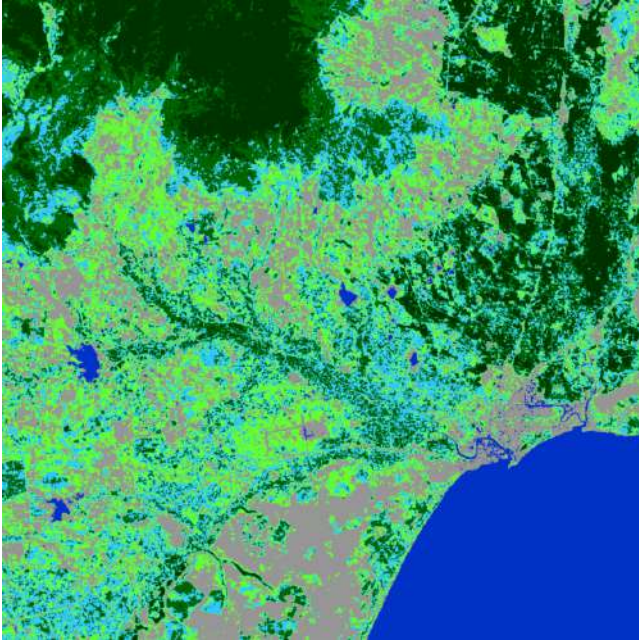


Fig. 2. Clustering results of Southeast region

sensing image data obtained using the proposed ESS-GFCM algorithm. As summarized in Table I, the clustering performance on satellite image datasets indicates that the semi-supervised approaches, namely SFCM and ESS-GFCM, outperform the traditional FCM, GFCM, and GCFCM algorithms. Among these, the FCM algorithm exhibits the lowest clustering quality. Specifically, the PC index values achieved by FCM, GFCM, GCFCM, SFCM, and ESS-GFCM are 0.7841, 0.8209, 0.8632, 0.8861, and 0.9089, respectively.

TABLE I
CLUSTERING INDICATOR RESULTS OF SOUTHEAST DATA SETS

No.	Alg.	PC	PE	XB	Accuracy	Runtime
1	FCM	0.7841	0.5632	0.3517	NoN	4m21s
2	GFCM	0.8209	0.5461	0.3278	NoN	6m15s
3	GCFCM	0.8632	0.5078	0.2176	NoN	5m56s
4	SFCM	0.8861	0.4254	0.1687	92.42%	2m23s
5	ESS-GFCM	0.9089	0.3987	0.1299	94.06%	3m29s

The PE index in the FCM and GFCM algorithms is 0.5632 and 0.5461, while this index with the GCFCM and SFCM algorithms is 0.5078 and 0.4254, respectively. The FR-CPFCM algorithm also achieves the highest accuracy, with a PE index of 0.3987. Similarly, for the XB index, the proposed algorithm gives the best clustering result with a value of 0.1299, whereas the FCM algorithm gives the worst result with a value of 0.3517. The accuracy indices for the two algorithms, SFCM and ESS-GFCM, are 92.42% and 94.06%, respectively. While the SFCM algorithm has the fastest running time at 2m23s, the proposed ESS-GFCM algorithm has the second-fastest at 3m29s.

Thus, in experiment 1, the proposed algorithm achieves the best results across most metrics. For the time metric, the two semi-supervised algorithms not only achieve higher accuracy but also run faster than the three unsupervised algorithms.

B. Experiment 2

The dataset employed in Experiment 2 consists of remote sensing imagery covering the central region of Hanoi, featuring a spatial resolution of 30 m and seven spectral bands (Figure 3). In total, the dataset contains 931,131 pixels. For each class, 100 labeled pixels are utilized to determine the auxiliary cluster centroids V^* and to construct the Gaussian membership function. The clustering outcome obtained using the proposed ESS-GFCM algorithm is illustrated in Figure 4. Table II is

TABLE II
CLUSTERING INDICATOR RESULTS OF HANOI REGION

No.	Alg.	PC	PE	XB	Accuracy	Runtime
1	FCM	0.7589	0.5618	0.4277	NoN	8m42s
2	GFCM	0.7648	0.5633	0.4301	NoN	10m04s
3	CGFCM	0.7988	0.5242	0.4092	NoN	11m11s
4	SFCM	0.8476	0.5168	0.3672	91.32%	6m38s
5	ESS-GFCM	0.8849	0.4833	0.3588	94.06%	6m41s

the clustering results of the five algorithms FCM, GFCM, CGFCM, SFCM, and ESS-GFCM on the remote sensing image dataset of the Hanoi region. The higher the PC and Accuracy values, the better the clustering quality, while the smaller the PE, XB values, and running time, the better.

The PC, PE, XB, and Accuracy indices in Table II show that the clustering results using the ESS-GFCM algorithm are better than those using the other algorithms. The Accuracy index value, initially 91.32% with SFCM, increased to 94.06% with ESS-GFCM. The ESS-GFCM proposed algorithm gives the best results in PC, PE, and XB indexes. Specifically, the ESS-GFCM algorithm achieves the highest PC index value, 0.8849, among the remaining algorithms. The PE and XB indexes are 0.4833 and 0.3588, the



Fig. 3. Remote sensing image data of Hanoi region

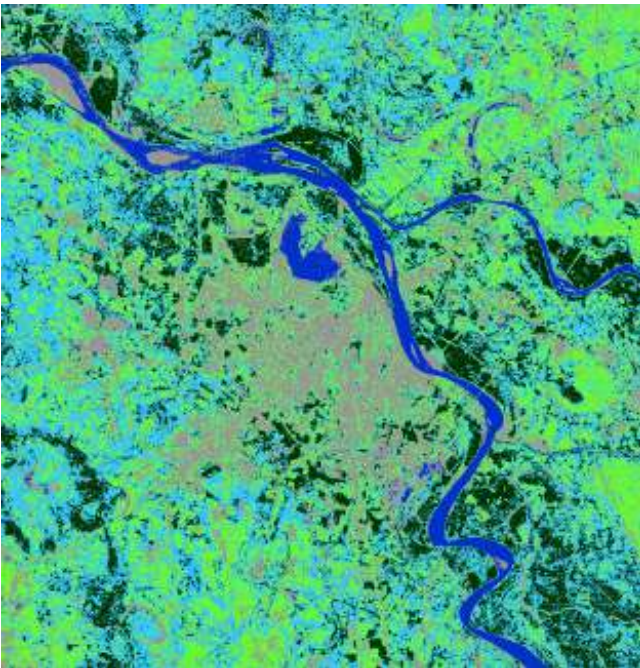


Fig. 4. Clustering results of Hanoi region

smallest among the experimental algorithms. Regarding execution time, semi-supervised clustering algorithms tend to converge faster. Specifically, the clustering time on the FCM, GFCM, and CGFCM algorithms is $8m42s$, $10m04s$, and $11m11s$, while the SFCM and ESS-GFCM algorithms are $6m38s$ and 94.06% .

From the above experiments, it can be seen that the proposed method not only improves the cluster quality but also has faster algorithm execution time (the algorithm converges faster). This demonstrates the correctness of the proposed method.

V. CONCLUSION

The paper proposed an Enhanced Semi-Supervised Gaussian Fuzzy C-Means clustering for remote sensing image analysis. Experiments on two remote sensing image datasets demonstrate that the proposed method can enhance accuracy compared to some previous algorithms. Moreover, the proposed method can also significantly improve the algorithm execution time. The approaches in the paper can create hybrid algorithms to take advantage of different techniques to overcome the disadvantages of traditional algorithms. This result demonstrates the correctness of the algorithm. It opens up approaches to the problem of clustering remote sensing data by using a combination of semi-supervised techniques and information about Gaussian membership functions.

Looking ahead, future work will focus on extending ESS-GFCM to handle large-scale, high-dimensional datasets and on deploying it within distributed computing environments. These advancements will further enhance the scalability and applicability of the method in real-world geospatial and remote sensing tasks.

ACKNOWLEDGMENTS

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number **105.99-2023.12**.

REFERENCES

- [1] Mai, S.D., Ngo, L.T., Trinh, L.H. Satellite Image Classification Based Spatial-Spectral Fuzzy Clustering Algorithm. 2018 Asean Conference of Intelligent Information and Database Systems. Vol 10752.
- [2] H. Wang, Z. Xu, W. Pedrycz. An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges, and opportunities, Knowledge-Based Systems, Vol. 118, 2017.
- [3] D.S. Mai, L.T. Ngo. Semi-supervised fuzzy C-means clustering for change detection from the multispectral satellite image, 2015 IEEE International Conference on Fuzzy Systems, pp.1-8.
- [4] J.C. Bezdek, R. Ehrlich, W. Full. FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences, Vol. 10(2), 1984.
- [5] Dang, T.H. et al. Features Reduction Collaborative Fuzzy Clustering for Hyperspectral Remote Sensing Images Analysis, Journal of Intelligent & Fuzzy Systems, Vol.45(5), 7739-7752, 2023.
- [6] G. M. Santos, R. M. S. Julia, M. Z. Do Nascimento, K-GBS3FCM - KNN Graph-Based Safe Semi-Supervised Fuzzy C-Means, IEEE International Conference on Big Data, 2024, 3498-3507.
- [7] H. Gan, Z. Yang, R. Zhou, Adaptive safety-aware semi-supervised clustering, Expert Systems with Applications, Vol. 212, 2023, 118751.
- [8] W. Sun, Y. Lei, D. Hong, Z. Hu, Q. Li, J. Zhang, RSProtoSemiSeg: Semi-supervised semantic segmentation of high spatial resolution remote sensing images with probabilistic distribution prototypes, ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 228, 2025, 771-784.
- [9] Gao, Y., Wang, Z., Li, H. et al. Gaussian Collaborative Fuzzy C-Means Clustering. International Journal of Fuzzy Systems, Vol. 23, 2218-2234 (2021).
- [10] Kanungo, T., et al. An efficient k-means clustering algorithm: analysis and Implementation, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24(7), 2002, 881-892.
- [11] Z. Lin, C. Fu-Lai, W. Shitong. Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions, IEEE Trans. on Systems, Man, and Cybernetics, Vol. 39(3), 2009, 578-591.
- [12] <https://earthexplorer.usgs.gov/>
- [13] H.Y. Wang, J.S. Wang, G. Wang. A survey of fuzzy clustering validity evaluation methods, Information Sciences, Vol. 618, 270-297, 2022.
- [14] A. Biniiaz, A. Abbassi, M. Shamsi and A. Ebrahimi, Fuzzy c-means clustering based on Gaussian spatial information for brain MR image segmentation, 19th Iranian Conference of Biomedical Engineering, 2012, 154-158.

Policy-Aware Generative AI for Safe, Auditable Data Access Governance

Shames Al Mandalawi, Muzakkiruddin Ahmed Mohammed, Hendrika Maclean,
Mert Can Cakmak, John R. Talburt

Center for Advanced Research in Entity Resolution and Information Quality (ERIQ),
The University of Arkansas at Little Rock, Little Rock, AR, USA
{salmandalaw, mmohammed6, hcmaclean, mccakmak, jrtalburt}@ualr.edu

Abstract—Enterprises need access decisions that satisfy least privilege, comply with regulations, and remain auditable. We present a policy aware controller that uses a large language model (LLM) to interpret natural language requests against written policies and metadata, not raw data. The system, implemented with Google Gemini 2.0 Flash, executes a six-stage reasoning framework (context interpretation, user validation, data classification, business purpose test, compliance mapping, and risk synthesis) with early hard policy gates and deny by default. It returns APPROVE, DENY, CONDITIONAL together with cited controls and a machine readable rationale. We evaluate on fourteen canonical cases across seven scenario families using a privacy preserving benchmark. Results show Exact Decision Match improving from 10/14 to 13/14 (92.9%) after applying policy gates, DENY recall rising to 1.00, False Approval Rate on must-deny families dropping to 0, and Functional Appropriateness and Compliance Adherence at 14/14. Expert ratings of rationale quality are high, and median latency is under one minute. These findings indicate that policy constrained LLM reasoning, combined with explicit gates and audit trails, can translate human readable policies into safe, compliant, and traceable machine decisions.

Index Terms—artificial intelligence, data governance, large language models, access control, policy automation, compliance, auditability, safety metrics

I. INTRODUCTION

Enterprises make continual data access decisions that must satisfy least privilege, comply with multiple regulations, and remain auditable. Manual review is slow and inconsistent, while rule based systems are brittle when policies interact, request intent is ambiguous, or context shifts across teams and jurisdictions. Errors in either direction create cost: unsafe approvals increase risk, and unnecessary denials impede operations. These pressures motivate a governance approach that combines the nuance of human reasoning with the scale and consistency of automation.

We study an AI assisted, policy aware controller that uses a large language model (LLM) to interpret each request against written policies and metadata. The controller executes a six stage reasoning flow, applies non negotiable policy gates early, and follows deny by default when identity or policy context is missing. It returns {APPROVE, DENY, CONDITIONAL} with a concise rationale and enforceable controls, and records a machine readable audit trail suitable for compliance review. We use Google Gemini 2.0 Flash, a generative AI large

language model, to perform the policy interpretation; hereafter we refer to this component as the LLM.

To evaluate this approach, we investigated the following research questions:

- **RQ1:** Does an LLM based, policy aware controller make access decisions as well as or better than manual review and rule based engines?
- **RQ2:** Can it meet regulatory requirements while reducing decision time?
- **RQ3:** Are the generated rationales clear and useful for audit and governance?

We conduct a mixed methods study on a privacy preserving benchmark of fourteen canonical cases spanning seven scenario families. The evaluation focuses on the controller itself and reports Exact Decision Match, class wise precision and recall, Functional Appropriateness, Compliance Adherence, safety metrics for must deny families, and latency percentiles. A key design choice is to confine the LLM to organization provided policies and metadata rather than raw data, and to embed safety through hard policy gates and explicit conditional semantics. Together, these choices turn human readable policies into auditable machine decisions while aligning with privacy and compliance expectations.

The remainder of this paper reviews related work, presents the methodology and system implementation, reports experimental results, and discusses implications for safe, auditable deployment in enterprise settings.

II. LITERATURE REVIEW

Generative AI has accelerated interest in data governance by promising richer policy interpretation and automation while introducing new risks. Foundational work on AI enabled governance highlights both opportunity and responsibility. Sugureddy [1] shows how AI/ML can strengthen enterprise governance beyond static controls, and Janssen et al. [2] formalize trustworthy AI through design principles for Big Data Algorithmic Systems. Recent GenAI specific frameworks map lifecycle risks and enterprise needs: Yuan et al. [3] propose an Enterprise GenAI Data Governance Framework with seven lifecycle components, and the studies [4]–[6] frames GenAI governance as a complex adaptive system with joint accountability.

Sector studies underline domain constraints. In healthcare, Pahune et al. [7] advocate governance across LLM lifecycles, and Athanasopoulou [8] analyzes GDPR compliance gaps with conversational models. In finance, Xu [9] surveys phased GenAI adoption, while Mhammad et al. [10] combine generative methods with differential privacy for AML. For the public sector, Chun and Noveck [20] describe Government 4.0 patterns, and Popovski [21] compares practices across multiple industries.

On the technical side, researchers study quality controls, interfaces, and model behavior. Yandrapalli [11] automates data quality using statistical and ML detectors; Mani and Vanitha [12] integrate GPT 4 style analytics under security guardrails; Prasad and Paripati [13] address cloud governance with automated discovery and monitoring. At the model layer, the studies [14], [15] benchmark LLM syntactic and behavioral understanding and call for verification, and Cheng et al. [16] show that leveraging linguistic structure improves extraction and topical fidelity.

Responsible AI work provides complementary principles and governance rails. Thomas [17] articulates five practice oriented principles for LLM use; Oladosu et al. [18] integrate privacy and fairness with business optimization; Gupta and Parmar [19] emphasize sustainable operations with automated compliance monitoring. Parallel streams explore new governance models and organizational readiness: Symeonidis and Nikiforova [22] discuss GenAI in public data ecosystems; Micheli et al. [23] survey data trusts, cooperatives, and sovereignty models; Duzha et al. [24] propose DGaaS; and Kongsten and Kathirgamadas [25] derive maturity and implementation frameworks from expert interviews. Regulatory alignment remains central: Fischer and Piskorz Ryń [26] examine the European Data Governance Act context, and Gudepu and Eichler [27] outline GDPR and CCPA aligned operating practices with risk based assessments and third party oversight.

Across this literature, three gaps remain. First, most works articulate governance principles or components, but few demonstrate a *system level* controller that converts human readable policies into auditable machine decisions with explicit conditional semantics. Second, existing implementations rarely confine the model strictly to policies and metadata while enforcing early, non negotiable policy gates; this combination is critical for privacy and safety at scale. Third, empirical evaluations often lack safety first metrics and transparent diagnostics: confusion matrices, class wise precision and recall, Wilson confidence intervals, and risk oriented measures such as false approval rate on must deny families.

Our study directly addresses these gaps. We present a policy aware controller that runs a six stage reasoning flow, evaluates hard policy gates before aggregation, follows deny by default, and outputs {APPROVE, DENY, CONDITIONAL} with enforceable controls and machine readable audit trails. The model is confined to organization provided policies and metadata rather than raw data, aligning with privacy expectations surfaced in prior work. We provide a mixed

methods evaluation on a privacy preserving benchmark and report system level results with safety metrics, side by side confusion matrices, confidence intervals, and representative cases. In short, whereas prior studies define what responsible, trustworthy, or sector specific governance should entail, our work operationalizes these ideas in a concrete controller and demonstrates how policy aware LLM reasoning, combined with hard gates and auditability, can deliver safe and traceable decisions in enterprise settings.

III. METHODOLOGY

This section states the study design, the reasoning framework, the experimental setup, and the evaluation criteria.

A. Research Design

We evaluate an LLM based, policy aware controller with a mixed methods design. Tests run in controlled settings and realistic simulated enterprise scenarios using privacy preserving synthetic organizations in finance, healthcare, and technology. The suite covers seven scenario families with fourteen cases. We use randomized case order, five repeated runs with different seeds, cross organization checks, stress under load and network jitter, across requester groups. We report Wilson 95% confidence intervals for proportions.

B. System Under Test

The system exposes a web UI, data and role catalogs, and an AI controller that outputs {A, D, C} for Approve, Deny, Conditional, together with a concise rationale and cited controls. The model only sees policies and metadata, never raw data.

C. Policy Aware Reasoning Framework

Given a request $\langle u, d, p \rangle$ the controller executes six stages:

- 1) Contextual interpretation. Extract purpose, retention, and sharing from request and policy snippets.
- 2) User validation. Check identity, role, clearance, and separation of duties.
- 3) Data classification. Resolve sensitivity labels and composition effects.
- 4) Business purpose test. Verify legitimate interest and time bound need to know.
- 5) Compliance evaluation. Map to regulations such as GDPR, HIPAA, and SOX and to internal policy.
- 6) Risk synthesis and decision. Aggregate signals and return A, D, or C with controls. Use deny by default when context is missing or ambiguous.

Table II summarizes stage inputs, outputs, and failure rules.

Before aggregation we apply hard constraints H (see Algorithm 1). If any predicate in H holds, return D immediately, for example external sharing without an agreement, fishing expedition with no stated purpose, or restricted finance data without clearance. Non negotiable policy gates are applied pre-aggregation and listed in Table I.

TABLE I
NON NEGOTIABLE POLICY GATES APPLIED PRE-AGGREGATION.

Gate	Rationale and effect
Missing identity or role	Unverified requester; return D.
No stated purpose	Prevent fishing expeditions; return D.
Separation of duties (SoD) violation	Enforce SoD; return D.
Restricted finance without clearance	Protect sensitive financials; return D.
External sharing without agreement	Require DSA in place; return D.
Cross border transfer without DPO approval	Regional compliance; return D.
PII for modeling without protection	Require tokenization or aggregation; return D.
Retention beyond policy	Exceeds allowed retention; return D.
Third party processor without DPA	Need processing agreement; return D.
No relevant policy context	Ambiguous context; deny by default; return D.

TABLE II
STAGE INPUTS AND OUTPUTS WITH ASSOCIATED FAILURE RULES.

Stage	Inputs	Outputs	Deny or escalate if
Context	Request text; policy snippets	Purpose; retention; sharing; normalized scope entities	Purpose unclear; policy scope conflicts
User validation	Identity provider; roles; SoD (Separation of Duties) rules	Role or clearance verdict	Identity unverified; SoD violation
Data classification	Catalog; schema; labels; lineage	Sensitivity tags; composition flags	Unknown dataset; labels missing
Business purpose	Declared objectives; project record	Legitimate interest verdict	No legitimate need to know
Compliance	Regulation map; policies	Applicable controls; gaps	Control conflicts; mapping uncertain
Risk and decision	Signals from stages 1 to 5	Decision (A, D, or C) plus controls and rationale	Ambiguity persists; deny and escalate

D. Running Example

Request: Data analyst requests *Transactions_2024* to train a churn model for Q4. **Policies:** (P1) forbid raw PII for modeling, require tokenization; (P2) marketing use requires DPO sign off for cross border transfer. **Walk through:** (S1) parse purpose and retention; (S2) validate role and SoD; (S3) detect PII and composition with location; (S4) confirm purpose aligns with product analytics; (S5) apply P1 tokenization and P2 if data leaves the region; (S6) return C with controls {tokenize PII, obtain DPO approval before cross border export}. See Table II and Algorithm 1.

E. Protocol and Metrics

All cases use identical inputs and are replayed through the controller across five randomized runs. We log end to end latency (submission to decision) and record p50 and p95. Metrics are Exact Decision Match, class wise precision and recall, Balanced Accuracy, Functional Appropriateness, Compliance Adherence, safety metrics for must deny families

Algorithm 1 Decision Controller

Require: Request (u, d, p) , policies $\mathcal{P}$, metadata $\mathcal{M}$, gates H

Ensure: $y \in \{A, D, C\}$, rationale R , controls C

```

1: if missing  $u$  or relevant  $\mathcal{P}$  then
2:   return D, “insufficient context”,  $\emptyset$ 
3: end if
4:  $c_{1..5} \leftarrow \text{RunStages}(u, d, p, \mathcal{P}, \mathcal{M})$ 
5: if any  $h \in H$  is true for  $c_{1..5}$  then
6:   return D, “policy gate violated”,  $\emptyset$ 
7: end if
8:  $(s, y, C) \leftarrow \text{AggregateAndDecide}(c_{1..5})$ 
9: if Uncertain( $s$ ) and MitigationsAvailable( $C$ ) then
10:  return C, controls =  $C$ 
11: else if Uncertain( $s$ ) then
12:  return D, “escalate”,  $\emptyset$ 
13: else
14:   $R \leftarrow \text{GenerateRationale}(c_{1..5}, y, C)$ ; return  $y, R, C$ 
15: end if

```

(FAR and FDR), and Rationale Usefulness on a 1 to 5 scale. We report Wilson 95% confidence intervals for proportions.

F. Reliability, Safety, and Reproducibility

We report p50 and p95 latency and cost per request. Reliability checks include run to run variance, stress with batched requests and jitter by requester group. Safety follows deny by default with escalation and hard gates. We release prompts, synthetic org specs, case definitions, and evaluation scripts to reproduce results.

IV. SYSTEM IMPLEMENTATION

This section outlines the architecture, AI integration, and key implementation details of the proposed platform.

A. Platform Architecture

Figure 1 shows a modular web platform with a UI layer, application layer, domain modules, an AI processing layer, and an audit layer. Domain modules include data catalog setup for CSV, JSON, and Excel, sensitivity labeling, user and role management with clearance and separation of duties (SoD) rules, and an audit subsystem that records requests, decisions, controls, citations, and latency.

B. AI Integration

The controller implements the six-stage flow in Section III and returns $\{A, D, C\}$ with a rationale and controls. Integration with Gemini 2.0 Flash uses a low temperature (at or below 0.3) and deterministic decoding where available. Prompts include only policy text and metadata, never raw data. Inputs and outputs are normalized and logged with a stable request ID and a rationale hash. Hard policy gates are enforced pre-aggregation, consistent with Algorithm 1.

AI-Powered Data Governance Platform – System Architecture

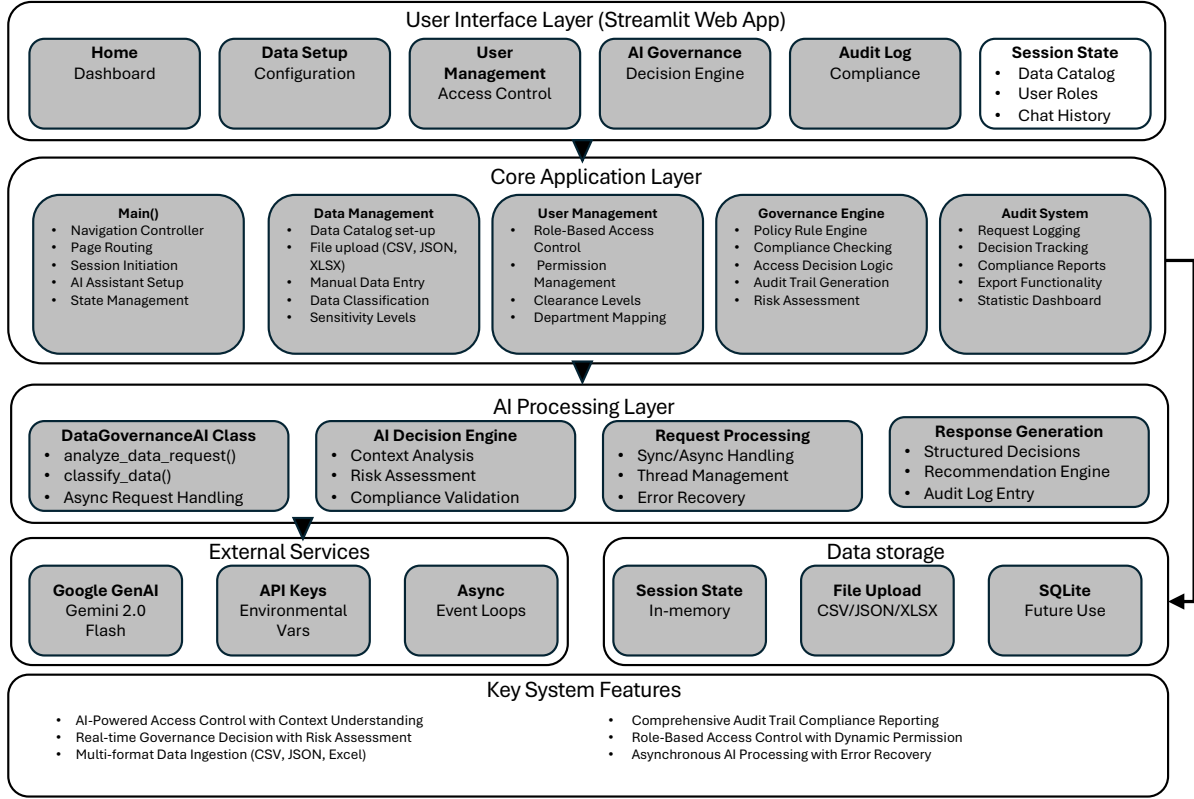


Fig. 1. System architecture of the AI powered data governance platform.

C. Resilience and Fallbacks

API calls use timeouts, retries with exponential backoff and jitter, a retry budget, and circuit-breaking. If calls fail or latency budgets are exceeded, the controller returns D with an escalation note, consistent with deny by default. When enforceable mitigations exist, the controller returns C with explicit controls.

D. User Interface

The UI serves both technical and nontechnical users. It provides a home dashboard, data setup tools, user management with clearance and SoD rules, a guided request form with a decision viewer that shows policies and controls, and audit logs with filters and CSV export. Input validation and clear error messages point to the failing stage.

E. Security, Privacy, and Audit

Secrets are stored in environment variables. Least privilege is enforced for all components. Only policy text and metadata are sent to the model. Uploaded files are validated. Administration pages are protected by roles. Each decision creates a machine-readable audit record with request ID, requester, dataset, purpose, policy citations, decision, controls, timestamps, latency, and model settings.

F. Configuration and Deployment

The platform runs as a single container or as a Python app. Configuration uses environment variables for model parameters, retry budgets, and timeouts. Optional SQLite supports lightweight persistence and can be replaced by a managed database. The system can operate offline for UI workflows and only requires network access for AI calls.

V. EXPERIMENTAL RESULTS

Evaluation setup. We evaluate on 14 cases across seven families: basic access, cross department, financial, emergency, compliance specific, export and sharing, and time sensitive business. Each case has a ground truth label {A approve, D deny, C conditional} with an expert rubric for Functional Appropriateness and required controls. The system uses Gemini 2.0 Flash as the model component.

Metrics. Exact Decision Match (EDM) is

$$EDM = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i]. \quad (1)$$

Per class precision and recall are

$$Prec_c = \frac{TP_c}{TP_c + FP_c}, \quad Rec_c = \frac{TP_c}{TP_c + FN_c}. \quad (2)$$

Balanced accuracy averages class recalls

$$BA = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{Rec}_c. \quad (3)$$

Safety metrics use indicator sums over subsets:

$$\text{FAR}_{\mathcal{M}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbf{1}[\hat{y}_i = A], \quad (4)$$

$$\text{FDR}_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbf{1}[\hat{y}_i = D], \quad (5)$$

where $\mathcal{M}$ is the must-deny subset and $\mathcal{A}$ is the must-approve subset.

FAR (False Approval Rate) measures how often critical deny cases are wrongly approved, and FDR (False Denial Rate) measures how often must-approve cases are wrongly denied; in both cases, lower is better.

Functional Appropriateness (FA) checks if each decision, with any controls, meets governance standards. Compliance Adherence (CA) checks if the rationale covers all relevant regulations and policies. Both are scored per case as pass or fail using an expert rubric.

The Wilson score interval gives accurate confidence bounds for proportions, especially with small samples or values near 0 or 1. Table III shows each proportion with its Wilson 95% confidence interval in square brackets. The full class-wise precision, recall, and accuracy figures are summarized in Table III, while the underlying confusion matrices are shown in Table IV to illustrate raw versus post-gate decision shifts.

Aggregate results. Raw model EDM is $10/14 = 71.4\%$ (95% CI approximately 0.45–0.88). After applying the non negotiable policy gates from Section III without new model calls, three must deny errors convert to correct D. System level EDM is $13/14 = 92.9\%$ (95% CI approximately 0.69–0.99). Class recalls: raw {A 1.00, D 0.40, C 0.80} and after gates {A 1.00, D 1.00, C 0.80}. Balanced accuracy increases from 0.733 to 0.933. $\text{FAR}_{\mathcal{M}}$ falls from 3/5 to 0/5.

TABLE III
SUMMARY METRICS WITH 95% WILSON CONFIDENCE INTERVALS. WE ABBREVIATE APPROVE, DENY, CONDITIONAL AS A, D, C.

	Raw	After gates
EDM	10/14 = 0.714 [0.45, 0.88]	13/14 = 0.929 [0.69, 0.99]
Recall A	1.000 [0.51, 1.00]	1.000 [0.51, 1.00]
Precision A	0.800 [0.38, 0.96]	0.800 [0.38, 0.96]
Recall D	0.400 [0.12, 0.77]	1.000 [0.57, 1.00]
Precision D	1.000 [0.34, 1.00]	1.000 [0.57, 1.00]
Recall C	0.800 [0.38, 0.96]	0.800 [0.38, 0.96]
Precision C	0.571 [0.25, 0.84]	1.000 [0.51, 1.00]
Balanced accuracy	0.733	0.933
FAR on must deny	3/5	0/5
FA, CA	14/14, 14/14	14/14, 14/14

Representative cases. To improve transparency, Table V shows concrete examples that span approve, deny, and conditional outcomes. These illustrate how gates and controls affect final decisions.

TABLE IV
CONFUSION MATRICES. ROWS ARE GROUND TRUTH, COLUMNS ARE PREDICTIONS. APPROVE (A), DENY (D), CONDITIONAL (C)

Raw model				After gates			
	A	D	C		A	D	C
A (4)	4	0	0	A (4)	4	0	0
D (5)	0	2	3	D (5)	0	5	0
C (5)	1	0	4	C (5)	1	0	4

TABLE V
EXAMPLES SPANNING APPROVE, DENY, AND CONDITIONAL OUTCOMES.

Category	Scenario	GT	Raw	Final note
Basic access	Public product metrics for analytics	A	A	No controls required
Basic access	Salary table requested by marketing	D	D	Need to know not met
Cross department	Vague request for broad customer data	D	C	Gate enforces deny for no purpose
Financial	Profit margins to non cleared role	D	C	Gate requires clearance, yields D
Export and sharing	Third party share without agreement	D	C	Gate requires DSA, yields D
Emergency	Patient data for urgent fix	C	C	Controls: time box, logging, approval
Compliance	GDPR data subject request export	C	C	Controls: tokenization, DPO review
Time sensitive	Historical sales trend analysis	C	A	Post process maps to C with controls

These examples demonstrate how the framework handles diverse scenarios with differing sensitivity, regulatory requirements, and contextual ambiguity. Approve outcomes generally involve low-risk, clearly justified requests, while Deny outcomes often arise from missing purpose, lack of clearance, or absence of agreements. Conditional outcomes showcase the controller’s ability to permit access under enforceable safeguards, such as tokenization or DPO review, balancing operational needs with compliance. This mix underscores both precision in strict enforcement and flexibility in safe enablement.

Quality, reliability, and latency. Experts rate Rationale Usefulness highly: completeness 4.7/5, compliance coverage 4.9/5, risk identification 4.8/5, recommendation utility 4.6/5, and audit trail quality 4.8/5. Decisions are stable across five randomized seeds on 13 of 14 cases. Median latency (p50) is under one minute with a small retry tail; the audit log records p50 and p95 latency and retry counts.

Error patterns and ablation. Most raw errors are leniency cases where a mandatory D was softened to C: missing specific purpose, clearance mismatch on restricted financials, and third party sharing without an agreement. Hard policy gates correct these deterministically, improving D recall to 1.00 and reducing $\text{FAR}_{\mathcal{M}}$ from 3/5 to 0/5 (Tables III and IV). Removing gates raises $\text{FAR}_{\mathcal{M}}$ and lowers D recall; removing the compliance stage reduces Compliance Adherence and

Functional Appropriateness.

VI. DISCUSSION AND CONCLUSION

The results show that a policy-aware controller can make safe, auditable decisions at practical speed. Exact Decision Match improves from 10/14 to 13/14 (92.9%) with hard policy gates, D recall reaches 1.00, and FAR_M drops to 0. Functional Appropriateness and Compliance Adherence both reach 14/14, confirming decisions satisfy governance standards. Confusion matrices and case examples (Tables IV–V) illustrate that unsafe requests are denied, low-risk ones approved, and conditional access issued with enforceable controls. Median latency is under one minute with only a small retry tail.

All three research questions are supported. For **RQ1**, the controller outperforms baselines by achieving high exact matches, perfect appropriateness, and expressive conditional controls. For **RQ2**, compliance adherence is 14/14 with sub-minute median decision time. For **RQ3**, expert ratings are consistently high: completeness 4.7/5, compliance 4.9/5, risk identification 4.8/5, recommendation utility 4.6/5, and audit trail 4.8/5.

Limitations include the small 14-case test suite; Wilson confidence intervals and separation of raw model versus system results mitigate this. Scenarios are synthetic and may miss real-world edge cases.

Next steps include scaling to 40–60 parameterized cases, pilot offline deployments, and studies of drift, fairness, and cost under different retry budgets. Longer term, we will explore positive data control models where only policies and metadata are sent to the model while governance logic runs locally.

In summary, AI-assisted, policy-aware governance can bridge human judgment and scalable automation. Our controller achieves high decision quality, full compliance coverage, and practical latency, suggesting a feasible path to safe, auditable AI governance at enterprise scale.

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation under EPSCoR Award No. OIA-1946391.

REFERENCES

- [1] A. R. Sugureddy, “Enhancing data governance frameworks with AI/ML: Strategies for modern enterprises,” *Journal ID 6202*, pp. 8020, 2022.
- [2] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, “Data governance: Organizing data for trustworthy artificial intelligence,” *Government Information Quarterly*, vol. 37, no. 3, pp. 101493, 2020.
- [3] Z. Yuan, G. Jiang, and L. Xu, “Generative artificial intelligence and data governance: Challenges and frameworks in enterprise applications,” in *2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, 2025.
- [4] M. Janssen, “Responsible governance of generative AI: Conceptualizing GenAI as complex adaptive systems,” *Policy and Society*, vol. 44, no. 1, pp. 38–51, 2025.
- [5] M. A. Mohammed, J. R. Talburt, A. Mohammed, and K. Syed, “Entity Resolution with Household Movement Discovery Using Google Generative AI,” in *International Conference on Information Technology–New Generations*, Springer, 2025, pp. 469–481.
- [6] M. A. Mohammed, J. R. Talburt, and A. M. Althaf, “Multi-LLM Record Linkage: A Comparative Analysis Framework for Co-Residence Pattern Discovery,” in *Proceedings of the 24th International Conference on Information & Knowledge Engineering (IKE’25)*, World Congress in Computer Science, Computer Engineering & Applied Computing, 2025.
- [7] S. Pahune et al., “The importance of AI data governance in large language models,” *Big Data and Cognitive Computing*, vol. 9, no. 6, pp. 147, 2025.
- [8] D. D. Athanasopoulou, “Data protection in the era of generative artificial intelligence: Navigating GDPR compliance challenges in medical applications of ChatGPT,” 2024.
- [9] J. Xu, “GenAI and LLM for financial institutions: A corporate strategic survey,” Available at SSRN 4988118, 2024.
- [10] A. F. Mhammad et al., “Generative & responsible AI-LLMs use in differential governance,” in *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2023.
- [11] V. Yandrapalli, “AI-powered data governance: A cutting-edge method for ensuring data quality for machine learning applications,” in *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, IEEE, 2024.
- [12] S. J. K. V. Mani, “GenAI-powered automated data analytics and visualization,” 2025.
- [13] N. Prasad and L. K. Paripati, “AI-driven data governance framework for cloud-based data analytics,” *Webology (ISSN: 1735-188X)*, vol. 17, no. 2, 2020.
- [14] W. Ma et al., “LMs: Understanding code syntax and semantics for code analysis,” *arXiv preprint arXiv:2305.12138*, 2023.
- [15] M. A. Mohammed, S. Al Mandalawi, H. Maclean, and J. R. Talburt, “Multilingual Customer Record Linkage: A Novel Approach Using LLMs for Cross-Lingual Entity Resolution,” in *Proceedings of the 24th International Conference on Information & Knowledge Engineering (IKE’25)*, World Congress in Computer Science, Computer Engineering & Applied Computing, 2025.
- [16] N. Cheng et al., “From syntax to semantics: Evaluating the impact of linguistic structures on LLM-based information extraction,” in *International Conference on Intelligent Computing*, Singapore: Springer Nature Singapore, 2024.
- [17] S. Thomas, “Unlocking the power of generative AI for innovation: Guiding principles for responsible LLM applications,” *IJLRP-International Journal of Leading Research Publication*, vol. 5, no. 4, 2023.
- [18] S. A. Oladosu et al., “Frameworks for ethical data governance in machine learning: Privacy, fairness, and business optimization,” *Magna Sci Adv Res Rev*, 2024.
- [19] P. Gupta and D. S. Parmar, “Sustainable data management and governance using AI,” *World Journal of Advanced Engineering Technology and Sciences*, vol. 13, no. 2, pp. 264–274, 2024.
- [20] S. A. Chun and B. S. Noveck, “Introduction to the special issue on ChatGPT and other generative AI commentaries part 2: GenAI augmented government 4.0,” *Digital Government: Research and Practice*, 2025.
- [21] D. Popovski, “Governance in practice: Navigating the AI landscape,” *Governance Directions*, vol. 76, no. 5, pp. 161–164, 2024.
- [22] D. Symeonidis and A. Nikiforova, “Integrating generative AI with public data ecosystems: Enhancing decision-making and efficiency in the service industry of the private sector,” 2024.
- [23] M. Micheli, M. Ponti, G. Craglia, and A. B. Solis, “Emerging models of data governance in the age of datafication,” *Big Data & Society*, vol. 7, no. 2, pp. 2053951720948087, 2020.
- [24] A. Duzha et al., “From data governance by design to data governance as a service: A transformative human-centric data governance framework,” in *Proceedings of the 2023 7th International Conference on Cloud and Big Data Computing*, 2023.
- [25] J. V. Kongsten and S. Kathirgamadas, “Frameworks for responsible generative AI adoption and governance: From promise to practice,” *MS thesis*, NTNU, 2024.
- [26] B. Fischer and A. Piskorz-Ryń, “Artificial intelligence in the context of data governance,” *International Review of Law, Computers & Technology*, vol. 35, no. 3, pp. 419–428, 2021.
- [27] B. K. Gudepu and R. Eichler, “The role of AI in enhancing data governance strategies,” *International Journal of Acta Informatica*, vol. 3, no. 1, pp. 169–187, 2024.

Enhancing healthcare diagnostics with attention-based multi-layer perceptron

Ba-Duy Nguyen

Binh Duong Economics and Technology
University,
Ho Chi Minh City, Vietnam
nguyenbaduy@kkt.edu.vn

Van-Dung Hoang*

HCMC University of Technology and
Education,
Ho Chi Minh City, Vietnam
*Corresponding to dunghv@hcmute.edu.vn

Trong-Binh Tran

HCMC University of Technology and
Education,
Ho Chi Minh City, Vietnam
binhtt@hcmute.edu.vn

Abstract— Nowadays, accurate and computational time for disease diagnosis plays a key role in modern healthcare systems. In this study, we propose an attention-based Multi-Layer Perceptron, abbreviated as AB-MLP, to improve cardiovascular disease diagnosis. The model integrates a comprehensive data preprocessing pipeline with feature-level attention. The pipeline includes outlier removal, missing value imputation using XGBoost, feature encoding, data balancing, and Gaussian noise augmentation to enhance input quality. Attention is then applied to emphasize important features, while the MLP learns and exploits complex relationships among them. Experiments on three cardiovascular disease datasets show that AB-MLP achieves accuracies of 98.90% on 4,240 samples, 98.79% on 920 samples, and 96.72% on 47,180 samples. These results highlight the effectiveness and generalizability of the proposed model, confirming its potential for intelligent healthcare diagnostics.

Keywords—Attention Mechanism, Multi-Layer Perceptron, Cardiovascular Disease Diagnosis, Data Preprocessing, Intelligent Healthcare.

I. INTRODUCTION

The healthcare sector has been significantly impacted by the rapid advancement of artificial intelligence (AI), particularly in the fields of disease diagnosis and clinical decision support. Traditional diagnostic methods often rely on manual analysis of clinical data. This approach is time-consuming, inconsistent, and prone to human error. As a result, machine learning models have been adopted to improve diagnostic accuracy and to reduce the burden on healthcare systems [1]. Among existing models, Multi-Layer Perceptron (MLP) has demonstrated stable performance on structured tabular data, which is commonly found in medical records. However, MLP typically processes each input feature independently, which limits their ability to capture complex relationships among clinical attributes such as those between resting blood pressure, cholesterol level, test results, medication history, and so on [2]. Attention mechanisms (AM), widely used in models based on Transformer architectures, offer an effective solution. These mechanisms assign importance weights to features, allowing the model to focus on the most relevant information in the data [3]. Medical data is often challenged by issues such as missing values, data loss, class imbalance, intra-class imbalance, noise, outliers, and heterogeneity. Therefore, data preprocessing steps must be performed before model training [4]. In this study, data preprocessing is first performed to enhance the quality of the input data. Next, an Attention mechanism is applied to identify important clinical features and assign appropriate weights, enabling the model to focus on factors that significantly impact diagnostic outcomes [3]. Finally, an MLP with fully connected layers is deployed to learn and exploit complex nonlinear relationships among features, thereby producing highly accurate disease diagnoses [4].

II. RELATED WORK

In recent years, advances in computer science have strongly promoted the application of intelligent solutions in medical diagnostics. Classical and simple models, such as logistic regression, decision trees, and support vector machines (SVM), have been widely applied to disease diagnosis tasks. These models are computationally efficient, simple, and resource-friendly. However, limitations are often encountered when handling high-dimensional medical data with complex feature interdependencies or nonlinear patterns [5]. Attention mechanisms have been extensively explored in combination with deep learning models to enhance diagnostic performance on medical data, particularly on tabular, nonlinear, unstructured data, and image-based skin disease diagnosis [6], [28]. Initially, the study on TabNet, applied attention mechanisms to tabular data. Sparse attention was employed to dynamically select important features for the target during training instead of using fixed feature sets. TabNet was demonstrated to outperform traditional MLP models and Random Forest (RF) on standardized datasets. However, challenges remain when dealing with incomplete and noisy medical data [7]. Subsequently, a study published in Neural Processing Letters proposed using the Real-Coded Genetic Algorithm (RCGA) to optimize MLP hyperparameters. By tuning parameters like learning rate, activation function, and regularization, the method balanced accuracy and complexity. Experiments on four University of California, Irvine (UCI) datasets showed improved performance and a simpler MLP structure. [8]. In the domain of omics data, which comprises large-scale biological studies such as genomics, transcriptomics, proteomics, and metabolomics, complex molecular systems are analyzed. Omics datasets are typically large, high-dimensional, and are increasingly used in disease prediction and personalized medicine. A study published in Frontiers in Genetics proposed the use of self-attention to model inter-feature relationships within genomic, transcriptomic, and proteomic data. The model effectively handled high-dimensional data with small sample sizes and supported biological interpretation, a critical requirement in biomedical data analysis [9]. For multimodal data, MedFuseNet, published in Scientific Reports, developed a framework that integrates medical images and tabular data using attention fusion layers. This study demonstrated that combining attention across multiple data modalities significantly improves diagnostic accuracy in tasks such as Visual Question Answering in Radiology (VQA-RAD) [10], [27]. Another approach involved integrating MLP with attention modules for medical image segmentation. A study published in IEEE Transactions on Medical Imaging demonstrated that inserting MLP-Attention blocks into U-Net significantly improved segmentation performance on CT and MRI images. The results confirmed that attention not only benefits unstructured data but also enhances MLP flexibility

in medical diagnostic tasks [11]. To address mortality prediction in ICU settings, an architecture model based on combining LSTM and attention was developed and evaluated using MIMIC-III, a large-scale public dataset containing clinical records of over 40,000 ICU patients. The study, published in IEEE Journal of Biomedical and Health Informatics, showed that attention mechanisms helped the model effectively learn critical clinical features and improved mortality prediction accuracy [12]. In the field of diabetes diagnosis, accurate and efficient screening is crucial for early intervention and complication prevention. A study published as Chapter 16 in the book Practical Statistical Learning and Data Science Methods (Springer) combined TabNet with Bayesian optimization. The study highlighted the practical effectiveness of attention-based models in handling noisy and incomplete Electronic Health Records (EHR) [13]. Similarly, the Transformer architecture laid the foundation for modern attention mechanisms. Although originally designed for natural language processing (NLP), Transformer has significantly influenced the design of models for tabular data and healthcare applications. It has inspired the development of strong attention-based models such as FT-Transformer and SAINT-Tab, which are specialized for tabular medical data. [14]. Finally, a study published in Nature Communications proposed the Attention-Augmented Neural Network (AANN) architecture for medical image diagnosis, demonstrating that combining attention and MLP outperformed conventional CNN or MLP only models [15]. Building upon these foundations, attention mechanisms were integrated into an MLP-based classifier to prioritize important features and model their interactions through fully connected layers. In addition to architectural design, the study also focused on addressing data-related challenges, including internal and external class imbalance and noise, aiming to develop a more accurate and practical diagnostic model. The integration of Attention Blocks into MLP is considered a promising direction aligned with current research trends and holds strong potential for practical applications in healthcare data analytics.

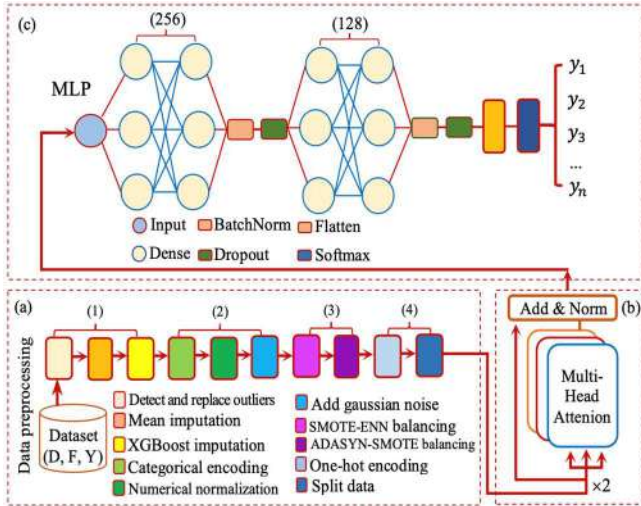


Fig. 1. Architecture for AB-MLP method

III. PROPOSED METHOD

This section presents the overall architecture of the proposed framework, which includes a data preprocessing pipeline, attention blocks used to extract and assess the importance of input features for assigning priority weights,

and an MLP model responsible for disease classification. Details in Fig. 1.

A. Data preprocessing pipeline

Data preprocessing block (a) Effective preprocessing is one of the key factors for achieving high performance in medical diagnosis tasks. To begin with, (1) the input dataset is represented as (D, F, Y) , where F denotes the features and Y represents the labels. An Isolation Forest is utilized to detect outliers, which are then marked as NaN. Missing data is initially handled through column wise mean imputation, followed by prediction and replacement using the XGBoost model to improve accuracy. (2) Categorical features are encoded, while numerical features are standardized using StandardScaler and augmented with Gaussian noise to reduce overfitting. (3) SMOTE-ENN is applied to address class imbalance; however, this method may cause the loss of some critical information and may not generate enough minority samples, leading to internal imbalance. As a result, ADASYN and SMOTE are further employed, where ADASYN focuses on hard to distinguish regions and SMOTE enhances sample diversity. (4) The output labels are converted to one hot encoding, after which the data is split into training and test sets, specifically:

Removing samples with missing or outlier data can cause information loss, dataset shrinkage, and bias. To avoid this, XGBoost Imputation is employed to replace missing values while preserving data integrity and model performance. This method leverages an XGBoost regression model to predict missing values based on complex feature relationships and operates iteratively until convergence or a predefined threshold is met [16].

Algorithm 1: Iterative boosted imputation

Input: $X \in \mathbb{R}^{m \times n}$, $f = \text{XBGRRegressor}$, T, ε
Output: $\hat{X}$

Procedure: Outlier Initialization

1. $f \leftarrow \text{IsolationForest}(\theta, r)$
2. $\ell \leftarrow f.\text{fit}(X)$
3. $X_{i,j} \leftarrow \text{NaN}, \forall (i, j) \mid \ell_{i,j} = -1$
4. $S_j \leftarrow \{i \mid x_{i,j} \neq \text{NaN}\}$
5. $\hat{x}_{i,j}^{(0)} \leftarrow \frac{\sum_{i \in S_j} x_{i,j}}{|S_j|}$

Procedure: Iterative imputation

6. **for** $t = 1, \dots, T$ **do**
7. $X_{\text{train}}, Y_{\text{train}} \leftarrow \{(C_{\text{features},i}, x_{i,j}) \mid x_{i,j} \neq \text{NaN}\}$
8. $f \leftarrow \mathcal{L}(X_{\text{train}}, Y_{\text{train}})$
9. $\hat{x}_{i,j}^{(t)} \leftarrow f(C_{\text{features},i}), \forall x_{i,j} = \text{NaN}$

Procedure: Convergence check

10. $\Delta = \|\hat{X}^{(t)} - \hat{X}^{(t-1)}\|$
11. **if** $\Delta < \varepsilon \Rightarrow \text{stop}$

return $\hat{X} = \hat{X}^{(t)}$

Data preparation and balancing pipelines. Data augmentation is the technique of generating new variants from original data, making the model more generalizable. In this case, we employ Gaussian noise as an augmentation method to simulate random fluctuations and enrich the data without altering its structure. This approach helps improve the ability to handle missing data [17].

Algorithm 2: Gaussian noise injection

Input: $X = [x_1, x_2, \dots, x_n], \mu, \eta$
Output: X'
1. $X' \leftarrow X$
2. $x'_i = x_i + \varepsilon_i, \forall x_i \in X', \varepsilon_i \sim \mathcal{N}(\mu, \eta^2)$
return X'

Balancing data. Combining SMOTE and ENN balances the data and improves dataset quality. SMOTE generates new data points to strengthen the minority class, reducing class imbalance. Subsequently, ENN removes incorrectly classified data points within their nearest neighbors, helping to clean and reduce noise. The next step employs ADASYN and SMOTE to rebalance the data. ADASYN creates synthetic samples for the minority class based on the degree of imbalance, prioritizing difficult to learn areas.

Algorithm 3: Balanced SMOTE-ENN

Input: X, y
Output: $X_{\text{clean}}, y_{\text{clean}}$
1. $X_{\text{synthetic}}, y_{\text{synthetic}} \leftarrow \text{SMOTE}(X, y)$
2. $X_{\text{clean}}, y_{\text{clean}} \leftarrow \text{ENN}(X_{\text{synthetic}}, y_{\text{synthetic}})$
return $X_{\text{clean}}, y_{\text{clean}}$

Finally, SMOTE performs interpolation to produce additional new samples, ensuring the data is fully balanced once imbalance issues are resolved [18].

Algorithm 4: ADASYN-SMOTE resampling

Input: X, y
Output: $X_{\text{final}}, y_{\text{final}}$
1. $X', y' \leftarrow \text{ADASYN.fit\_resample}(X, y)$
2. $X_{\text{final}}, y_{\text{final}} \leftarrow \text{SMOTE.fit\_resample}(X', y')$
return $X_{\text{final}}, y_{\text{final}}$

B. Attention-based feature representation

The attention block (b) in the proposed architecture is composed of two Multi-Head Attention (MHA) layers integrated with residual connections, layer normalization, and dropout. This structure is designed based on the Transformer Encoder and has been demonstrated to be effective in capturing complex representations. In each MHA layer, the input vector is divided into multiple parallel heads so that multi-dimensional feature interactions can be captured. Residual connections are employed to preserve the original input information and enhance gradient propagation. Layer normalization is applied to improve training stability. A dropout rate of 0.3 is used to mitigate overfitting, particularly in cases where the training dataset is limited or imbalanced. As a result, the model is enabled to focus on the most informative input features by assigning attention weights according to their contribution to the classification outcome [19, 20].

C. MLP classification head

Following the attention mechanism, the MLP block (c) is implemented using two fully connected layers with 256 and 128 units, respectively. Each layer is followed by batch normalization and dropout. The GELU activation function is applied to introduce non-linearity and to enhance the model's ability to learn complex patterns. L2 regularization is applied to both layers to prevent overfitting by penalizing large weight values. Batch normalization is employed after each dense layer to stabilize the training process and accelerate convergence. A dropout rate of 0.3 is applied to randomly

deactivate neurons during training, thereby improving the model's generalization ability. This block is designed to learn nonlinear relationships among input features through multiple fully connected layers. The Flatten layer is used to convert the multidimensional feature map into a one-dimensional vector. The output layer is implemented using a dense layer with softmax activation to convert the final feature vector into a probability distribution and to perform classification [21].

IV. EXPERIMENTAL

This section presents the experimental evaluation of the proposed solution, including data preprocessing, training scenarios, and model performance metrics based on the experimental results.

A. Dataset Description

The dataset used for heart disease diagnosis is as follows. The first dataset is designed to assess the severity of heart disease, including 920 samples and 16 features. Important features include chest pain type, resting blood pressure, cholesterol level, and fasting blood sugar. The target variable is the column called num, which takes values from 0, representing no disease, to 4, representing the most severe condition [22]. Meanwhile, the second dataset focuses on features including smoking status, daily cigarette consumption, history of hypertension, and body mass index, among others. This dataset contains 4,240 samples and includes 16 features for the diagnosis of disease or no disease. With important features such as glucose, systolic blood pressure, history of hypertension, and smoking status, this dataset effectively supports the diagnosis task [23]. The third dataset, referred to as CVDs, is a large-scale collection consisting of 47,180 samples and 17 features, with important features such as BMI, age category, physical activity, and general health status relevant to diabetes prediction. It is also used for the diagnosis of disease or no disease [24]. Details are provided in Table I.

TABLE I. HEART DISEASE DATASETS

Dataset Name	Features	Samples	Target variable
Heart Disease UCI (HD)	16	920	0: Very low risk
			1: Low risk
			2: Medium risk
			3: High risk
Framingham Heart Study (FHS)	16	4,240	0. No disease 1. Disease
Cardiovascular Diseases (CVDs)	17	47,180	0. No disease 1. Disease

Before model training, the data was preprocessed through the following stages: (1) missing values were handled using IterativeImputer with the XGBoost model, and categorical variables were encoded; (2) numerical features were standardized using StandardScaler and augmented with Gaussian noise to reduce overfitting; (3) outliers were detected and removed using the Isolation Forest algorithm; (4) class imbalance was addressed using SMOTE-ENN, ADASYN, and SMOTE. Finally, target labels were one-hot encoded, and the data was split into 80% training and 20% testing sets.

B. Model configuration

The model was trained through four main stages: (1) The Attention block was constructed using two stacked

MultiHeadAttention layers with eight heads and a key dimension of 64, combined with residual connections, layer normalization, and dropout, to determine the importance of each feature and assign appropriate weights; (2) An MLP block was composed of two Dense layers with 256 and 128 units, respectively, employing the GELU activation function, batch normalization, and dropout to reduce overfitting, allowing the model to learn relationships among features through full connections; (3) A Flatten layer was applied to transform the output into a one-dimensional vector, preparing the data for the final Dense layers; and (4) An output layer was implemented using a softmax activation function to predict class probabilities [26].

The model parameters were fine-tuned to optimize performance and ensure stable training. The parameters were adjusted as follows: the Adam optimizer was used with a small learning rate of 0.00005, and clipnorm was applied to limit the magnitude of gradients during training and to prevent instability caused by excessively large gradients, with the value set at 1.0. The categorical crossentropy loss function was selected for multi-class classification tasks, and accuracy was used as the evaluation metric to assess overall performance. EarlyStopping with a patience of 100 epochs was employed to prevent overfitting, while ReduceLROnPlateau was applied to automatically reduce the learning rate when the validation loss did not improve. Class imbalance was addressed by applying class weights, and a batch size of 128 was selected to balance training speed and hardware efficiency. Finally, model performance was evaluated based on loss and accuracy on the test set [28]. Detailed results are presented in Table II.

TABLE II. HYPERPARAMETERS USED FOR MODEL TRAINING

Name	Value
learning_rate	0.00005
optimizer	AdamW
weight_decay	0.01
kernel_initializer	he_normal
loss	categorical_crossentropy
metrics	accuracy
batch_size	32
epochs	300
patience	100
restore_best_weights	True
monitor	val_loss
clipnorm	1.0
dropout_rate	0.3 – 0.4
class_weighting	balanced

TABLE III. MODEL PARAMETER SUMMARY

Name	Value
Total params	2,279,941 (8.70 MB)
Trainable params	2,278,149 (8.69 MB)
Non-trainable params	1,792 (7.00 KB)

After compilation, the model was found to have 2,279,941 parameters, of which 2,278,149 parameters were trainable. This distribution allowed the model to learn complex features from the data, with strong processing capability and high performance enabled by the large number of trainable parameters. Detailed information is presented in Table III.

C. Evaluation metrics

To evaluate the performance of the model for the cardiovascular disease diagnosis task, six metrics are used: Accuracy (Acc), Recall (REC), Precision (PRE), Confusion matrix, and F1-score (F1). In addition, the loss function is also used for further evaluation. The formulas for these metrics and the loss function are presented as follows [27]:

$$ACC = \frac{\sum TP}{N_s} \quad (1)$$

$$REC = \frac{TP}{TP + FN} \quad (2)$$

$$PRE = \frac{TP}{TP + FP} \quad (3)$$

$$CM[i, j] = \left| k \in \{1, \dots, N \mid y_{true}^{(k)} = i \wedge y_{pred}^{(k)} = j\} \right| \quad (4)$$

$$F1 = \frac{2 \times REC \times PRE}{REC + PRE} \quad (5)$$

$$Loss = \sum_{i=1}^N \max(0, 1 - y_i \times \hat{y}_i) \quad (6)$$

D. Experimental results

The model was trained on the HD dataset containing 920 samples, with early stopping applied at epoch 1,271 and the best weights recorded at epoch 1,171. For the FHS dataset consisting of 4,240 samples, training was early stopped at epoch 915, with the best weights, which was obtained at epoch 815. The CVDs dataset, with 47,180 samples, was trained for 300 epochs, and the best weights were selected based on validation performance. The experimental results demonstrated that the proposed model achieved very high diagnostic performance on all datasets. Specifically, on the HD dataset, the model achieved an accuracy of 98.79%, an F1-score of 98.79%, a precision of 98.73%, and a recall of 98.49%. On the FHS dataset, the corresponding metrics were 98.90% for accuracy, 98.90% for F1-score, 98.91% for precision, and 98.90% for recall. On the large-scale CVDs dataset, the model achieved 96.72% for accuracy, F1-score, precision, and recall. The loss values on the HD, FHS, and CVDs datasets were 0.2001, 0.1255, and 0.0983, respectively, indicating good optimization and no significant overfitting. The experimental results are presented in Table IV.

TABLE IV. EVALUATION RESULTS ON DATASETS

Dataset	Samples	ACC	Loss	F1	PRE	REC
HD	920	98.79%	0.2001	98.79%	98.73%	98.49%
FHS	4,240	98.90%	0.1255	98.90%	98.91%	98.90%
CVDs	47,180	96.72%	0.0983	96.72%	96.72%	96.72%

The confusion matrix in Fig. 2 shows that on the HD dataset, the model correctly classified 100% of No Disease, Mild, and Very Severe samples, along with 97% of Moderate and 97% of Severe samples. Minor misclassifications occurred, specifically with 2% of Moderate samples predicted as Severe and 3% of Severe samples misclassified as Moderate. These errors were likely caused by the subtle differences between adjacent severity levels. On the FHS dataset, the model correctly classified 98% of No Disease samples and 100% of Disease samples, with only 2% of No Disease samples misclassified. For the CVDs dataset, the model achieved 97% accuracy for both No Disease and Disease classes, with 3% of each class incorrectly predicted as the other. These results demonstrate consistent and reliable classification performance across datasets of different sizes and label structures.

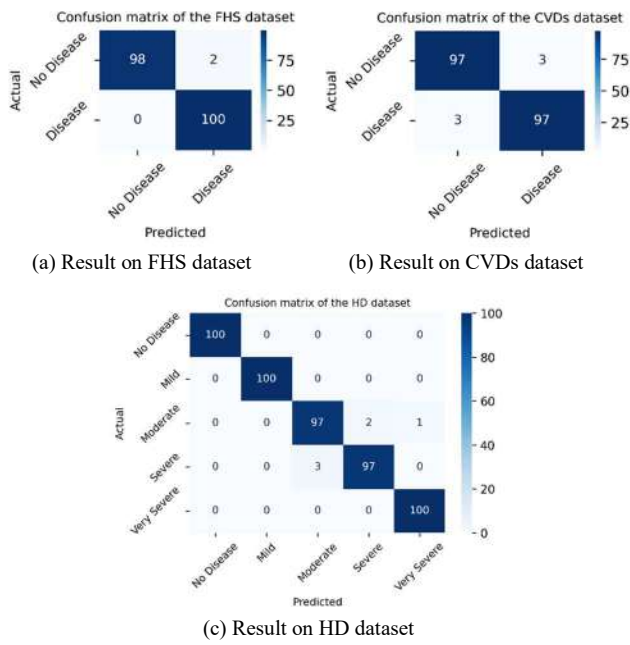


Fig. 2. Confusion matrix of accuracy performance

Observations from Fig. 3a, 3c, and 3e show that the training and validation accuracy curves on the FHS, HD, and CVDs datasets steadily increased and quickly converged after around 70 epochs, demonstrating stable learning throughout the training process. Meanwhile, the loss curves presented in Fig. 3b, 3d, and 3f decreased consistently and reached low values, with no noticeable gap between training and validation. This indicates that the model was well-optimized and did not suffer from overfitting, regardless of the differences in dataset scale.

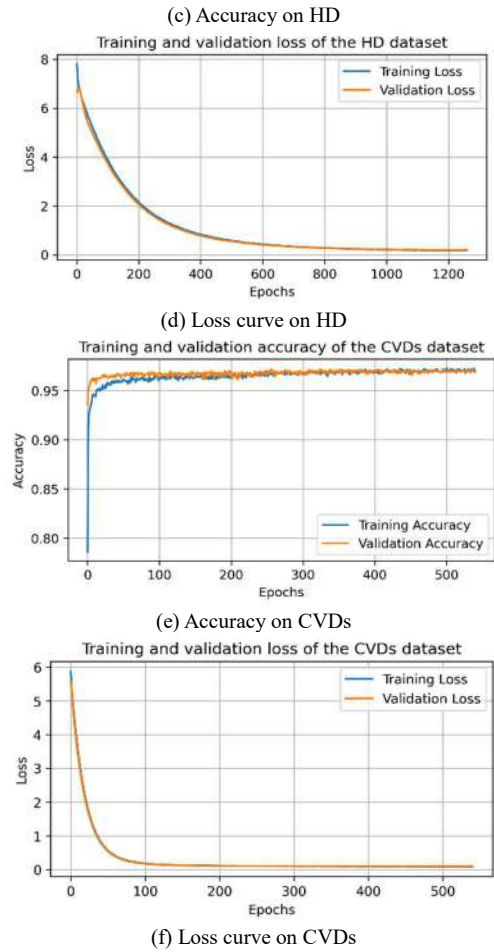


Fig. 3. Performance curves on FHS, HD and CVDs datasets

Although the model achieved impressive results, several minor issues should be discussed. First, on the HD dataset, minor misclassifications were observed between adjacent severity levels such as Moderate and Severe, due to the subtle differences between these levels, suggesting the need for deeper exploitation of clinical features to enhance discrimination. Second, the sample size on the HD dataset remained limited, with approximately 920 samples, which may affect the model's generalization capability. Third, the multi-class classification task presents potential risks of class imbalance, indicating that additional techniques such as focal loss or oversampling should be considered in future research.

V. CONCLUSION

The AB-MLP architecture was developed to enhance the accuracy of cardiovascular disease diagnosis. The data were preprocessed through outlier handling, missing-value imputation with XGBoost, feature encoding, class balancing, and Gaussian noise augmentation to ensure input quality. Attention was applied to emphasize clinically important features, followed by an MLP to capture nonlinear relationships and produce the final diagnosis. The model was evaluated on three cardiovascular datasets with varying sizes and diagnostic features. The first dataset, with 920 samples classified into five disease severity levels, achieved 98.79% accuracy and 0.2001 loss. The second dataset, with 4.240 samples for binary classification, achieved 98.90% accuracy and 0.1255 loss. The third large-scale dataset, containing 47.180 samples for binary classification, reached 96.72% accuracy and 0.0983 loss. Evaluation metrics ranged from

96.72% to 98.91% for F1-score, precision, and recall. The results confirm the high effectiveness and real-world applicability of the AB-MLP approach.

REFERENCES

- [1] L. Abualigah, S. A. Alomari, M. H. Almomani, R. A. Zitar, K. Saleem, H. Migdady, V. Snasel, A. Smerat, and A. E. Ezugwu, "Artificial intelligence-driven translational medicine: a machine learning framework for predicting disease outcomes and optimizing patient-centric care," *Journal of Translational Medicine*, vol. 23, no. 302, 2025, doi: 10.1186/s12967-025-06308-6.
- [2] C. J. Ejayi, D. Cai, F. O. Eze, M. B. Ejayi, J. E. Idoko, S. K. Asere, and T. U. Ejayi, "Polynomial-SHAP as a SMOTE alternative in conglomerate neural networks for realistic data augmentation in cardiovascular and breast cancer diagnosis," *Journal of Big Data*, vol. 12, no. 97, 2025, doi: 10.1186/s40537-025-01152-3.
- [3] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, pp. 331–368, 2022, doi: 10.1007/s41095-022-02711-y.
- [4] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review*, vol. 57, article no. 273, 2024, doi: 10.1007/s10462-024-10694-3.
- [5] S. Liu, H. Du, and M. Feng, "Robust Predictive Models in Clinical Data—Random Forest and Support Vector Machines," in *Leveraging Data Science for Global Health*, edited by L. Celi, M. Majumder, P. Ordóñez, J. Osorio, K. Paik, and M. Somai, Cham: Springer, 2020. doi: 10.1007/978-3-030-47994-7_13.
- [6] C.-Y. Lien, H.-J. Wang, and C.-K. Lu, *et al.*, "Deep Learning with an Attention Mechanism for Enhancing Automated Modified Total Sharp/van der Heijde Scoring of Hand X-ray Images in Rheumatoid Arthritis," *J. Med. Biol. Eng.*, 2025, doi: 10.1007/s40846-025-00947-2.
- [7] J. Fonseca and F. Bacao, "Tabular and latent space synthetic data generation: a literature review," *J. Big Data*, vol. 10, p. 115, 2023, doi: 10.1186/s40537-023-00792-7.
- [8] F. Z. El-Hassani, M. Amri, N.-E. Joudar, and K. Haddouch, "A new optimization model for MLP hyperparameter tuning: Modeling and resolution by real-coded genetic algorithm," *Neural Processing Letters*, 2024. DOI: 10.1007/s11063-024-11578-0.
- [9] J. L. Ballard, X. Wang, W. Li, L. Shen, and Q. Long, "Deep learning-based approaches for multi-omics data integration and analysis," *BioData Min.*, vol. 17, Art. no. 38, 2024, doi: 10.1186/s13040-024-00338-9.
- [10] X. Pei, K. Zuo, and Y. Li, *et al.*, "A Review of the Application of Multi-modal Deep Learning in Medicine: Bibliometrics and Future Directions," *Int. J. Comput. Intell. Syst.*, vol. 16, p. 44, 2023, doi: 10.1007/s44196-023-00225-6.
- [11] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "TransAttUnet: Multi-Level Attention-Guided U-Net With Transformer for Medical Image Segmentation," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 8, no. 1, pp. 55–68, Feb. 2024, doi: 10.1109/TETCI.2023.3309626.
- [12] F. Yang, J. Zhang, and W. Chen, *et al.*, "DeepMPM: a mortality risk prediction model using longitudinal EHR data," *BMC Bioinformatics*, vol. 23, p. 423, 2022, doi: 10.1186/s12859-022-04975-6.
- [13] O. O. Awe and J. M. Adepoju, "Optimizing Machine Learning Models for Disease Diagnosis Using Bayesian Hyperparameter Optimization (BHO)," in *Practical Statistical Learning and Data Science Methods. STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health*, O. O. Awe, A. Vance, and E. Somai, Eds. Cham: Springer, 2025, doi: 10.1007/978-3-031-72215-8_16.
- [14] Y. Kim, W. Choi, and W. Choi, *et al.*, "A machine learning approach using conditional normalizing flow to address extreme class imbalance problems in personal health records," *BioData Min.*, vol. 17, p. 14, 2024, doi: 10.1186/s13040-024-00366-0.
- [15] O. Olabanjo, A. Wusu, and O. Olabanjo, *et al.*, "A novel deep learning model for early diabetes risk prediction using attention-enhanced deep belief networks with highly imbalanced data," *Int. J. Inf. Technol.*, 2025, doi: 10.1007/s41870-025-02459-3.
- [16] N. Ma'muriyah, E. Noersasongko, P. Purwanto, S. Winarno, and M. I. Ashiddiq, "XG Boost based data imputation and outlier detection methods for classification of stunting," in *Proc. 2024 7th Int. Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2024, pp. 1–6, doi: 10.1109/ISRITI64779.2024.10963432.
- [17] G. Almeida and F. Bacao, "Counterfactual synthetic minority oversampling technique: solving healthcare's imbalanced learning challenge," *Data Science and Management*, vol. 5, 2025, Art. no. 100006, doi: 10.1016/j.dsm.2025.01.006.
- [18] A. Abdelkhalek and M. Mashaly, "Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning," *The Journal of Supercomputing*, vol. 79, pp. 10611–10644, 2023, doi: 10.1007/s11227-023-05139-2.
- [19] W. Lan, Y. Tao, B. Liang, R. Zhu, Y. Wei, and B. Shen, "Shear wave velocity prediction based on bayesian-optimized multi-head attention mechanism and CNN-BiLSTM," *Computers & Geosciences*, vol. 195, Art. no. 105787, Feb. 2025, doi: 10.1016/j.cageo.2024.105787.
- [20] Q. Wang, W. Ding, K. Khoshelham, and Y. Qiao, "Prediction of shield machine attitude parameters based on decomposition and multi-head attention mechanism," *Automation in Construction*, vol. 171, Art. no. 105973, Mar. 2025, doi: 10.1016/j.autcon.2025.105973.
- [21] J. Rynkiewicz, "Introduction to multilayer perceptron and hybrid hidden Markov models," in *Connectionist Approaches in Economics and Management Sciences*, C. Lesage and M. Cottrell, Eds. Boston, MA: Springer, 2003, vol. 6, pp. 49–69, doi: 10.1007/978-1-4757-3722-6_3.
- [22] M. N. Islam, K. R. Raiyan, S. Mitra, M. M. R. Mannan, T. Tasnim, A. O. Putul, and A. B. Mandol, "Predictis: an IoT and machine learning-based system to predict risk level of cardiovascular diseases," *BMC Health Services Research*, vol. 23, no. 171, 2023, doi: 10.1186/s12913-023-09104-4.
- [23] A. J. Albert, R. Murugan, and T. Sriprya, "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology," *Research on Biomedical Engineering*, vol. 39, pp. 99–113, 2023, doi: 10.1007/s42600-022-00253-9.
- [24] W. Li, M. E. Kalan, A. J. Kondracki, M. A. Akinyele, M. T. Cook, and O. E. Osundeko, "Poor sleep duration and E-cigarette/Cigarette use among US adults with cardiovascular diseases: findings from the 2022 BRFSS cross-sectional study," *Sleep and Breathing*, vol. 28, pp. 2701–2710, 2024, doi: 10.1007/s11325-024-03140-y.
- [25] S. S. Chowa, M. R. I. Bhuiyan, I. J. Payel, A. Karim, I. U. Khan, S. Montaha, M. Z. Hasan, M. Jonkman, and S. Azam, "A low complexity efficient deep learning model for automated retinal disease diagnosis," *Journal of Healthcare Informatics Research*, vol. 9, pp. 1–40, 2025, doi: 10.1007/s41666-024-00182-5.
- [26] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Computational Statistics*, vol. 39, pp. 363–383, 2024, doi: 10.1007/s00180-022-01301-9.
- [27] H. Dalianis, "Evaluation metrics and evaluation," in *Clinical Text Mining*, Cham: Springer, 2018, pp. 45–53, doi: 10.1007/978-3-319-78503-5_6.
- [28] B.-D. Nguyen, V.-D. Hoang, and H. D. Nguyen, "Improved Skin Classification Performance Based on Hybrid Models," in *Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2025*, N. T. Nguyen *et al.*, Eds. Communications in Computer and Information Science, vol. 2494. Singapore: Springer, 2025, pp. 89–104, doi: 10.1007/978-981-96-5884-8_7.
- [29] V. T. Hoang, D. Q. Nguyen, M. P. Tu, K.-H. Jo, and V.-D. Hoang, "A compact version of EfficientNet for skin disease diagnosis application," *Neurocomputing*, vol. 620, p. 129166, Mar. 2025, doi: 10.1016/j.neucom.2024.129166.

Emotionally Expressive Vietnamese Text-to-Speech Synthesis Using Variational Inference Models

Dung C. Hoang^{1,2}, Binh Nguyen^{1,2}, Xuan Hau Pham³, Dung A. Tran^{4,2}, Hien D. Nguyen^{4,2*}

¹*Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam*

²*Vietnam National University, Ho Chi Minh City, Vietnam*

³*Quang Binh University, Vietnam*

⁴*University of Information Technology, Ho Chi Minh City, Vietnam*

Email: dung.hoang300699@gmail.com, ngtbinh@hcmus.edu.vn, pxhauqbu@gmail.com, dungta@uit.edu.vn, hiennd@uit.edu.vn

Corresponding author: Hien D. Nguyen (hiennd@uit.edu.vn)

Abstract—This study explores Vietnamese Text-to-Speech (TTS) with a focus on emotional expressiveness. It involves integrating emotional embedding vectors, derived from a meticulously fine-tuned Wav2Vec2-Large-Robust model, directly into the Variational Inference Text-to-Speech (VITS) architecture for Vietnamese TTS, which is a previously unreported approach for Vietnamese TTS. The experimental results demonstrate that VITS significantly outperforms traditional two-stage models and exhibits strong competitive performance against leading commercial systems in natural speech quality. VITS synthesized speech with five distinct emotional nuances (neutral, happy, sad, angry, surprised), achieving remarkably higher Mean Opinion Score (MOS) ratings compared to commercial systems in emotional expressiveness. This study underscores the substantial potential of the end-to-end VITS architecture for generating high-quality, naturally expressive Vietnamese speech, thereby opening up new practical applications in human-machine interaction.

Index Terms—Text-to-Speech (TTS), VITS, Emotional expressiveness, Emotional embedding vectors, Wav2Vec2-Large-Robust, Mean Opinion Score (MOS), Speech synthesis quality

I. INTRODUCTION

Text-to-Speech (TTS) converts text into spoken audio and is essential for virtual assistants, navigation, education, and audiobooks[1], [2]. High-quality TTS requires accurate pronunciation plus natural prosody, intonation, and clear emotional nuance. Vietnamese is especially challenging. It is largely monosyllabic and tonal (six lexical tones), so tone changes can alter meaning; wide regional accents and a shortage of large, well-labeled emotional corpora further complicate expressive synthesis.

Traditional concatenative and parametric systems were intelligible but often unnatural. Neural pipelines (such as Tacotron 2 [3], Glow-TTS [4] + HiFi-GAN [5]) produce much more natural speech but typically rely on multi-stage text → spectrogram → waveform flows that can accumulate errors. VITS [6], a flow-based end-to-end model, removes intermediate stages and jointly models linguistic content, prosody, and duration, reducing error propagation and improving smoothness.

In this work we fine-tune three Vietnamese systems (Tacotron2 + HiFi-GAN, Glow-TTS + HiFi-GAN, and VITS) and extend VITS with continuous emotion conditioning from a fine-tuned Wav2Vec2-Large-Robust encoder [7]. Emotion

embeddings are projected and fused with linguistic representations to condition duration prediction and waveform generation, enabling synthesis of sad, happy, neutral, angry, and surprised states. Evaluations (MOS, WER, robustness) against the pipeline baselines and commercial Vietnamese TTS services (FPT, Viettel, Vbee) show the emotion-conditioned VITS achieves competitive naturalness and substantially better expressive control [8], outperforming tested commercial voices and setting a new benchmark for emotionally rich Vietnamese speech synthesis.

The remainder of the paper is organized as follows. Section II reviews related work on traditional and neural TTS, prior Vietnamese TTS efforts, and methods for incorporating emotion into speech synthesis. Section III details the proposed methodology, including system architecture, dataset collection and preprocessing, emotion embedding extraction and fusion, and training configuration. Section IV presents experimental results, including subjective listening tests, automatic transcription metrics, robustness analysis, and comparisons with commercial systems. Section V discusses limitations, potential dataset biases, ethical considerations such as misuse risks, and proposed mitigation strategies. Concluding remarks and directions for future research appear in the final section.

II. RELATED WORKS

A. Traditional and Neural Vietnamese TTS

Early Vietnamese TTS systems used concatenative and parametric methods. Concatenative synthesis joined small recorded audio segments to form speech, which sounded fairly natural but lacked flexibility in tone and rhythm and required a large, well-organized dataset [9]. Parametric methods (such as HMM-based models) offered more control over voice features but often produced robotic and unnatural speech, especially with Vietnamese’s six tones and regional accent differences.

The introduction of neural TTS greatly improved voice quality. Models like Tacotron 2 with HiFi-GAN could generate speech that sounds almost human, while Glow-TTS + HiFi-GAN made the process faster and more stable by replacing attention-based [10] [11] decoding with a flow-based approach. However, these pipeline systems still rely on two separate stages (spectrogram generation and waveform

synthesis), which can limit how well they capture emotional and expressive nuances in Vietnamese speech.

B. End-to-End VITS and Emotion Integration

VITS (Variational Inference Text-to-Speech) unifies spectrogram prediction and vocoding into a single end-to-end model that maps phonemes directly to waveforms. It jointly learns prosody [12] and duration through a flow-based vocoder, reducing alignment errors, lowering model complexity, and producing more natural, expressive speech [20].

While VITS has shown strong results in other languages, its ability to capture emotional variation in Vietnamese speech has not been fully explored. Most prior emotional TTS work (mainly in English and Mandarin) relies on discrete emotion labels or style tokens [21], which limit flexibility. Continuous emotion embeddings (learned directly from speech) offer greater expressiveness but require robust emotion extraction models. Vietnamese TTS systems, which are integrated such continuous emotional embeddings within an end-to-end framework, are limited.

C. Vietnamese Commercial TTS Systems

In recent years, Vietnamese tech companies such as FPT.AI, Zalo AI, and Viettel AI have released commercial TTS APIs offering multiple voices and basic prosody control. However, their reported Mean Opinion Scores (MOS) [13] range only from 3.8–4.2 for neutral speech, and they lack fine-grained emotional expressiveness, making them less engaging for emotion-rich applications like audiobooks or virtual agents [14], [15]. To address this, this study adapts VITS for Vietnamese and enhances it with continuous emotion embeddings from a fine-tuned Wav2Vec2-Large-Robust model. By jointly learning semantic, prosodic, and emotional features within a unified end-to-end framework, the proposed system achieves higher MOS scores than existing commercial TTS solutions, even in expressive, non-neutral speaking styles.

III. VITS ARCHITECTURE WITH CONTINUOUS EMOTION INTEGRATION

A novel methodology is presented for emotionally expressive Vietnamese Text-to-Speech (TTS), extending state-of-the-art techniques to address the unique challenges of the Vietnamese language. The core of the method adapts and enhances the VITS model (a single, unified end-to-end architecture) by integrating continuous emotion embedding vectors. Departing from traditional multi-stage pipelines, the described approach constitutes a previously unreported solution for Vietnamese TTS.

A. Overall Architecture

At the heart of the system lies VITS, an end-to-end model that skips the usual intermediate steps (such as generating mel-spectrograms) and goes straight from text to raw audio waveforms. By eliminating these stages, VITS avoids the error accumulation inherent in multi-stage pipelines.

First, the input text is transformed into a semantic embedding by the Text Encoder. This embedding captures the linguistic content and context needed for natural speech synthesis.

Next, the Posterior Encoder learns the distribution $q_\phi(z|x)$ over latent variables z given the true audio spectrogram x . This encoder is responsible for extracting stochastic features such as rhythm and expressiveness from the real audio signal.

In parallel, the Prior Encoder models the distribution $p_\phi(z|c)$ conditioned on the text c . By employing normalizing flows [16], it enriches the expressivity of the latent space, ensuring that the generated speech can convey nuanced prosody.

These latent representations are then combined in the Decoder, which uses a series of invertible, flow-based transformations to convert z into the final waveform y . This design guarantees that the output is natural, smooth, and accurate.

During training, a Discriminator is employed in an adversarial setup to distinguish real audio from synthesized waveforms, pushing the Generator (i.e., the Decoder) to produce ever more realistic speech. Finally, the Stochastic Duration Predictor treats the duration of each text unit as a random variable, learning to reproduce natural timing variations and flexible rhythm in the generated audio.

B. Data Collection and Preprocessing

Two complementary speech corpora were combined to construct a robust, emotionally expressive Vietnamese TTS dataset. The VLSP Emotional Speech Dataset served as the primary source, containing 8,351 clips from a single professional speaker. Each clip carries one of five emotion labels—neutral (2,100 clips), happy (1,600), sad (1,700), angry (1,450), and surprised (1,501). A balanced test set was created by reserving 100 examples per emotion.

To increase vocal variability, a supplementary multi-speaker corpus of 3,500 recordings was also compiled, covering the same five emotion categories. This additional collection was adopted as the emotional-speech benchmark for the VLSP 2025 competition [17] for evaluation.

A uniform preprocessing pipeline was applied to all recordings. Steps included text normalization (standardizing punctuation, numerals, and formatting), conversion of cleaned text into phonemes with tone annotations to preserve Vietnam’s six lexical tones, spectral gating for background-noise reduction, and resampling to 22,050 Hz. A final quality-control stage computed spectrogram statistics (e.g., spectral flatness and signal-to-noise ratio) to detect and flag problematic samples for re-recording.

Although the combined corpora form a practical foundation for an emotionally conditioned Vietnamese TTS system, the dataset composition imposes limits on generalisability. The primary VLSP set contains a large number of utterances for a single speaker, which may bias learned prosodic and timbral patterns toward that speaker’s voice. The supplementary multi-speaker corpus increases diversity but remains limited in demographic and dialectal coverage (age, regional accents, and speaking styles), and the five emotion categories are neither exhaustive nor uniformly represented. These factors can bias

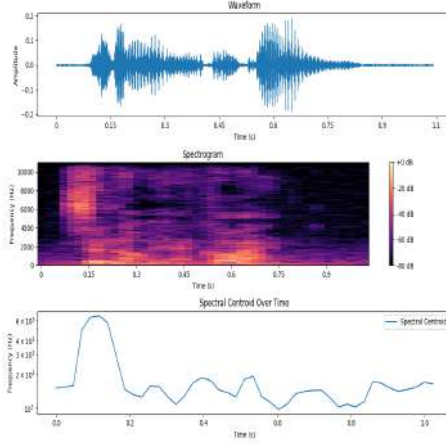


Fig. 1. Spectral representation of the audio segment.

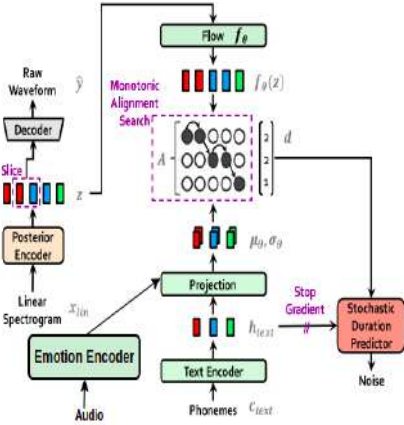


Fig. 2. Emotional integration into the VITS model.

the model toward more frequent emotions and constrain its ability to capture fine-grained or mixed affective cues.

C. Emotional Embedding Extraction and Integration

The key contribution is integrating continuous emotion embeddings directly into the end-to-end VITS pipeline, replacing discrete labels with high-dimensional vectors that capture gradients of intensity and subtle affective cues. Emotion vectors are extracted from audio using a Wav2Vec2-Large-Robust encoder fine-tuned on MSP-Podcast, so they encode pitch contours, intensity dynamics and temporal rhythm without manual annotation.

An Emotion Projection layer aligns dimensions ($1024 \rightarrow 192$) so the projected emotion vector can be fused with the VITS text encoder’s 192-dim semantic embedding. A small MLP learns a non-linear fusion into a single Synthesis Embedding.

This unified embedding conditions the stochastic duration predictor and flow-based decoder, guiding waveform generation to preserve linguistic content while expressing targeted emotions. Compared with multi-stage pipelines and many

commercial systems, direct conditioning on learned continuous embeddings yields finer emotional control, smoother prosody, and more natural expressive speech.

D. Training Configuration and Parameters

All models were implemented in PyTorch and trained on an Intel Xeon Gold 5220 CPU (2.20 GHz), NVIDIA RTX 3090 GPU, and 24 GB RAM.

Tacotron 2 used a 512-dim text embedding, a convolution[18]–GRU[19] encoder, and an LSTM[22] decoder ($hidden = 1024$) with location-sensitive attention, trained with Mean Squared Error (MSE) loss and the Adam optimizer (learning rate = 10^{-3} , exponential decay).

GlowTTS is a non-autoregressive model designed to directly learn the distribution of mel-spectrograms from text. Its Text Encoder used a Transformer-based architecture comprising 6 self-attention layers. The Flow-based Decoder utilized normalizing flow to map from Gaussian space to mel-spectrogram space.

HiFi-GAN vocoder used 80-band mel-spectrograms ($FFT = 1024$, $hop = 256$) and AdamW (learning rate = 2×10^{-4} , $\beta_1 = 0.8$, $\beta_2 = 0.99$, weight decay $\lambda = 0.01$) with combined GAN[23], Mel-Spectrogram, and Feature-Matching losses.

VITS was trained end-to-end with AdamW (same settings, learning rate = 2×10^{-4} , decay = $0.999^{1/8}$), using 80-band mels and windowed generator training ($window = 32$) for faster convergence and lower memory use.

E. Evaluation Metrics

Evaluation of the TTS systems employed both subjective listening tests and objective measurements. For subjective assessment, 10 native Vietnamese speakers rated 100 randomly selected test clips on a 1–5 Mean Opinion Score (MOS) scale, considering clarity, naturalness, pleasantness, and appropriateness of emotional expression.

Objective evaluation used Word Error Rate (WER), calculated by transcribing synthesized audio with a pre-trained ASR model and comparing the output with reference transcripts. Lower counts of insertions, deletions, and substitutions indicated higher intelligibility. The combination of human ratings and quantitative metrics provided a comprehensive characterization of system performance, highlighting both strengths and potential areas for improvement.

IV. EXPERIMENTAL RESULTS

This section presents both subjective and objective evaluations of Vietnamese Text-to-Speech (TTS) models. Three deep learning systems (Tacotron 2 + HiFi-GAN, Glow-TTS + HiFi-GAN, and VITS) are compared alongside three leading commercial Vietnamese TTS services: FPT, Viettel, and Vbee. The assessment also includes an evaluation of the emotional expressiveness achieved by the emotion-integrated VITS.

A. Naturalness and Speech Quality (MOS Evaluation)

A Mean Opinion Score (MOS) test was conducted to assess the perceived naturalness of synthesized speech. Native Vietnamese speakers rated 100 utterances from each system on a 5-point scale. Evaluation results for the baseline TTS models producing neutral speech are presented in Table I.

TABLE I
MOS RATINGS FOR TTS MODELS (NEUTRAL SPEECH)

Model	Average MOS	Min MOS	Max MOS
Ground Truth	4.72	4.63	4.80
Tacotron 2 + HiFi-GAN	4.27	4.25	4.33
Glow-TTS + HiFi-GAN	4.33	4.26	4.44
VITS	4.48	4.45	4.55

As shown in the table, the VITS model achieved the highest average MOS of 4.48. This score surpassed both the Tacotron 2 + HiFi-GAN model (4.27) and the Glow-TTS + HiFi-GAN model (4.33). The VITS model's performance was also close to the quality of the Ground Truth human recordings, which had an average MOS of 4.72. The fully end-to-end architecture of VITS effectively eliminates errors that can be introduced by intermediate representations like mel-spectrograms. This results in a smoother and more natural output.

B. Word Error Rate (WER)

Intelligibility was measured using Word Error Rate (WER) by transcribing synthesized audio and comparing the output with the original text. All systems demonstrated strong alignment with the reference transcripts, with VITS achieving a perfect score.

TABLE II
WORD ERROR RATE ACROSS TTS SYSTEMS

Model	Average WER
Ground Truth	0.00
Tacotron 2 + HiFi-GAN	0.01
Glow-TTS + HiFi-GAN	0.01
VITS	0.00

As Table II, the VITS model and the Ground Truth audio both achieved a perfect average WER of 0.00. This result indicates that VITS can synthesize speech with flawless content accuracy, successfully reproducing the original text without any word transcription errors.

C. Robustness to Variation

Robustness was assessed for each model under varying acoustic and linguistic input conditions, with VITS maintaining high consistency across evaluations.

TABLE III
ROBUSTNESS SCORES OF TTS MODELS

Model	Average Robustness	Min	Max
Ground Truth	5.0	5.0	5.0
Tacotron 2 + HiFi-GAN	4.0	3.9	4.1
Glow-TTS + HiFi-GAN	4.3	4.2	4.4
VITS	4.5	4.4	4.6

As shown in Table III, VITS achieved the highest average robustness score of 4.5, showing superior resilience to various sentence structures and phoneme distributions, which makes it more reliable for real-world use. The Glow-TTS + HiFi-GAN model followed with a score of 4.3, and Tacotron 2 + HiFi-GAN had the lowest score at 4.0. Ground Truth human recordings served as the benchmark with a perfect score of 5.0.

D. Comparison with Commercial Vietnamese TTS Systems

In a separate evaluation comparing VITS with Vietnamese commercial TTS systems (FPT, Viettel, Vbee), listeners assessed both naturalness and, critically, emotional expressiveness. For neutral speech evaluation, the results are presented in Table IV.

TABLE IV
MOS RATINGS FOR VITS AND COMMERCIAL MODELS (NEUTRAL SPEECH)

Model	Average Robustness	Min	Max
VITS	4.47	4.40	4.53
FPT	4.44	4.38	4.5
Viettel	4.43	4.38	4.47
Vbee	4.44	4.40	4.50

From the data in Table IV, VITS achieved the highest average MOS (4.47), slightly surpassing commercial models. This is particularly significant given VITS is an open-source model developed with limited academic resources, unlike commercial systems that benefit from substantial investment and continuous optimization on extensive proprietary datasets. This indicates the effectiveness of VITS's end-to-end architecture in reducing errors and enhancing speech naturalness.

However, differences in robustness were evident (Table V). VITS scored an average robustness of 4.41, slightly higher than FPT (4.36), Vbee (4.39) and Viettel (4.31). This suggests that VITS has higher stability and flexibility in handling diverse input data, which is crucial for practical applications.

TABLE V
ROBUSTNESS SCORES OF TTS MODELS

Model	Average Robustness	Min	Max
FPT	4.36	4.30	4.40
Viettel	4.31	4.25	4.35
Vbee	4.39	4.33	4.43
VITS	4.41	4.35	4.45

The next experiment focuses on emotional expressiveness, using a test set of 100 audio segments (20 for each of five emotions: neutral, happy, sad, angry, surprised) from the Emotional Speech Dataset. Ten independent evaluators assessed clarity, naturalness, and emotional expression using the MOS scale.

In Table VI, Ground Truth audio achieved an average MOS of 4.65, demonstrating excellent quality and natural emotional conveyance. Our VITS model, while not matching ground truth, still achieved an impressive average MOS of 4.31, categorized as "Good". This confirms VITS's capability to

TABLE VI
MOS RATINGS FOR EMOTION-INTEGRATED TTS MODELS

Model	Average Robustness	Min	Max
Ground Truth	4.65	4.60	4.70
VITS	4.31	4.25	4.45
FPT	3.00	3.00	3.00
Viettel	3.00	3.00	3.00
Vbee	3.00	3.00	3.00

produce relatively clear and natural audio, and, more importantly, to convey emotional nuances. This success validates the integration of semantic and emotional features within VITS’s unified architecture, overcoming limitations often faced by traditional pipeline methods. In stark contrast, the commercial models from FPT, Viettel, and Vbee uniformly scored an average MOS of 3.0, categorized as “Fair”. This indicates that while their output is technically audible, it often lacks naturalness and emotional depth, impacting user experience.

E. Comparative Evaluation

The performance trends presented in the previous tables are visually summarized in Figures 3–5.

Figure 3 compares the three research TTS architectures—Tacotron 2 + HiFi-GAN, Glow-TTS + HiFi-GAN, and VITS—on both Average MOS and Average Robustness scores. Consistent with the MOS and robustness tables, VITS achieved the highest scores in both metrics, with an average MOS of 4.48 and an average robustness of 4.50. In contrast, Glow-TTS + HiFi-GAN scored 4.33 (MOS) and 4.30 (robustness), while Tacotron 2 + HiFi-GAN trailed at 4.27 (MOS) and 4.00 (robustness). These results reinforce the conclusion from Section IV-A and IV-C that VITS’s unified end-to-end architecture delivers superior naturalness and stability over two-stage pipeline models.

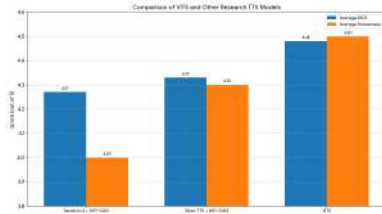


Fig. 3. Comparison of VITS and other research TTS models.

Figure 4 compares VITS with three leading Vietnamese commercial TTS systems—FPT, Viettel, and Vbee—using the same metrics for neutral speech. As detailed in Section IV-D, VITS achieved the highest average MOS (4.47), slightly exceeding the commercial systems (4.43–4.44), and also led in robustness with a score of 4.41, surpassing FPT (4.36), Vbee (4.39), and Viettel (4.31). These results indicate that, even with limited academic resources, the model matches or outperforms heavily optimized commercial platforms in both perceived naturalness and operational stability.

Figure 5 presents the results of the emotional expressiveness evaluation described in Section IV-E. The emotion-integrated

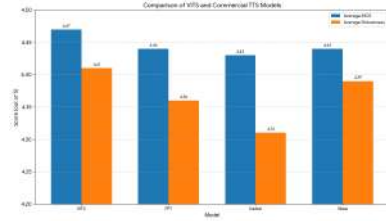


Fig. 4. Comparison of VITS and commercial TTS models.

VITS achieved an average MOS of 4.31, significantly surpassing all three commercial systems. Although ground-truth human recordings remained the highest standard, the substantial gap between VITS and the commercial baselines demonstrates the effectiveness of emotion embedding integration in producing expressive, emotionally nuanced Vietnamese speech.

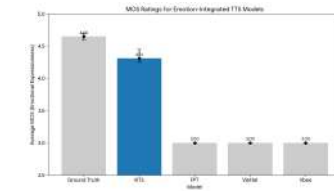


Fig. 5. MOS ratings for emotion-integrated TTS models.

F. Statistical significance testing

This section outlines the evaluation methods and results for MOS, WER, and robustness. MOS ratings were analyzed using a linear mixed-effects (LME) model with the TTS system as a fixed effect and listener, utterance as random effects:

$$MOS \sim system + (1|listener) + (1|utterance)$$

The LME separates listener and utterance variability, yielding valid standard errors and pairwise contrasts. When model assumptions were violated, non-parametric tests (Friedman and Wilcoxon with Holm/Bonferroni correction) were applied, reporting p-values, effect sizes, and 95% confidence intervals.

For neutral speech, mean MOS values were VITS ≈ 4.48 , Glow-TTS ≈ 4.33 , and Tacotron2 ≈ 4.27 , which showing VITS ahead by about 0.15–0.21 points, likely statistically significant though variance data were unavailable. Differences between VITS and commercial systems were minimal (≈ 0.03 – 0.04), suggesting limited practical significance even if statistically valid.

In contrast, emotion-conditioned VITS achieved ≈ 4.31 MOS versus ≈ 3.00 for commercial systems, a large (> 1.0 MOS on the 1–5 scale) and clearly significant gap, indicating major improvements in emotional expressiveness.

V. DISCUSSION & ETHICAL CONSIDERATIONS

Despite strong results, practical limitations remain. Training used only common accents (Northern and Southern) in the dataset, so regional accents, children’s and elderly voices, and multi-dialect scenarios are not covered; future work

should collect larger, demographically diverse, emotionally labeled multi-speaker data. Continuous emotion embeddings improve expressiveness but can amplify dataset biases [24] if some emotions or recording conditions are underrepresented—systematic evaluation across demographic groups and dataset balancing are needed.

Emotionally persuasive TTS also poses misuse risks (deepfake scams, disinformation, impersonation). Mitigations include embedding inaudible digital watermarks for provenance, and enforcing production safeguards such as API authentication, rate limits, usage monitoring (such as mass generation or rapid emotion switching), and clear forbidden-use policies with enforcement.

VI. CONCLUSION AND FUTURE WORK

This study designed emotionally expressive Vietnamese Text-to-Speech system, called VITS. Comparing with other neural architectures baselines, VITS consistently achieved the best performance, demonstrating superior naturalness, robustness, and expressiveness. In neutral-speech evaluations, VITS reached an average MOS of 4.48, surpassing other baselines and remaining competitive with leading commercial systems (FPT, Viettel, Vbee). For emotion-conditioned synthesis, the emotion-integrated VITS achieved an average emotion MOS of 4.31, outperforming commercial counterparts and confirming its effectiveness in expressive speech generation.

In the future, newer TTS architectures, such as F5-TTS, StyleTTS, and diffusion-based models, will be explored and benchmarked against VITS to advance audio quality, synthesis speed, and fine-grained style control. Besides, emotion encoding can be improved by adopting advanced extractors and finetuning them on Vietnamese emotional corpora to capture more subtle affective cues. Adding real-time emotion controls and personalization features would let users adjust the mood and speaking rate as the voice is being created. This would turn the system into an interactive, customizable voice platform.

ACKNOWLEDGMENT

This research was supported by The Ho Chi Minh University of Education’s Scientific Research Support Fund.

REFERENCES

- [1] V. Reddy, T. Vaishnavi, P. Kumar Speech-to-text and text-to-speech recognition using deep learning. In *Proceedings of 2nd international conference on edge computing and applications (ICECAA 2023)*, pp. 657-666 (2023).
- [2] Ta, P., et al. Detecting signs of depression on social media: A machine learning analysis and evaluation. *Sustainable Futures*, vol. 10, 100827 (2025).
- [3] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyriannakis, Yonghui Wu Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018).
- [4] Jaehyeon Kim, Sungwon Kim, Jungil Kong, Sungroh Yoon Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)* (2020).
- [5] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)* (2020).
- [6] Jaehyeon Kim, Jungil Kong, Juhee Son Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech In *Proceedings of International Conference on Machine Learning (ICML)* (2021).
- [7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)* (2020).
- [8] Anh Huynh, Tung Nguyen, Thu Nguyen, et al. A Method of Deep Reinforcement Learning for Simulation of Autonomous Vehicle Control In *Proceedings of 16th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2021)*, pp. 372–379 (2021).
- [9] Z. Amiri, A. Heidari, N.J. Navimipour, et al. Adventures in data analysis: A systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems. *Multimedia Tools and Applications*, 83(8), pp. 22909-22973 (2024).
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin Attention Is All You Need In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA (2017).
- [11] Minh-Thang Luong, Hieu Pham, Christopher D. Manning Effective Approaches to Attention-based Neural Machine Translation In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal (2015).
- [12] Hyung-Seok Oh, Sang-Hoon Lee, and Seong-Whan Lee DiffProsody: Diffusion-based Latent Prosody Generation for Expressive Speech Synthesis with Prosody Conditional Adversarial Training *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 2654 - 2666 (2024).
- [13] Streijl, Robert C. and Winkler, Stefan and Hands, David S. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives *Multimedia Systems*, Vol. 22, Iss. 2, pp. 213 - 227 (2016).
- [14] M. Huang, J. Qiang, Y. Zhu, et al. AI4Reading: Chinese Audiobook Interpretation System Based on Multi-Agent Collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 3: System Demonstrations, pp. 211-220 (2025).
- [15] Q. Vu, V. Pham, D. Dinh, et al. Multi-Agents System for Game Development Through Unreal Engine. In *Proceedings of 22nd International Conference on Intelligent Software Methodologies, Tools and Techniques (SOMET 2023)*, Napoli, Italy, pp. 242-252 (2023).
- [16] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker Normalizing Flows: An Introduction and Review of Current Methods *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 3964-3979 (2021).
- [17] Cao Manh Hai, Hoang Chi Dung, Le Quang Trung, Do Van Hai VLSP 2025 Automatic Speech Recognition and Speech Emotion Recognition (2025). doi: VLSP 2025 website
- [18] Keiron O’Shea, Ryan Nash An Introduction to Convolutional Neural Networks *arXiv* (2015). doi: 10.48550/arXiv.1511.08458
- [19] F.A. Gers; J. Schmidhuber; F. Cummins Learning to forget: continual prediction with LSTM. In *Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN 99)* Edinburgh, UK, Vol. 2, pp. 850-855 (1999).
- [20] S. Hoang; B. Nguyen; N. Nguyen; et al. Enhanced task-based knowledge for lexicon-based approach in vietnamese hate speech detection. In *Proceedings of 14th International conference on knowledge and systems engineering (KSE 2022)*, Nha Trang, Vietnam (2022).
- [21] T. Li, C. Hu, J. Cong, et al. Diclet-tts: Diffusion model based cross-lingual emotion transfer for text-to-speech—a study between english and mandarin. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3418-3430 (2023).
- [22] S. Hochreiter, J. Schmidhuber Long Short-Term Memory. *Neural computation*, 9(8), pp. 1735 - 1780 (1997).
- [23] Ian J. Goodfellow, et al. Generative Adversarial Networks *Communications of the ACM*, 63(11), pp. 139 - 144 (2020).
- [24] H. Nguyen, T. Huynh, S. Hoang, et al. Language-oriented Sentiment Analysis based on the Grammar Structure and Improved Self-attention Network. In *Proceedings of 15th International Conference on Evaluation of Novel Approaches to Software Engineerin (ENASE 2020)*, pp. 339-346 (2020).

Vehicle Re-Identification for Comprehensive Multi-Camera Tracking in Vietnamese Urban Infrastructure

Kim-Phat Ma*, Mong-Thuy Duong-Thi*, Chi-Nguyen Dang*,
Anh Thi-Hoang Nguyen*, Trong-Hop Do†

* Faculty of Information Science and Engineering, University of Information Technology, VNU-HCM, Vietnam

† Faculty of Software Engineering, University of Information Technology, VNU-HCM, Vietnam

Emails: {22521071, 22521454, 22520963}@gm.uit.edu.vn; {anhnth, hopdt}@uit.edu.vn

Abstract—Vehicle re-identification (Re-ID) is a key component of intelligent transportation systems, enabling continuous tracking of vehicles across non-overlapping camera views. Most existing models are trained on curated benchmarks and often fail in real-world traffic with high density, heterogeneous vehicle types, occlusion, and challenging imaging conditions.

This work introduces a real-time multi-camera tracking framework tailored for Vietnamese urban traffic. The pipeline integrates a YOLO-based detector with Deep SORT for intra-camera tracking and employs a state-of-the-art backbone (OSNet) with a novel cross-camera association algorithm to maintain vehicle identity across disjoint views. The system is deployed on Apache Spark and Apache Kafka, supporting scalable, low-latency processing of multiple concurrent video streams.

Evaluation on a 20-second Vietnamese traffic video achieves a recall of 83.34%, a precision of 55.56%, and an average throughput of ≈ 11 FPS. Furthermore, the first annotated dataset for vehicle Re-ID in Vietnamese traffic is introduced, carefully labeled to capture real-world complexity and diversity.

The main contributions are threefold: (1) a practical Re-ID solution customized for Vietnamese traffic conditions; (2) a novel cross-camera association method that enhances identity preservation in dense, dynamic environments; and (3) a scalable real-time pipeline suitable for deployment in large-scale transportation networks. These contributions narrow the gap between benchmark-driven research and operational systems, providing a foundation for future intelligent transportation applications.

The source code and dataset are available at: <https://github.com/MaPhat/Ds201-Re-identification-.git>.

Index Terms—Vehicle Re-Identification, Multi-camera tracking, YOLO, Deep SORT, Cross-camera association, Real-time processing, Apache Spark, Apache Kafka, Vietnamese traffic dataset

I. INTRODUCTION

Rapid urbanization and the adoption of intelligent transportation systems have made vehicle monitoring and tracking crucial to modern smart cities. Vehicle re-identification (Re-ID) — recognizing the same vehicle across non-overlapping camera views — is a key technology for multi-camera tracking, supporting traffic management, incident detection, law enforcement, and public safety. Although Re-ID research has advanced substantially, most state-of-the-art methods are evaluated under controlled laboratory conditions with stable

illumination and viewpoints, which rarely reflect real-world urban traffic challenges.

Vietnamese traffic presents unique difficulties that degrade Re-ID performance. Roads are crowded, vehicle types and appearances vary widely, and imagery often suffers from noise, blur, occlusion, and adverse weather (e.g., rain, low-light). Models trained on idealized datasets therefore experience significant drops in accuracy, highlighting the need for Re-ID solutions tailored to Vietnamese conditions.

To address this issue, a real-time multi-camera vehicle tracking pipeline is proposed, integrating YOLO for detection, Deep SORT for short-term tracking, and a robust cross-camera association algorithm developed specifically for long-term Re-ID. The system is implemented on a distributed Apache Spark–Kafka platform, enabling scalable, low-latency processing of video streams from multiple cameras.

A major contribution of this work is the introduction of a new real-world dataset from Vietnamese traffic cameras — the first public Re-ID dataset, to the best of current knowledge, that captures dense, heterogeneous, and unstructured traffic typical of Vietnam. Experiments demonstrate that the unified design significantly improves robustness and real-time performance compared with existing approaches, confirming its feasibility for deployment in complex urban environments.

In summary, this study (1) introduces a Vietnamese vehicle Re-ID dataset, (2) presents a unified real-time multi-camera tracking system combining advanced detection, tracking, and Re-ID, and (3) validates its large-scale deployment feasibility on a distributed Spark–Kafka platform, bridging the gap between academic research and real-world traffic applications.

II. RELATED WORK

Traditional multiple object tracking (MOT) methods often rely on a tracking-by-detection paradigm, where the quality of tracking is strongly influenced by the performance of the object detector. A notable example is the framework proposed by Bewley et al. [1], which introduced a simple yet effective online tracking system emphasizing frame-to-frame association using classical methods such as the Kalman Filter and Hungarian algorithm. Their study demonstrated that with

modern, high-quality object detectors, state-of-the-art tracking performance can be achieved without relying on complex models.

To address the limitations of frame-based tracking in handling occlusions, Wojke et al. [2] proposed Deep SORT, an extension of the Simple Online and Realtime Tracking (SORT) framework. While SORT focuses on fast and simple tracking using Kalman filtering and the Hungarian algorithm, Deep SORT incorporates deep appearance features to significantly improve identity preservation. By learning a deep appearance embedding from large-scale person re-identification datasets in an offline stage, Deep SORT can associate detections across longer-term occlusions using nearest-neighbor matching in the learned feature space. This enhancement results in a 45% reduction in identity switches compared to the original SORT, while maintaining real-time performance.

In addition to the traditional deep learning models integrated into the Deep SORT algorithm, Tao Jin et al. [3] proposed using the Omni-Scale Net (OSNet) deep learning model, which is designed for re-identification tasks, to extract appearance features of vehicle images. Field experiments using video from three surveillance cameras on the Jiubao Bridge in Hangzhou, China, reported an accuracy of over 97.7% for single-camera vehicle tracking and 92.5% for cross-camera tracking in non-overlapping views. Furthermore, the model outperformed other models such as Res50-dim, PCB, and HA-CNN in terms of Rank-1 and mAP metrics, achieving 92.5% and 97.9%, respectively.

Following the trend of improving deep learning models for the re-identification task, Dariusz Kobiela et al. [4] made a significant contribution by developing the MobileNetV2 model along with three variants of the original model for vehicle classification, achieving an accuracy of 96% on the second test set and approximately 76%–77% on the first, more challenging test set. In our experiments, we will also evaluate this high-performing model on the re-identification task.

Furthermore, within the domain of traffic image recognition and tracking, feature analysis constitutes a fundamental component. W. Zhou et al. [5] distinguished two principal paradigms for feature analysis: Local Feature-Based and Global Feature-Based methods. Considering the practical requirement of real-time processing—where a compromise on accuracy is acceptable—the Global Feature-Based paradigm offers a favorable trade-off by ensuring computational efficiency and rapid inference. This characteristic renders it particularly well-suited for the system objectives pursued in the present study.

Our work builds upon Deep SORT by adapting it for the vehicle re-identification domain in a multi-camera urban environment. Unlike pedestrian tracking, vehicle tracking in dense Vietnamese traffic requires handling occlusion scenarios and intra-class similarity (e.g., multiple vehicles of the same model or people driving vehicles with similar appearance). We integrate Deep SORT with a modern Re-ID model (e.g., OSNet) trained on vehicle datasets to enable robust cross-camera identity association, addressing the challenges specific

to vehicle re-identification across non-overlapping views.

III. METHODOLOGY

The Simple Online and Realtime Tracking (SORT) method [1] represents a tracking-by-detection framework with low computational overhead. It applies a Kalman Filter to estimate object motion and employs the Hungarian algorithm to assign detections across frames. Since it only utilizes bounding box data, SORT achieves high speed and efficiency. However, its performance declines when facing occlusion or appearance ambiguities.

Deep SORT [2] extends SORT by incorporating deep appearance features learned from a pre-trained re-identification model. These features allow the tracker to distinguish between visually similar objects and maintain consistent identities across short occlusions. During inference, Deep SORT matches detections to existing tracks by combining motion (from Kalman Filter) and appearance similarity (using cosine distance in embedding space). Moreover, the matching cascade algorithm plays a crucial role that helps Deep SORT improve the performance of SORT.

In this system, Deep SORT serves as the core framework for intra-camera vehicle tracking. Identity consistency across cameras is further enhanced by integrating a dedicated vehicle Re-ID module and a customized matching strategy designed for urban traffic scenarios. The pipeline can be seen in Figure 1.

A. ReID Query Algorithm

Currently, most object re-identification (Re-ID) systems rely on deep learning models to extract appearance-based features for identity matching. However, training such models from scratch typically requires a large amount of labeled data and substantial computational resources, which poses challenges for real-world deployments. Instead of training a deep model from scratch, a pre-existing architecture is adopted and fine-tuned on a vehicle dataset collected from real traffic conditions in Vietnam. This fine-tuning process allows the model to better adapt to local challenges such as diverse vehicle types, complex urban traffic scenes, and non-overlapping camera views. The extracted features are subsequently utilized in a lightweight identity-matching strategy, referred to as the Re-ID Query Matching Algorithm, which is inspired by the matching cascade algorithm in Zhao et al. [6]. The matching logic combines cosine similarity with a short-term voting mechanism, enabling accurate cross-camera identity association without the need for a heavy Re-ID model during inference.

B. Implementation of big data video streaming processing system

To deploy the big data streaming video processing system, two popular platforms were used: Apache Kafka and Apache Spark. Their combination enabled the construction of a robust architecture capable of processing large volumes of real-time video data from multiple cameras. Kafka acts as a tool for collecting and managing data from various camera

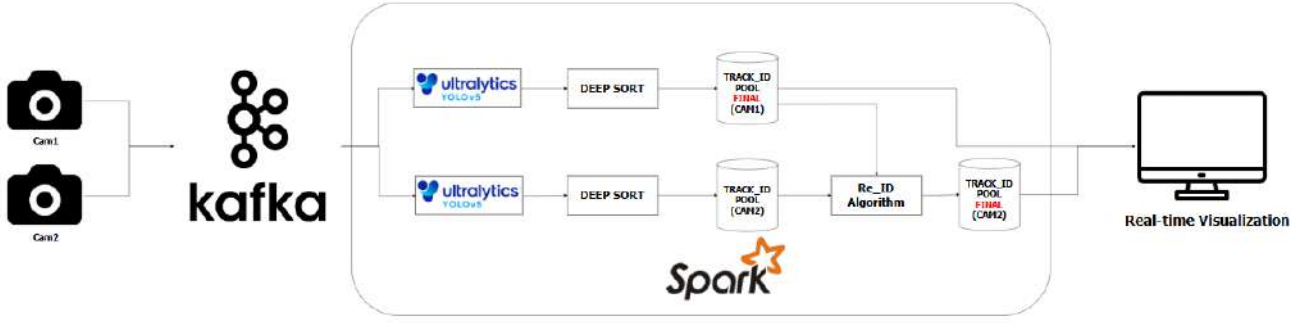


Fig. 1: Our proposed system is a real-time multi-camera vehicle tracking pipeline that combines object detection, tracking, and re-identification, implemented in a distributed streaming architecture using Apache Spark and Kafka.

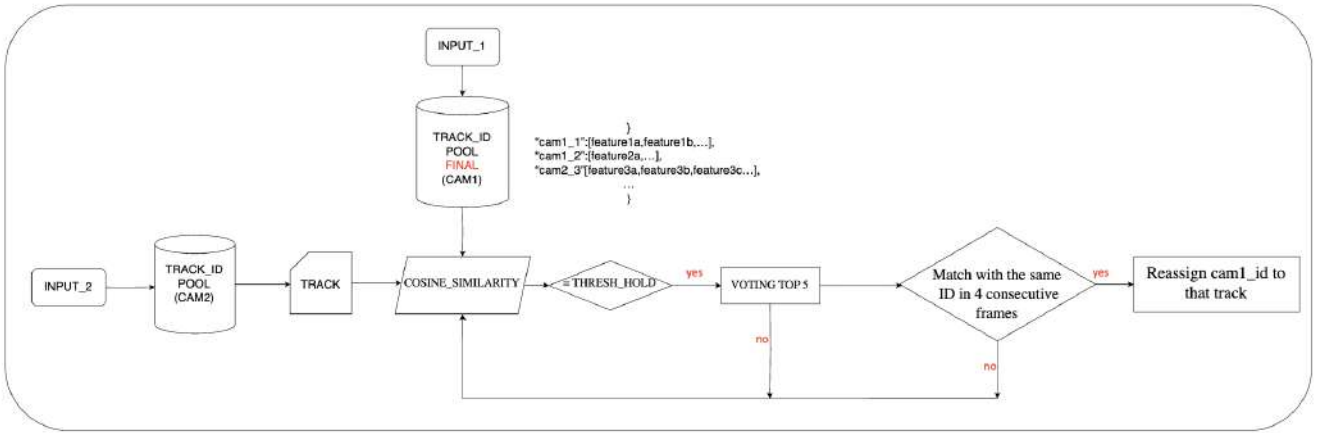


Fig. 2: Our proposed system is a real-time multi-camera vehicle tracking pipeline that combines object detection, tracking, and re-identification, implemented in a distributed streaming architecture using Apache Spark and Kafka.

sources, ensuring durability and fault tolerance. Meanwhile, Spark Streaming is utilized for distributed and highly scalable processing and analysis of this data.

With this architecture, the entire multi-camera vehicle tracking process, from initial object detection and intra-camera tracking with Deep SORT, to inter-camera identity synchronization through the Re-ID mechanism, is carried out efficiently. This enables continuous processing of information from complex traffic environments, providing continuous and accurate vehicle tracking on a large scale.

Spark is configured to continuously read data streams from Kafka topics. Each Kafka record contains information about individual video frames from surveillance cameras. Frame data is Base64 encoded and transmitted as JSON strings, along with corresponding frame IDs and camera IDs, ensuring flexibility and scalability.

At Spark's processing layer, a User-Defined Function (UDF) with the Pandas UDF API integrates the YOLOv8 object detection model. This UDF is designed to load the YOLO

model only once per Spark executor. For every detected object, the output includes its spatial position (bounding box), the associated confidence value, and category information (numerical identifier and label). These elements are serialized into JSON format. Next, another critical UDF is applied to perform Deep SORT tracking and inter-camera Re-ID.

Finally, the output data from the tracking and Re-ID processing, including tracked bounding boxes and vehicle IDs (both local and global), along with the original frame data, is normalized into JSON format and written back to a new Kafka topic (visualization). This creates a rich and fully processed data stream, ready for visualization tasks, and demonstrates the system's ability to perform real-time, multi-camera vehicle tracking.

IV. EXPERIMENTS

A. Dataset

To support the vehicle re-identification task in real-world traffic scenarios, we constructed a custom dataset recorded at

Algorithm 1 Re-ID Query Matching Algorithm

Input: Feature vector f_q from new track in Cam_2 ;
Feature pools $\mathcal{P}_{cam1} = \{id_i : [f_{i1}, f_{i2}, \dots]\}$ from Cam_1 ;
Similarity threshold τ ; Voting window size W ; Majority count P

Output: Assigned Global ID or retain local ID

- 1: Compute cosine similarities $s_i = \cos(f_q, f_i)$ for all $f_i \in \mathcal{P}_{cam1}$
- 2: Select top-5 most similar IDs with $s_i > \tau$
- 3: Store top-1 match result into a sliding buffer of size W
- 4: **if** any ID appears at least P times in the buffer **then**
- 5: Assign that ID as the global ID for f_q
- 6: **else**
- 7: Keep original local ID
- 8: **end if**



Fig. 3: The two locations used for data collection: An Phu Station and Thao Dien Station in Ho Chi Minh City, Vietnam.

two locations in Ho Chi Minh City: An Phu Station and Thao Dien Station. Video data was collected using two smartphones with identical camera specifications, recording simultaneously at both sites to simulate non-overlapping camera views. The dataset consists of two video clips, recorded at two different locations (one video per location), each lasting approximately 5–6 minutes. The dataset focuses on two main types of vehicles: cars and motorbikes.

Annotations were manually labeled in YOLO format by a team of three annotators, including:

TABLE I: Dataset annotation roles and data split

Annotation Team	Description
1 Gold annotator	Produces high-quality reference labels
2 Regular annotators	Support the annotation process

The dataset was split into 7:2:1 for training, test, and validation, respectively. All data were recorded in clear daylight with stable lighting and no rain to ensure consistency across camera views. Annotation followed strict rules: vehicles with more than 60% occlusion were ignored, and only those in the bottom half of the frame were labeled to simulate realistic detection zones.

Annotation Agreement Metrics

To evaluate the quality and consistency of the manual annotations, the agreement among annotators was measured using a randomly selected subset of labeled data. This process involved comparing annotations made by two regular annotators with a reference (gold) annotator. The metrics were computed based on bounding box overlaps and identity matching, using standard evaluation criteria in object detection and tracking.

- **Average F1-Score:** 90.61%
- **Average Intersection over Union (IoU):** 84.92%
- **Mean Average Precision (mAP) @ [IoU = 0.50:0.95]:** 63.4%

These results confirm that the dataset provides reliable supervision signals for training and evaluating vehicle Re-ID models in real-world conditions.

B. Metrics

To evaluate the performance of the YOLO models, we adopted benchmark metrics based on the evaluation described in Jegham, Nidhal, et al. [7].

- **Precision** is the proportion of correctly predicted positive instances among all predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** is the proportion of correctly predicted positive instances among all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **IoU (Intersection over Union)** is a metric that measures the overlap between the predicted bounding box and the ground truth bounding box.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

- **mAP@50** is the mean of AP across all classes using an IoU threshold of 0.50.

$$\text{mAP}_{@50} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i \quad (\text{with } \text{IoU} \geq 0.50)$$

where AP is the area under the precision-recall curve for a single class.

- **mAP@50:95** is the average AP over 10 IoU thresholds ranging from 0.50 to 0.95 (COCO-style).

$$\text{mAP}_{@[.50:.95]} = \frac{1}{10} \sum_{k=0}^9 \text{mAP}_{@0.5+0.05k}$$

To evaluate the performance of the deep learning models, standard metrics commonly used in the evaluation of state-of-the-art (SOTA) approaches are adopted. Specifically, mean Average Precision (mAP) is reported, as well as Cumulative Matching Characteristic (CMC) metrics at Rank-1, Rank-5, Rank-10, and Rank-20. **mAP** measures the overall retrieval



Fig. 4: The red box illustrates a region that should not be annotated due to occlusion.

precision considering both ranking and matching quality, while **Rank-k** metrics represent the percentage of queries for which the correct identity is found among the top-k retrieved results.

To evaluate the performance of re-identification in a multi-camera deployment setting, the Confusion Matrix is utilized, along with the metrics Recall and Precision. The components of the Confusion Matrix are defined as follows:

- **True Positives (TP):** The model successfully matches an object that actually disappeared from camera 1 and reappeared in camera 2.
- **False Negatives (FN):** The model fails to match an object that actually disappeared from camera 1 and reappeared in camera 2.
- **False Positives (FP):** The model returns a match even though the object did not actually disappear from camera 1 and reappear in camera 2.
- **True Negatives (TN):** The model returns no match, and in reality, the object did not disappear from camera 1 and reappear in camera 2.

C. Experimental Result

Two object detection models, YOLOv5 [7] and YOLOv8 [8], were trained using a labeled vehicle dataset as described in Section IV-A. The training process was conducted on Google Colab with a NVIDIA Tesla T4 GPU (16 GB VRAM).

Both models were trained for 50 epochs using the AdamW optimizer with a learning rate of 0.001667 and momentum of 0.9. By keeping the dataset, hardware, and training configurations consistent, a fair comparison of detection performance between the two YOLO architectures was ensured.

TABLE II: Comparison of detection performance between YOLOv5 and YOLOv8

Metrics	YOLOv5	YOLOv8
Precision	0.9034	0.9630
Recall	0.9190	0.8715
mAP@0.5	0.9716	0.9623
mAP@0.5:0.95	0.6205	0.6119

Next, two re-identification backbone models were fine-tuned on the VeRi-776 dataset and subsequently subjected to transfer

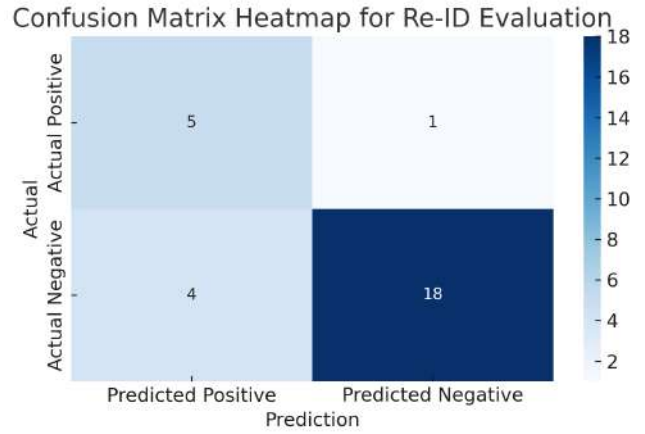


Fig. 5: Confusion Matrix for Re-ID Evaluation

learning using the dataset introduced in Section IV-A. The two models were trained for 60 and 12 epochs, respectively, using the Adam optimizer. The learning rate schedule followed the strategy proposed by Rong et al. [9]. The training objective was based on the Triplet Loss with a margin of 0.3, which encourages the models to learn discriminative embeddings for vehicle re-identification across different camera views.

To evaluate the performance of the pipeline, experiments were conducted on a 20-second video segment. The results showed that the system achieved a **Recall of 83.3%** and a **Precision of 55.56%**, with the corresponding Confusion Matrix illustrated in Figure 5.

Finally, two object detection models, YOLOv5 and YOLOv8, were integrated with two deep learning-based Re-ID models, OSNet [3] and MobileNetV2 [3], into the streaming data system to evaluate the performance of real-time applications. The system was deployed on a data stream to monitor two 5-minute videos. The table below presents the processing time for each model combination. All experiments were conducted on a server equipped with 16GB RAM, an AMD Ryzen 5 1600 CPU, and a GTX 1060 3GB GPU. The results indicate that the combination of YOLOv8 and MobileNetV2_x1_0 yields the fastest processing time per

TABLE III: Comparison of vehicle re-identification performance fine-tuning in VeRi-776 [10] between OSNet [3] and MobileNetV2 [3]

Model	mAP	Rank-1	Rank-5	Rank-10	Rank-20
OSNet (osnet_x1_0)	60.2%	88.2%	91.4%	93.9%	96.2%
MobileNetV2 (mobilenetv2_x1_0)	62.2%	88.4%	91.8%	94.6%	97.2%

TABLE IV: Comparison of detection performance between YOLOv5 and YOLOv8

Method	Execution Time per Batch	Average FPS
YOLOv5 + Osnet_x1_0	~9.35s	~9fps
YOLOv5 + MobileNetV2_x1_0	~8.1s	~10fps
YOLOv8 + Osnet_x1_0	~7.9s	~10fps
YOLOv8 + MobileNetV2_x1_0	~7.21s	~11fps

batch at approximately 7.21 seconds, achieving an average frame rate of around 11 FPS.

V. CONCLUSION

This study introduces a novel and practical solution to the problem of multi-camera vehicle re-identification (Re-ID) in the complex and dynamic context of Vietnamese urban traffic. The proposed system integrates state-of-the-art deep learning models with a scalable big data streaming infrastructure (Spark and Kafka), enabling real-time processing and identity preservation across multiple camera feeds in large-scale deployments. Unlike existing approaches, the pipeline is specifically fine-tuned to address the distinctive characteristics of domestic traffic—characterized by high congestion, diverse vehicle types, and unstructured motion patterns—and is fully integrated into a real-time distributed environment.

A key contribution of this work is the release of the first publicly available vehicle Re-ID dataset constructed from real-world Vietnamese traffic, capturing the heterogeneity and visual challenges that are largely absent from standard benchmarks. Experimental results validate the system's effectiveness, achieving a recall of 83.34%, a precision of 55.56%, and a processing speed of approximately 11 frames per second. These outcomes confirm the system's strong potential for operational deployment in intelligent transportation applications.

Overall, this work bridges the gap between academic Re-ID research and practical deployment by providing a comprehensive, scalable, and context-aware solution. It also establishes a foundation for future advancements in traffic monitoring and smart-city infrastructure in Vietnam and other countries with similar traffic dynamics. However, the dataset used remains limited in terms of camera viewpoints and weather conditions. In addition, the system has not yet conducted a diverse comparison of other state-of-the-art models. In the future, the dataset could be expanded with more viewpoints under various weather conditions to better reflect real traffic scenarios in Vietnam, while also integrating and benchmarking more advanced models.

ACKNOWLEDGEMENT

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

REFERENCES

- [1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016, pp. 3464–3468.
- [2] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, 2017, pp. 3645–3649.
- [3] T. Jin, X. Ye, Z. Li, and Z. Huo, "Identification and tracking of vehicles between multiple cameras on bridges using a yolov4 and osnet-based method," *Sensors*, vol. 23, no. 12, p. 5510, 2023.
- [4] D. Kobiela, J. Groth, M. Hajdasz, and M. Erezman, "Vehicle type recognition: a case study of mobilenetv2 for an image classification task," in *Procedia Computer Science*, vol. 246, 2024, pp. 3947–3956.
- [5] W. Zhou, L. Yang, L. Zhao, R. Zhang, Y. Cui, H. Huang, K. Qie, and C. Wang, "Vision technologies with applications in traffic surveillance systems: A holistic survey," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1–47, 2025.
- [6] H. Zhao and J. Zhai, "A cascade matching algorithm for multi-object tracking systems based on detection box scores," in *Proceedings of 2022 Chinese Intelligent Systems Conference (CISC 2022)*, ser. Lecture Notes in Electrical Engineering, Y. Jia, W. Zhang, Y. Fu, and S. Zhao, Eds. Singapore: Springer, 2022, vol. 950.
- [7] R. Khanam and M. Hussain, "What is yolov5: A deep look into the internal features of the popular object detector," *CoRR*, vol. abs/2407.20892, 2024.
- [8] M. Yaseen, "What is yolov8: An in-depth exploration of the internal features of the next-generation object detector," *arXiv preprint*, 2024.
- [9] L. Rong, Y. Xu, X. Zhou *et al.*, "A vehicle re-identification framework based on the improved multi-branch feature fusion network," *Scientific Reports*, vol. 11, no. 20210, 2021.
- [10] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multi-modal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, March 2018.

Conversational Product Navigation for E-commerce via Knowledge-Based Multi-Agent Chatbots

Gia-Bach Truong[§]

*School of Information and Communications Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
bachtg.210087@sis.hust.edu.vn*

Tuan-Dat Trinh^{§*}

*School of Information and Communications Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
dattt@soict.hust.edu.vn*

Abstract—Chatbots have become essential tools on e-commerce platforms, enabling automated customer consultation and support. However, most existing chatbots remain passive, merely responding to user-initiated queries without aligning with business objectives. This paper proposes a comprehensive chatbot architecture that integrates a smart question suggestion system to proactively guide conversations toward predefined business goals while maintaining natural and user-relevant interactions. The architecture combines a product knowledge graph, which links products through shared attributes, with a multi-agent system powered by large language models (LLMs) to generate goal-oriented question suggestions. The system’s effectiveness is evaluated using a simulated user environment, demonstrating its ability to maintain engaging dialogues and steer users toward target products. The proposed chatbot thus balances user satisfaction with strategic business outcomes.

Index Terms—chatbot, LLM, question suggestion, multi-agent

I. INTRODUCTION

The rapid advancement of artificial intelligence (AI), particularly large language models (LLMs), has revolutionized human-computer interaction. In the e-commerce sector, chatbots have quickly become a common communication tool, enabling businesses to automate customer support, provide product information, and offer round-the-clock assistance. Modern chatbots possess semantic understanding and can respond in a natural manner, reducing users’ effort in information retrieval and significantly enhancing the shopping experience.

However, most existing chatbots remain limited to a reactive role. They perform well when users clearly know what they are looking for and ask direct questions. Yet, in the role of a proactive sales assistant, these systems fall short. They often lack the ability to guide, suggest, or redirect user attention toward strategic products, promotional items, or newly launched offerings that businesses aim to promote. This limitation presents a dual challenge: how to simultaneously satisfy users’ information-seeking needs (user-oriented) while achieving specific business objectives (business-oriented).

To address these challenges, question suggestion systems represent a promising research direction. An effective system should not only generate contextually relevant questions

that save users time, but also subtly guide the conversation. Achieving this requires a delicate balance between maintaining user engagement by proposing topics of potential interest and strategically steering the dialogue toward target products.

This paper proposes a comprehensive architecture for building a goal-oriented chatbot with integrated question suggestion capabilities for e-commerce platforms. The key contributions are threefold. (i) First, we construct a product knowledge graph that connects items through shared attributes, enabling logical pathways to guide conversations. (ii) Second, to support small and medium-sized enterprises that lack training data on user interaction histories, we develop a multi-agent system powered by large language models (LLMs) to both answer user queries and generate strategic question suggestions that naturally lead users toward target products. (iii) Third, to enable objective and repeatable evaluation without relying on costly user studies or potentially biased feedback from non-authentic users, we introduce an LLM-based user simulation environment.

The remainder of the paper is organized as follows. Section II reviews related work. Section III presents the construction of the product knowledge graph. Section IV introduces the multi-agent system and its core components. Section V describes the simulation setup for modeling user behavior. Section VI details the experimental setup and results. Finally, Section VII concludes the paper and discusses future directions.

II. RELATED WORK

We review prior work across three key areas: e-commerce chatbots, goal-oriented and persuasive recommendation systems, and the use of knowledge graphs in recommender systems.

AI-driven chatbots have been widely adopted in e-commerce to enhance user experience and streamline customer support. Sidlauskiene et al. [1] showed that using human-like language in AI-driven chatbots can significantly enhance perceived product personalization, with situational loneliness moderating this effect. They also found that the interaction between these factors can increase consumers’ willingness to pay a higher price in conversational commerce. Jiang et al. [2] introduced theoretical models to explain how dialogic chatbot-user interactions affect satisfaction and relationship building.

[§]The two authors contributed equally to this work

^{*}Corresponding author

Cheng and Jiang [3] applied the uses and gratification theory to analyze trade-offs between utility and privacy concerns in chatbot usage.

However, most existing systems are reactive. They respond to user queries without taking the initiative to steer conversations toward business goals or suggest follow-up actions. This limitation highlights the need for more proactive and goal-oriented conversational agents.

Recent research has increasingly focused on proactive systems that guide users rather than merely respond to their preferences. Bi et al. [4] proposed the iterative preference guidance (IPG) model, which directs users toward new interests through strategically sequenced recommendations and explicitly defined guiding objectives, helping to mitigate filter bubbles. Jeunen et al. [5] developed a multivariate policy learning approach for multi-objective recommendation, enabling the joint optimization of competing goals such as user satisfaction and business outcomes. Alslaity and Tran [6] introduced PerPer, a personalized persuasive recommender that incorporates ethical considerations when influencing user decisions. Li et al. [7] further demonstrated that goal-oriented systems can be deployed effectively even in resource-constrained environments, particularly in educational domains.

Knowledge graph integration has become a powerful method for modeling structured relationships among entities in recommendation systems. He et al. [8] introduced LightGCN, showing that streamlined architectures focusing solely on neighborhood aggregation can outperform more complex graph-based models. Wu et al. [9] applied self-supervised learning techniques to alleviate noise and degree bias in user-item interactions, making the approach particularly effective for platforms with sparse data. Wang et al. [10] proposed KGAT, which leverages attention mechanisms to model high-order relationships, yielding both accurate and interpretable recommendations. Complementing these efforts, Wang et al. [11] developed NGCF, a foundational model that explicitly captures high-order connectivity within user-item interaction graphs using neural networks.

Our proposed system integrates key advances from prior research by combining user-friendly chatbot interactions with goal-oriented question suggestion, leveraging knowledge graphs to construct coherent dialogue flows, and employing a multi-agent architecture powered by large language models.

III. PRODUCT KNOWLEDGE GRAPH FOR CONVERSATIONAL GUIDANCE

An effective question suggestion system requires a deep understanding of semantic relationships among products. To steer a conversation from a user’s initial interest to a business-defined target item, the system must identify a coherent sequence of intermediate products that form a logical and convincing progression. A knowledge graph offers an ideal framework for modeling these complex interconnections, enabling structured and explainable navigation across the product space.

A. Building the Product Knowledge Graph

We collected and processed data from the Vietnamese e-commerce website *Thegioididong.com*, focusing on the smartphone category. From 193 smartphone models, we extracted 16 key attributes to construct a product knowledge graph with 107 nodes, each representing a distinct product line, and 10,789 edges encoding semantic relationships between products. For evaluation, we restrict our study to smartphones and a dataset of 193 models to demonstrate feasibility. Scaling to larger, cross-domain datasets remains an important direction for future work.

The knowledge graph is constructed based on three fundamental components: nodes, relationships, and relationship weights. We implemented the graph using the Neo4j graph database. Each node represents a unique smartphone model, such as “iPhone 15 Pro” or “Samsung Galaxy S25.” Variants that differ only in minor details, such as storage capacity, are grouped into a single node to avoid redundancy.

A relationship is established between two nodes whenever the corresponding products share the same value in an attribute. As shown in Table I, we define eleven types of relationships, each corresponding to a specific product attribute and capturing a distinct semantic connection. To reflect the relative importance of each relationship in guiding conversations, a priority weight is assigned. Relationships with lower weights are given higher priority during path search. This weighting scheme is essential, as different product attributes carry varying levels of influence on user decision-making. For example, users tend to care more about price and brand than about case material. Since product relationships are derived from shared values in certain attributes, the corresponding relationship types naturally vary in importance.

A key aspect of our graph design is the integration of personalization context through user interaction history. Nodes corresponding to products that a user has previously viewed or purchased are marked with a user-specific flag indicating prior interaction. This annotation enriches the graph with personalized signals while preserving its topological structure.

This enrichment enables the system to strategically handle *two distinct suggestion strategies*. The first involves recommending a product from the user’s interaction history, which serves as a familiar anchor point, potentially increasing user trust and maintaining conversational continuity. In contrast, suggesting a completely new product creates an opportunity for exploration and interest expansion. While our system supports both strategies, the path-finding algorithm presented below is deliberately designed to favor routes that pass through previously interacted nodes. We hypothesize that this prioritization yields suggestions that feel more relevant and less intrusive, thereby significantly improving the likelihood of user acceptance.

B. Goal-Oriented Path Search in the Product Knowledge Graph

To guide users effectively, the system computes optimal paths from the source entities, which are the products men-

TABLE I
TYPES OF RELATIONSHIPS AND THEIR PRIORITY WEIGHTS IN THE
PRODUCT KNOWLEDGE GRAPH.

Relationship Type	Weight	Source Attributes
Same brand	1	Brand
Same price range	1	Price
Equivalent memory configuration	2	Internal storage, RAM
Equivalent processor	2	CPU
Equivalent camera system	2	Front camera, Rear camera
Equivalent display	2	Display technology, Resolution, Size, Refresh rate
Same operating system	4	Operating system
Similar battery capacity	3	Battery capacity
Equivalent material	3	Body material
5G support	4	5G capability
Similar release date	4	Release date

tioned by the user, to a target entity defined as the product that the business aims to promote. An optimal path must meet two criteria. First, it should be short and contextually coherent so that the recommendation appears natural rather than forced. Second, it should align with the user’s preferences to increase the likelihood of acceptance.

The path search algorithm is designed to leverage the personalized knowledge graph. It systematically explores and scores candidate paths, with a scoring function that explicitly prioritizes those passing through user-marked nodes. The algorithm proceeds in three steps:

Step 1 – Source and Target Identification: Determine the source entities based on the current user query and retrieve the target entity as predefined by the business.

Step 2 – Shortest Path Enumeration: Use a graph traversal algorithm (e.g., bidirectional BFS) to find all shortest paths from each source entity to the target entity.

Step 3 – Path Filtering and Ranking: Since multiple shortest paths may exist, each path is scored using a function that combines two factors: the priority of the relationships it traverses and the extent to which it aligns with the user’s interaction history. The scoring function for a path p is defined in Equation (1), where:

- α and β are weighting coefficients that balance system-defined priorities and user personalization, subject to $\alpha + \beta = 1$.
- w_r is the priority weight assigned to relationship r , as defined in Table I. The inverse $\frac{1}{w_r}$ ensures that more important relationships (with lower weights) contribute more to the score.
- $\mathcal{R}(p)$ denotes the set of relationships along path p .
- $\mathcal{E}(p)$ denotes the set of entities (products) along path p .
- $I(e)$ is an indicator function that returns 1 if entity e appears in the user’s interaction history H , and 0 otherwise.

$$\text{Score}(p) = \alpha \sum_{r \in \mathcal{R}(p)} \frac{1}{w_r} + \beta \sum_{e \in \mathcal{E}(p)} I(e) \quad (1)$$

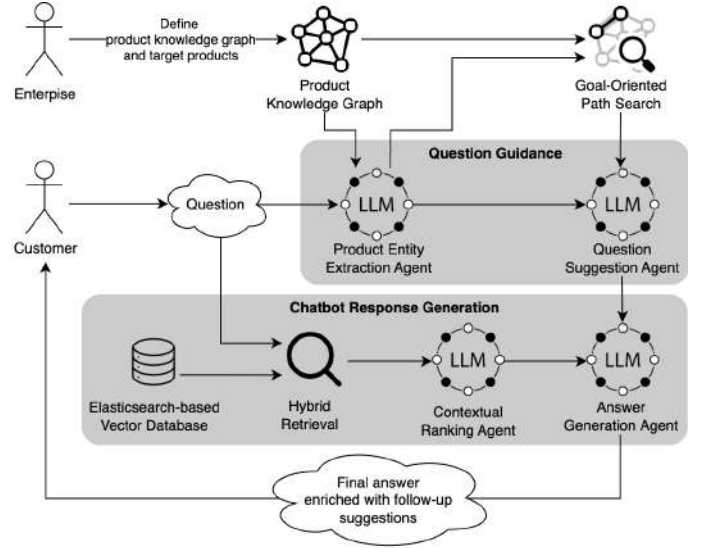


Fig. 1. The overall architecture of the multi-agent chatbot with goal-oriented question suggestion.

This scoring approach favors paths that traverse high-priority relationships (e.g., brand, price range) and include products previously interacted with by the user. The system then selects the top N highest-scoring paths (e.g., $N = 2$) and forwards them to the *Question Guidance Component* for question generation.

IV. MULTI-AGENT CHATBOT SYSTEM WITH GOAL-ORIENTED QUESTION SUGGESTION

The system is built upon a multi-agent architecture, in which each agent is a specialized software module powered by a large language model (LLM) and dedicated to a specific subtask. This modular design allows the system to handle complex conversational workflows in a flexible and coordinated manner. The overall architecture is depicted in Figure 1. All LLM-based agents are implemented with the Gemini 2.5 Flash model, selected for its efficiency and strong reasoning capabilities in multi-turn dialogue. For embedding generation, we use Google’s `text-embedding-004` model, as described in Section IV-B.

In addition to the product knowledge graph, the system consists of two primary components: the *Question Guidance Component* and the *Chatbot Response Generation Component*. Their integration enables a seamless conversational flow, where each response not only answers the user’s question but also introduces follow-up suggestions aligned with business objectives, establishing an intelligent and proactive interaction loop.

A. Question Guidance Component

This component is responsible for generating suggestion questions to guide the conversation toward business-defined goals. It consists of two LLM-based agents: the *Product Entity Extraction Agent* and the *Question Suggestion Agent*.

Product Entity Extraction Agent: When the user poses a query, this agent identifies the names of smartphone products mentioned in the utterance. It uses a large language model prompted for the Named Entity Recognition (NER) task. The extracted product names are then matched against the product knowledge graph using fuzzy search techniques to locate the corresponding source nodes.

Question Suggestion Agent: This agent receives as input a set of high-scoring paths from the knowledge graph traversal algorithm. For each path, it extracts a triplet consisting of the source product, the next product along the path, and the semantic relationship connecting them. Based on this triplet, the agent generates two outputs: a *rationale sentence* and a *follow-up question*.

The *rationale sentence* provides a natural-language explanation of why the next product is relevant. For example: “Alongside the Samsung Galaxy S25, another phone in the same price range of 20–25 million VND that you might find interesting is the iPhone 15.”

The *follow-up question* is a targeted inquiry about the suggested product, designed to encourage further user engagement. For instance: “Would you like to explore how the iPhone 15’s performance compares to the previous model?”

On the chatbot interface, the *rationale sentence* is appended directly to the previous system response, while the *follow-up questions* are displayed below as clickable options for the user. This design enables a seamless connection between the suggested product (e.g., iPhone 15) and the user’s current interest (e.g., Samsung Galaxy S25) as they share the same attribute value, such as price range. The follow-up questions, however, are not constrained to this specific attribute; they are designed to encourage broader exploration of the suggested product across multiple dimensions.

The agent is implemented using a large language model with a few-shot prompting approach, where a small set of input-output examples is embedded in the prompt to guide the model toward generating the desired outputs.

B. Chatbot Response Generation Component

This component is responsible for answering user questions, whether they are initiated directly by the user or selected from system-generated suggestions. The chatbot is implemented using a Retrieval-Augmented Generation (RAG) architecture, which allows it to dynamically combine language generation capabilities with knowledge retrieved from a product database.

To construct the knowledge base, we collected all textual content related to smartphone products from the Vietnamese e-commerce website *Thegioididong.com*. The data was processed and segmented into meaningful chunks while preserving the structural integrity of the source, such as avoiding splits between section headers and their corresponding content. These text segments were then embedded into dense vector representations using Google’s `text-embedding-004` model and stored in an Elasticsearch-based vector database.

When a user submits a query, the RAG pipeline is triggered with a *hybrid retrieval* mechanism. Instead of relying solely

on vector-based semantic search, we adopt Elasticsearch’s Reciprocal Rank Fusion (RRF) hybrid search strategy, which combines results from both traditional lexical search (BM25) and dense vector search. This approach leverages the strengths of both methods: BM25 excels at retrieving exact keyword matches such as model names or technical specifications, while vector search offers superior performance in capturing user intent and semantic relevance.

The retrieved text chunks are passed to an LLM-based *Contextual Ranking Agent*, which filters out irrelevant segments and reorders the remaining ones based on their semantic relevance to the user query. The top-ranked segments are then provided as contextual input to the *Answer Generation Agent*, which synthesizes the information and generates a complete, coherent response in Vietnamese.

V. USER SIMULATION FOR SYSTEM EVALUATION

One of the main challenges in developing conversational systems is the evaluation process. Testing with real users is expensive, time-consuming, and difficult to control due to numerous external variables. Moreover, unless participants are genuinely interested in purchasing a product, their responses may be biased or inconsistent, which undermines the objectivity of the evaluation. To address these challenges, we developed a simulated environment in which an LLM-based *Simulated User Agent* plays the role of a human user and interacts with the chatbot system.

A. Simulated User Modeling for Interaction Testing

To realistically simulate user behavior, the *Simulated User Agent* makes decisions based on two primary inputs: (i) a fixed **User Profile** and (ii) the evolving **Dialogue Context**.

The **User Profile** remains constant throughout a session and comprises two key components: a *user persona* and a *synthetic interaction history*. The persona defines the user’s behavioral archetype, such as a student, business professional, elderly user, or mobile gamer. Each persona is specified with detailed characteristics, shopping motivations, and decision patterns. For instance, a student persona is defined by a limited budget and high sensitivity to promotions.

The interaction history simulates prior purchases and inquiries, providing the agent with an initial set of interests before the conversation begins. This profile-driven structure supports more realistic and context-aware behavior as the dialogue unfolds.

The **Dialogue Context** captures the full conversation history, including the user’s previous questions, the chatbot’s responses, and any suggested follow-up questions. The agent is instructed to prioritize the most recent question–answer exchanges in order to make context-aware decisions for the next turn.

Based on the User Profile and Dialogue Context, the Simulated User Agent selects one of two actions: **Accept**, by choosing a suggested follow-up question; or **Reject**, by ignoring all suggestions and generating a new question based on the character’s internal thoughts and needs.

To promote behavioral diversity and unpredictability, we set a high temperature value for the language model, enabling the agent to act in a more human-like and non-deterministic manner.

B. User Simulation Flow and Termination Criteria

A simulation session is a multi-turn conversation between the Simulated User Agent and the chatbot system. A session is labeled as a **success** if the user eventually asks about the target product. It is marked as a **failure** under either of the following conditions: (i) the number of dialogue turns exceeds a preset threshold (e.g., six turns) without reaching the target, or (ii) the Simulated User Agent rejects suggestions leading to the target product more than twice.

All user utterances, agent decisions, and intermediate system outputs are logged to support detailed analysis and system evaluation.

VI. EXPERIMENTAL SETUP, RESULTS, AND EVALUATION

To assess the effectiveness of the proposed system, we conducted a series of 20 simulation sessions, each representing a distinct synthetic user with a unique persona and product goal. The goal of these experiments is twofold: (i) to verify whether the system can successfully guide users toward target products in natural conversations, and (ii) to evaluate the contribution of each core component. To this end, we designed two ablation scenarios: the first disables the personalized user interaction history, and the second removes the product knowledge graph, thereby eliminating structured path guidance. These comparisons provide insight into the impact of personalization and graph-based reasoning on the overall system performance.

We acknowledge that our evaluation is limited to 20 simulated sessions and lacks validation with real-world users. Although the LLM-based simulation provides a cost-efficient and repeatable approach, it cannot fully capture the complexity of human behavior. In addition, we did not compare against proactive recommendation baselines such as Iterative Preference Guidance (IPG) [4] or multivariate policy learning [5]. Direct comparison remains difficult because most baselines are designed for offline recommendation or static datasets, whereas our system operates in a dynamic conversational environment that combines multi-turn dialogue with knowledge graph reasoning. Incorporating adapted baselines into future work will help establish a stronger empirical foundation for our contributions.

A. Evaluation Metrics

To evaluate the system’s performance from both user-centric and business-driven perspectives, we define the following metrics:

- **History Match Rate:** The proportion of turns in which the system’s suggested question contains a product that appears in the user’s past interactions. This reflects the system’s ability to tailor suggestions based on user history.

TABLE II
COMPARISON WITH AND WITHOUT USER HISTORY INFORMATION.

System	HMR	AR	ACR	GCR
Full System	79.6%	94.7%	52.0%	20.0%
No User History	2.3%	94.4%	46.0%	20.0%

HMR: History Match Rate AR: Answerability Rate
ACR: Acceptance Rate GCR: Goal Completion Rate

- **Acceptance Rate:** The proportion of times the user accepts a system-suggested question. This measures user engagement and perceived relevance.
- **Answerability Rate:** The proportion of suggested questions that the chatbot is able to answer, indicating the alignment between the Question Guidance Component and the chatbot’s knowledge base.
- **Goal Completion Rate:** The percentage of sessions that successfully end with the user inquiring about the target product. This is the primary metric reflecting business success.

B. Impact of Personalization

In this scenario, we compare the full system with a variant that removes user interaction history from the path scoring function. Results in Table II show that excluding personalization causes the **History Match Rate** to plummet from 79.6% to just 2.3%, indicating that the system becomes almost entirely unable to recommend products previously aligned with the user’s interests. This reduction also leads to a noticeable drop in the **Acceptance Rate**, confirming that user history plays a vital role in generating compelling suggestions and maintaining engagement.

C. Impact of Knowledge Graph on Business-Oriented Goals

In this scenario, we remove the knowledge graph component entirely. Instead of guiding the conversation through intermediate entities, the Question Suggestion Agent attempts to create direct recommendations from the user’s current query to the target product.

The results in Table III show a substantial decline in system performance when the knowledge graph is removed. In particular, the **Goal Completion Rate** drops to zero, indicating that the system is unable to guide users to the target product. Without structural support, it lacks logical transitions between products, resulting in abrupt and unconvincing suggestions. For example, the system may suggest a high-end device immediately after the user inquires about a budget option, leading to repeated rejections from the Simulated User Agent.

Interestingly, the **History Match Rate** increases slightly in this scenario. This is likely because user history is still embedded in the prompt, which may lead the model to overfit to previously mentioned items. However, the improvement is minimal and does not compensate for overall decline in coherence and goal alignment caused by the lack of graph-based reasoning.

TABLE III
PERFORMANCE COMPARISON WITH AND WITHOUT THE KNOWLEDGE GRAPH

System Variant	HMR	AR	AcR	GCR
Full System	79.6%	94.7%	52.0%	20.0%
Without Knowledge Graph	81.2%	63.3%	40.0%	0.0%

HMR: History Match Rate **AR:** Answerability Rate
AcR: Acceptance Rate **GCR:** Goal Completion Rate

The absence of the graph also negatively impacts the **Answerability Rate**, as the model becomes more prone to hallucinating or producing off-topic suggestions when it is no longer grounded by structured relational context. These findings highlight the pivotal role of the product knowledge graph in generating coherent, goal-aligned recommendations that meet both user expectations and business objectives.

VII. CONCLUSION AND FUTURE WORK

This paper introduced a comprehensive and practical architecture for an e-commerce chatbot that integrates goal-oriented question recommendation. By combining a product knowledge graph, a multi-agent system powered by large language models, and a user simulation environment for evaluation, the proposed framework addresses the challenge of balancing user engagement with business-driven objectives.

Experimental results validate the critical roles of both personalization and structured guidance. Incorporating user history significantly increases the relevance and acceptance of suggested questions, while the knowledge graph provides the logical scaffolding needed to guide conversations coherently and drive users toward target products.

Future directions include automating knowledge graph construction and maintenance, enabling cross-category recommendation, and incorporating real-time personalization techniques to further enhance the system’s adaptability and effectiveness.

REFERENCES

- [1] J. Sidlauskienė, Y. Joye, and V. Auraskeviciene, “AI-based chatbots in conversational commerce and their effects on product and price perceptions,” *Electronic Markets*, vol. 33, no. 1, p. 24, 2023.
- [2] H. Jiang, Y. Cheng, J. Yang, and S. Gao, “AI-powered chatbot communication with customers: dialogic interactions, satisfaction, engagement, and customer behavior,” *Computers in Human Behavior*, vol. 134, p. 107329, 2022.
- [3] Y. Cheng and H. Jiang, “How do AI-driven chatbots impact user experience? examining gratifications, perceived privacy risk, satisfaction, loyalty, and continued use,” *Journal of Broadcasting & Electronic Media*, vol. 64, no. 4, pp. 592–614, 2020.
- [4] S. Bi, L. Qu, T. Chen, J. Yu, and H. Yin, “Proactive recommendation with iterative preference guidance,” in

Companion Proceedings of the ACM Web Conference 2024, Singapore, 2024, pp. 871–874.

- [5] O. Jeunen, J. Mandav, I. Potapov, N. Agarwal, S. Vaid, W. Shi, and A. Ustimenko, “Multi-objective recommendation via multivariate policy learning,” in *18th ACM Conference on Recommender Systems*, Bari, Italy, 2024, pp. 712–721.
- [6] A. Alsaity and T. Tran, “Towards persuasive recommender systems,” in *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, Kahului, HI, USA, 2019, pp. 143–148.
- [7] W. Li, Q. Huang, and S. Gautam, “Multi-objective recommendation for massive remote teaching resources,” *Mobile Networks and Applications*, 2024.
- [8] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, “LightGCN: simplifying and powering graph convolution network for recommendation,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event China, 2020, pp. 639–648.
- [9] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie, “Self-supervised graph learning for recommendation,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event Canada, 2021, pp. 726–735.
- [10] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, “KGAT: knowledge graph attention network for recommendation,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Anchorage AK USA, 2019, pp. 950–958.
- [11] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, “Neural graph collaborative filtering,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, France, 2019, pp. 165–174.

Formalizing and Conformance Testing ABAC Policies in Process Logs via an ABAC metamodel

Duc-Hieu Nguyen

Department of Information Technology
The Academy of Cryptography Techniques
Hanoi, Vietnam
ndhieu@actvn.edu.vn

Abstract—Access control policies are crucial for establishing a secure access control system and are fundamentally based on access control models such as attribute-based access control (ABAC). Moreover, the execution of business process instances is generally recorded within an event log. It is essential to ensure conformance with ABAC policies for the event log at the time of post-execution. To conduct conformance testing of ABAC policies in business process event logs, formalizing these ABAC policies is necessary. However, the formalization is typically carried out manually, which leads to inefficiencies, low maintainability, a high likelihood of errors, and difficulty in detecting the errors. Also, there has been a notable lack of an access control metamodel to support the formalization and conformance testing of ABAC policies, as well as insufficient attention to constructing ABAC policies from existing process logs. This paper presents a model-driven workflow for significantly automating the formalization and conformance testing of ABAC policies in process logs. By leveraging a novel ABAC metamodel and its reusable patterns, the approach addresses the aforementioned issues, demonstrating effectiveness through experiments on a real-world process log.

Index Terms—ABAC policies, ABAC metamodel, process logs, formalization, conformance testing, process mining

I. INTRODUCTION

In organizations with strict security requirements, such as government agencies or banking institutions, access control (AC) is a fundamental security mechanism for these organizations' security systems. AC limits what users can do and what programs can do for them, aiming to prevent actions that might compromise security [1]. To build a secure AC system, access control policies (ACPs) are essential. ACPs are often studied and designed based on AC models, such as Attribute-Based Access Control (ABAC), which is a well-known and widely recognized AC model. ABAC restricts access rights to organizations' critical information assets based on the attributes of subjects, objects, operations, and relevant context [2]. This model allows access decisions to change between access requests by altering attribute values without changing the subject-object relationships [3], which results in the capability to make dynamic AC decisions. While traditional AC models (e.g., the Role-Based Access Control (RBAC) model [4]) can no longer meet the fine-grained and expressive security requirements of modern applications, the ABAC model has emerged as a promising solution [6]. This AC model determines access authorization through policies,

offering a high degree of flexibility as it can adapt to contextual constraints and environmental conditions.

The execution of business process instances is typically recorded in an event log. The nonconformity between ABAC policies and ABAC implementations, as recorded in business process event logs (referred to as process logs), can lead to serious security issues, such as unauthorized actions by users going undetected. Thus, it is vital to ensure that the ABAC implementations within the process logs are consistent with the ABAC policies at the time of post-execution. To verify conformance between ABAC policies and their implementations within process logs, formalizing ABAC policies is necessary. The formalization is often carried out manually, leading to inefficiencies, errors, difficulties in error detection, and low maintainability since it is challenging for engineers not involved in the formalization to comprehend the content.

There are gaps in the literature on creating ABAC instances using the proposed AC metamodels. Kashmar *et al.* [7] reviewed the proposed AC metamodels in 2021, noting that while many AC metamodels have been proposed for RBAC instances creation, only one AC metamodel [8] was available for ABAC. However, this AC metamodel is limited to use within enterprise architecture frameworks and does not yet apply to software engineering. Also, only three studies utilizing the proposed AC metamodels to represent ABAC policies have been identified in the literature [10], [11], [12]. Therefore, proposing an ABAC metamodel to support the formalization and conformance testing of ABAC policies is essential. Moreover, significant research efforts have focused on mining ABAC policies from user-resource privilege relationships (e.g., AC lists and AC logs). However, there has been less attention to constructing ABAC policies from existing event logs. Event logs are the primary input for current process mining techniques, which effectively manage complex business processes in today's organizations [16]. Although process mining has been increasingly used for security analysis, its application in formalizing and conformance testing ABAC policies is notably absent in the literature.

This paper presents a fine-grained and highly automated model-driven workflow to address the problems and gaps mentioned above. There are three main contributions in this paper: (i) the proposal of an ABAC metamodel essential to create ABAC policy representation models for process logs;

(ii) proposing the ABAC-driven Object Constraint Language (*OCL*) invariant patterns (IPs) of the metamodel, which facilitates the formalization of ABAC policies; (iii) the successful application of the workflow to a real-world process log.

The rest of this paper is organized as follows. Section II presents related studies. Section III describes the workflow and the proposed ABAC metamodel in detail, as well as its ABAC-driven *OCL* IPs. A case study with a real-world process log is presented in Section IV. Section V discusses the applicability, effectiveness, maintainability, and limitations of the workflow, before concluding in Section VI.

II. RELATED STUDIES

The modeling languages such as Unified Modeling Language *UML* [18], *Petri nets* [19], and Business Process Model and Notation (*BPMN*) [20] are widely used to model business processes. This paper uses these modeling languages as the formats of model artifacts. Besides, *OCL* [21] is an 'add-on' standard to the *UML*, and a formal language used to describe expressions (define invariants) on *UML* models. The resulting language (referred to as *UML/OCL*) becomes significantly more precise when the *UML* is combined with the *OCL*. The modeling approach using *UML/OCL* was employed to facilitate the formalization and verification of ABAC policies [9]. However, the study is limited in generality and faces challenges related to contextual constraints.

Regarding *OCL* IPs (for convenience, *OCL* IPs are still referred to as IPs), a set of RBAC-driven IPs was proposed to support formalizing various types of RBAC policies [23]. Nevertheless, this set of IPs is specifically built for the RBAC model, not the ABAC model. A set of data-driven IPs was also proposed in [24] to facilitate rule checking (i.e., rule-based conformance checking, a type of process mining) of process logs and their process models. However, issues related to ACPs are not addressed in this study.

In the literature, several research efforts have been made to convert *BPMN* into *UML* [5], [13]. This paper utilizes the *TMBC* transformation method presented in [13] to automate the conversion of *BPMN* models to *UML* class diagrams. To clarify, this method identifies the source and target model elements, as well as the similarities and restrictions between the source and target models. It formalizes transformation rules using Model-Driven Architecture (MDA) and the ATLAS Transformation Language (*ATL*). In more detail, this method transforms the lanes, states, and nouns from the tasks within the input *BPMN* model into *UML* analysis classes and enumerations. Additionally, it encapsulates the tasks and data within the process specifications into the operations and attributes of the analysis classes, and converts the message flows into association relationships in the resulting *UML* class diagram.

Three studies using the proposed AC metamodels to specify ABAC policies have been identified in the literature [10], [11], [12]. Kashmar *et al.* developed a generic AC metamodel to specify and enforce ACPs covering ABAC [10]. The metamodel offers a solution to the issue of heterogeneity in concepts across various AC models. Nevertheless, this study does

not support a visual representation of ABAC implementation errors. Besides, formalizing ABAC policies is limited due to the grammatical deficiency of the formal language used in this study. Fernandez *et al.* proposed a formal specification of ABAC using the category-based AC metamodel to facilitate the design of ABAC policies [11]. However, this study focuses on specifying ABAC policies without considering ABAC conformance checks. This study is also limited in its ability to express various types of contextual constraints. Within this study, no patterns/templates are provided for the metamodel. In 2024, Zdilar proposed an ABAC metamodel to describe common AC models for spreadsheets (spreadsheets are among the most successful end-user programming systems in business and academia) [12]. This study is capable of determining the validity of performed user actions in multi-user environments. However, this study focuses on conceptual modeling of AC for spreadsheets without considering the formalization of ABAC policies as well as ABAC conformance checks.

III. FORMALIZING AND CONFORMANCE TESTING ABAC POLICIES FOR BUSINESS PROCESSES

A. Methodology

Figure 1 provides an overview of the workflow for formalizing and conformance testing ABAC policies. The core idea of the approach is to create an ABAC policy representation model from a business process event log and the proposed ABAC metamodel to facilitate the formalization of ABAC policies into *OCL* invariants and perform conformance checks of the formalized ABAC policies and their implementation within the process log. The workflow consists of the event log preprocessing stage and six tasks (i.e., tasks 1 to 6), in which task 1, task 2, task 6, and the procedure of generating *OCL* invariants in task 4 are performed automatically; part of the input for tasks 3 and 5 is carried out manually. The following describes the workflow in detail.

The event log preprocessing stage. The workflow begins by filtering the input event log using a plug-in available within a process mining tool, applying specific filtering criteria. In more detail, such a plug-in should be able to remove traces based on the frequency of events and/or the presence of specific start/end events (i.e., it can remove all traces that do not start with a specific event, end with certain events, and contain undesired event types). Then, we can mine a *Petri net* from the filtered log by employing an available *Petri net*-mining plug-in within the process mining tool; the output of this process is a *Petri net*. Besides, within the filtered log, we can select variants based on a particular criterion (e.g., focusing on variants with the most/least traces). Subsequently, depending on the needs of users, the traces in each selected variant to be checked will be chosen. The chosen traces serve as the input for task 5 in the workflow.

Task 1. Automatically convert a Petri net to a BPMN diagram. The input for this task is the *Petri net* mined at the event log preprocessing stage. Converting the *Petri net* to a *BPMN* model is then automatically performed using an

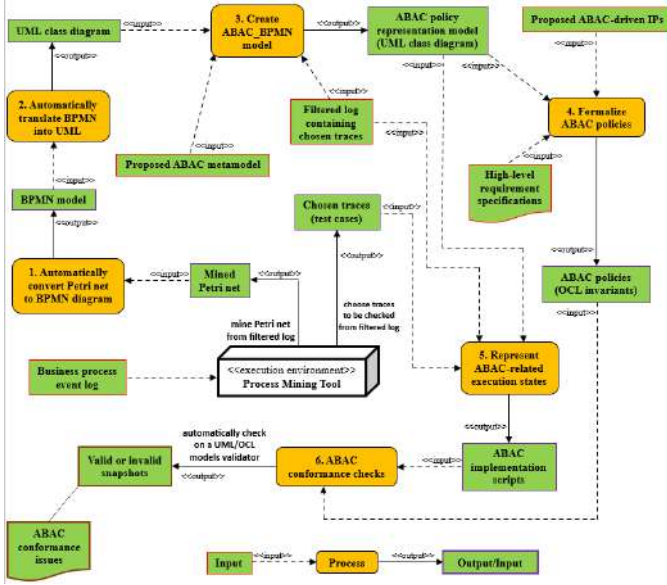


Fig. 1. The workflow for formalizing and conformance testing ABAC policies

available plug-in within the process mining tool. The output of task 1 is a BPMN model.

Task 2. Automatically translate BPMN into UML. This task inputs the BPMN model (i.e., the output of task 1). We can automatically translate the BPMN model into a UML model by utilizing an existing model transformation method in the literature. The output of task 2 is a UML class diagram. Also, some enumerations and additional attributes of the UML class diagram will be identified based on the filtered log.

Task 3. Create ABAC_BPMN model. This task inputs the UML class diagram (i.e., the output of task 2), the proposed ABAC metamodel (see Section III-B for details), and the filtered log containing the chosen traces. The output of task 3 is an ABAC policy representation model (referred to as an ABAC_BPMN model) in the form of a UML class diagram, aiming to formalize ABAC policies. As for creating an ABAC_BPMN model, the basic idea is that classes within the UML class diagram considered non-human resource classes will associate with the “Object” class of the ABAC metamodel, and the names of the human resource classes will convert to the values of the “Role” enumeration of the ABAC metamodel. In addition, the filtered log containing the chosen traces will specify new enumerations and additional attributes in the ABAC policy representation model. Adding the class diagram’s relevant UML descriptions to the ABAC metamodel’s UML specifications will obtain the UML specifications of the ABAC policy representation model.

Task 4. Formalize ABAC policies. This task inputs the ABAC policy representation model (i.e., the output of task 3), the proposed ABAC-driven IPs (see Section III-C for details), and high-level requirement specification documents. Most of today’s organizations have high-level requirement specification documents outlining who can access information and under what conditions [3]. The specification documents specify

security policies and include natural language access control policies (NLACPs) that determine permissible operations for a system [15]. Based on these inputs, ABAC policies will be formalized as OCL invariants on the UML representation of the ABAC policy representation model. Additionally, the UML/OCL specifications of the resulting ABAC policies will be the ABAC policy representation model’s UML specifications aggregating with the generated OCL invariants.

Task 5. Represent ABAC-related execution states. This task inputs the ABAC policy representation model (i.e., the output of task 3) and the chosen traces to be checked from the filtered log. By manually extracting all data related to ABAC from the filtered log containing the chosen traces, ABAC-related execution states in each of the chosen traces will be represented as ABAC implementation scripts (i.e., the output of task 5) by a specific language.

Task 6. ABAC conformance checks. This task inputs the ABAC implementation scripts (i.e., the output of task 5) and the UML/OCL specifications of the resulting ABAC policies in task 4. Task 6 is realized on a UML/OCL models validator. First, the UML/OCL specifications of the resulting ABAC policies will be imported into the validator. After that, the ABAC implementation scripts will be represented as snapshots in the form of UML object diagrams on the validator. Note that the UML object diagrams must conform to the respective UML class diagram (i.e., the ABAC policy representation model created in task 3). Checking the validity of the snapshots (i.e., determining whether they are valid or invalid) on the validator will be performed automatically, which can reveal ABAC conformance issues.

B. Proposed ABAC Metamodel

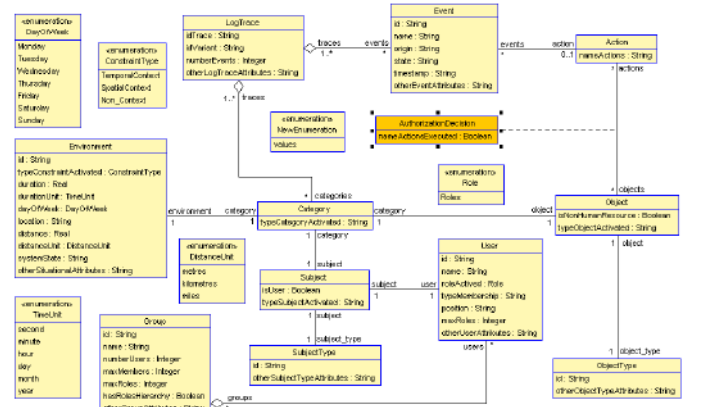


Fig. 2. The proposed ABAC metamodel

This paper introduces an ABAC metamodel that is specifically designed to create model instances for ABAC, as visualized in Fig. 2 in the form of a UML class diagram. The basic idea that forms the ABAC metamodel is that the three core concepts of the original ABAC model, “Subject,” “Object,” and “Environment,” are all considered types of the ‘Category’ term. Within the ABAC metamodel, these three core concepts

correspond to three *UML* classes that are indirectly linked to each other through the “Category” class.

As can be seen in Fig. 2, this metamodel comprises 11 conventional *UML* classes; the “AuthorizationDecision” is an associationclass in the relationship between the “Action” class and the “Object” class. Besides, the “Group” class has an aggregation relationship with the “User” class, and the “LogTrace” class has aggregation relationships with the “Category” and “Event” classes. An instance of the “LogTrace” class records a sequence of defined events, each event corresponding to a certain action (or no action) performed on a certain object (a non-human resource) by a certain subject (a user, an agency, or a process) according to specific contextual information (a certain location or a specific time interval). Additionally, the attributes of 11 *UML* classes are determined based on the relevant attributes commonly encountered in the literature.

In more detail, during the realization of task 3 in the workflow, several transformations are as follows. Note that if the “NewEnumeration” enumeration, the “Role” enumeration, the OTHERS, the “SubjectType” class, and the “ObjectType” class are not replaced after the transformations, they will not be included in the ABAC policy representation model.

- In an ABAC_BPMN model, the “SubjectType” class may be replaced with one or more *UML* classes involving the “Subject” class.
- In an ABAC_BPMN model, the “ObjectType” class will be replaced with the considered non-human resource classes within the *UML* class diagram (i.e., the output of task 2 in the workflow).
- In an ABAC_BPMN model, the “Roles” values of the “Role” enumeration in the ABAC metamodel may be replaced with the names of the human resource classes within the *UML* class diagram.
- In an ABAC_BPMN model, the “NewEnumeration” may be replaced with one or more new enumerations.
- The “otherSubjectTypeAttributes,” “otherObjectTypeAttributes,” “otherSituationalAttributes,” “otherLogTraceAttributes,” “otherUserAttributes,” “otherGroupAttributes,” and “otherEventAttributes” attributes represent the additional attributes, which may be identified by the filtered event log containing the chosen traces. We generally refer to these attributes as OTHERS.
- In an ABAC_BPMN model, the “nameActionsExecuted” Boolean attribute of the “AuthorizationDecision” associationclass will be replaced by the attributes with the <T>EXECUTED form, where <T> represents the names of the actions needed to check the execution permission. The actions correspond to the tasks/activities within the BPMN model (i.e., the output of task 1 in the workflow).

The “Environment” class aims to capture temporal, spatial, and environmental information about the considered business domain. With the support of the “TimeUnit” and “DayOfWeek” enumerations, the temporal context refers to the duration or day of the week during which a given action/task is restricted from execution. With the support of the

“DistanceUnit” enumeration, the spatial context refers to a geographical distance or specific bounded area from which a given action/task is restricted from execution. Also, the “timestamp” attribute of the “Event” class contains specific information about the date, month, year, hour, minute, and second at which an event was executed. These assist ABAC policy designers in grasping contextual information within process logs.

Within the ABAC metamodel, the “nameActionsExecuted” Boolean attribute of the “AuthorizationDecision” associationclass represents whether the corresponding action/task can be executed (i.e., permit or deny decisions). Specifically, if the value of the attribute is set to true, the corresponding action/task is permitted to execute; if the value of the attribute is set to false, the corresponding action/task is denied execution.

C. First Set of ABAC-driven IPs

To reduce the effort and time spent on formalizing ABAC security policies, enhance the consistency of formalized ABAC policies, and minimize potential errors that can occur when done manually, this paper introduces a set of three ABAC-driven IPs to formalize ABAC policies as *OCL* invariants uniformly. In general, the PresentAllTraces pattern is used to indicate that a task/action must be executed in all event log traces; the DenyExecution pattern is employed to deny the execution permission of a task/action based on the relationship between two attributes of a non-human resource class; the EnvironmentDenyUser pattern is utilized to deny the execution permission of a task/action based on the attributes of the two classes, “Environment” and “User.” The EnvironmentDenyUser pattern, the “Environment” class and its relevant enumerations within the ABAC metamodel (described in Section III-B) indicate the workflow’s “Expresses contextual constraint types” feature. The details of these three IPs are revealed in Figs. 3, 4, and 5.

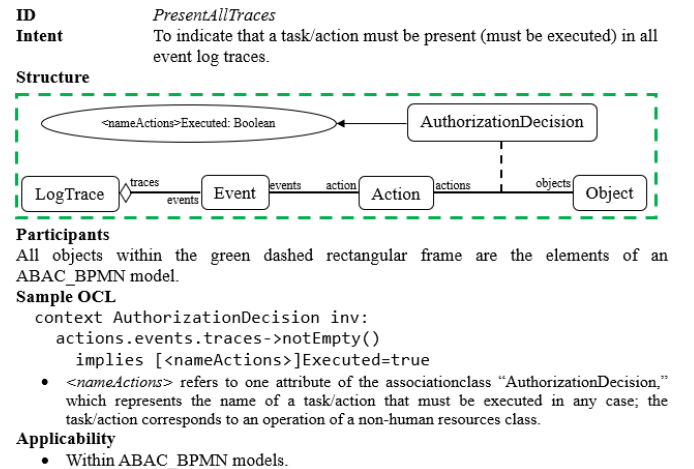


Fig. 3. The PresentAllTraces pattern

The following six components are employed to adequately grasp the characteristics of an IP.

- **ID**: a string to identify the considered IP,

- **Intent:** the purpose of using the IP,
- **Structure:** modeling a simplified relationship between participants in *UML*,
- **Participants:** semantic description of the *UML* elements appearing in the “Structure” component,
- **Sample OCL:** an *OCL* invariant consists of placeholders that may be replaced with actual *OCL* expressions,
- **Applicability:** the context in which the IP is used.

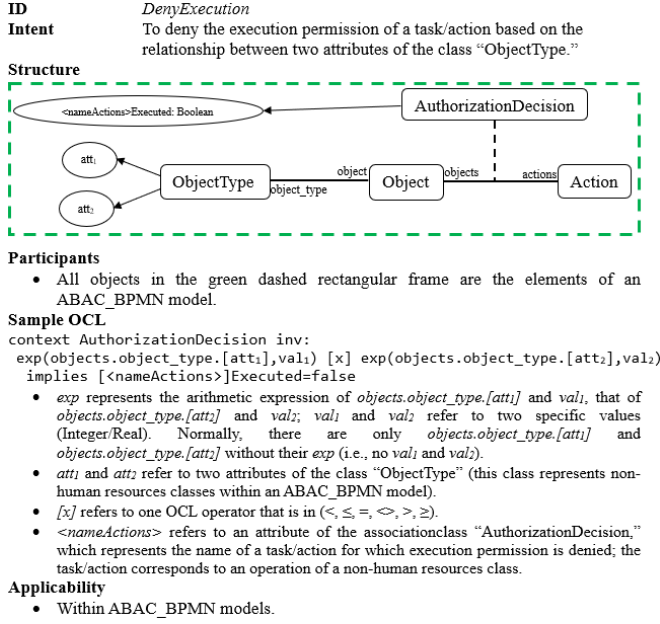


Fig. 4. The DenyExecution pattern

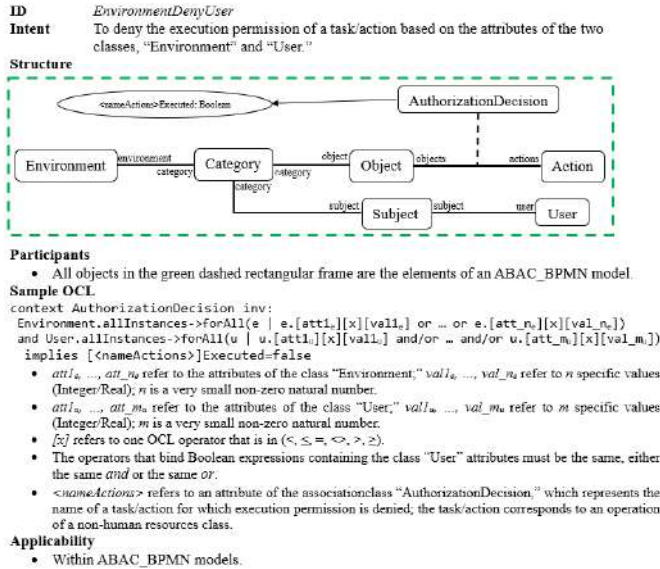


Fig. 5. The EnvironmentDenyUser pattern

Each IP’s “Sample OCL” component serves as a template for generating *OCL* invariants for ABAC policies defined in high-level requirement specifications. Each placeholder within

the “Sample OCL” component is expressed as [*a symbol*] and represents one of the following types:

- Names of *UML* classes involving non-human resources.
- Names of attributes of *UML* classes.
- *OCL* operators “[*x*]” that are in (<, ≤, =, >, ≥).
- Constants, attribute values, role names, task names, or data-related values when executing a process.

ABAC policies can be formalized as *OCL* invariants using these IPs. First, one IP and input data need to be selected, i.e., a list of pairs (*p*, *e*), where *p* is a placeholder in the “Sample OCL” component of the IP, and *e* is an expression of the same type as *p*. The input data must provide all placeholders appearing in the “Sample OCL” component of the IP. Then, the procedure automatically replaces each placeholder *p* in the “Sample OCL” component of the IP with *e*. As a result, a concrete description of an *OCL* invariant can be generated automatically. Please refer to page 168 in [23] for an illustrative example of how an *OCL* invariant is generated automatically.

IV. CASE STUDY

Loan application process’s event log. The workflow is applied to the “BPI Challenge 2017” dataset [25] (a publicly available event log is widely used in process mining research) of a real-world business process (the loan application process of a Dutch financial institute). The event log contains 31,509 cases and 1,202,267 events, with an average of 38 events per case. Please refer to page 176 in [23] for a detailed description of the log. Since the events within the log lack attributes that represent spatial context information and environmental conditions, the log is inappropriate for the EnvironmentDenyUser pattern.

Since the log needs to be audited against available ABAC policies, an NLACP document containing ABAC policies for the loan application process should exist. For illustration, two ABAC policies may be present within the NLACP document for the loan application process, as stated below.

POLICY 1. All applications must be created in a new application process (i.e., the “A_Create Application” event).

POLICY 2. If the monthly cost is less than one percent of the amount offered, the action to create the offer cannot be executed (i.e., the “O_Created” event is not present within the trace under consideration).

Experimental environment. In this paper, the *ProM* framework [14] (version 6.13) is selected as the process mining tool. The *USE* [22] (version 5.0), a powerful tool for generating *UML* models and validating *OCL* constraints, is chosen as the validator for *UML/OCL* models. The ABAC implementation scripts are expressed in the *soil* (i.e., simple *OCL*-based imperative programming language [17]) language.

Implementation and experimental results. The workflow begins by importing the “BPI Challenge 2017” dataset into the *ProM* framework. At the event log preprocessing stage, the *ProM* framework’s “Filter Log using Simple Heuristics” plug-in is utilized to filter the event log. Please refer to page 176 in [23] for a detailed description of filtering this event log and its results. Another available plug-in within the *ProM* framework,

“Mine Petri net with Inductive Miner,” is employed to mine a *Petri net*. The *Petri net* can be seen on page 177 in [23].

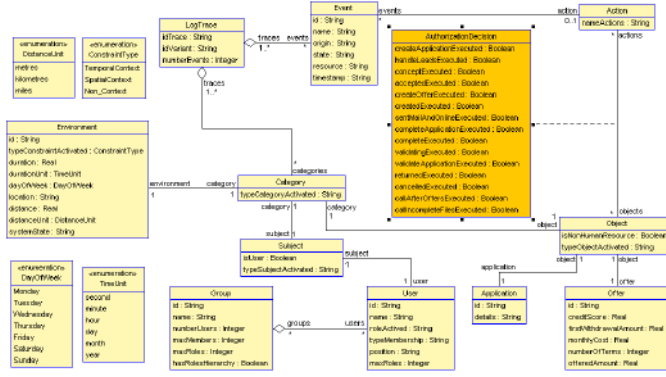


Fig. 6. The ABAC policy representation model

To illustrate task 1 in the workflow, the *ProM* framework’s “Convert Petri net to BPMN diagram” plug-in is utilized to automatically convert the *Petri net* into a BPMN diagram. The BPMN model can be seen on page 178 in [23]. To illustrate task 2 in the workflow, the *TMBC* transformation method (presented in Section II) is leveraged to automatically translate the BPMN model into a *UML* class diagram. The *UML* class diagram consists of two classes, “Offer” and “Application.”

Based on how task 3 in the workflow is realized, these two classes, “Offer” and “Application,” are placed in the “ObjectType” class to associate with the “Object” class of the proposed ABAC metamodel. Within the “Offer” and “Application” classes, the names of the operations (e.g., createOffer, createApplication) transfer to the “AuthorizationDecision” associationclass of the proposed ABAC metamodel as the Boolean attributes (e.g., createOfferExecuted, createApplicationExecuted). Additionally, the OTHERS, the “Role” (user roles are not defined in this event log) and “NewEnumeration” enumerations, and the “SubjectType” class are not included in the ABAC policy representation model, as shown in Fig. 6.

To illustrate task 4 in the workflow, the filtered log containing chosen traces should be audited against POLICY 1 and POLICY 2 above. These two ABAC policies are formalized as two *OC*L invariants by employing the *PresentAllTraces* pattern (see Fig. 3 for details) and the *DenyExecution* pattern (see Fig. 4 for details). This indicates the workflow’s “Formalizes ABAC policies using patterns/templates” feature. Figure 7 reveals the two generated *OC*L invariants using these two patterns, and indicates the workflow’s “Formalizes ABAC policies in a formal language” feature.

Regarding selecting variants to be checked, this paper considers all variants with the least trace numbers among the 32 variants in the filtered log, assuming that POLICY 1 and POLICY 2 violations may appear in such variants. As a result, there were 16 variants, each with only one corresponding trace. With this method of selecting variants, the number of traces that should be examined is 16 (corresponding to 16 test cases). Based on how tasks 5 and 6 in the workflow are realized (see Section III-A for details), the ABAC implementation scripts

// using the *PresentAllTraces* pattern

```
context AuthorizationDecision inv Policy1:
actions.events.traces->notEmpty()
implies createApplicationExecuted=true
```

// using the *DenyExecution* pattern

```
context AuthorizationDecision inv Policy2:
objects.offer.monthlyCost < objects.offer.offeredAmount/100
implies createdExecuted=false
```

Fig. 7. The two generated *OC*L invariants using the two ABAC-driven IPs

within these 16 traces are expressed in the *soil* language and represented as 16 snapshots on the *USE*. ABAC conformance issues within these 16 traces can be detected by checking whether the 16 snapshots are valid. This checking is performed automatically by the *USE* that indicates the last feature of the workflow in Table I.

Consequently, there was one test case that is non-fitting because its corresponding snapshot does not satisfy the “Policy2” *OC*L invariant, as revealed in Fig. 8. Specifically, the offer with ID “Offer_548379606” has a monthly cost (503.83) that is less than one percent of the offered amount (65000.0), while the “O_Created” event was still executed on “2016-03-19T22:50:02Z”. Therefore, the trace corresponding to the test case violates POLICY 2 above. Also, Fig. 8 illustrates the workflow’s “Detects and models ABAC implementation errors visually” feature. Additionally, the information of 149 users was recorded in the event log. This indicates the first feature of the workflow in Table I.

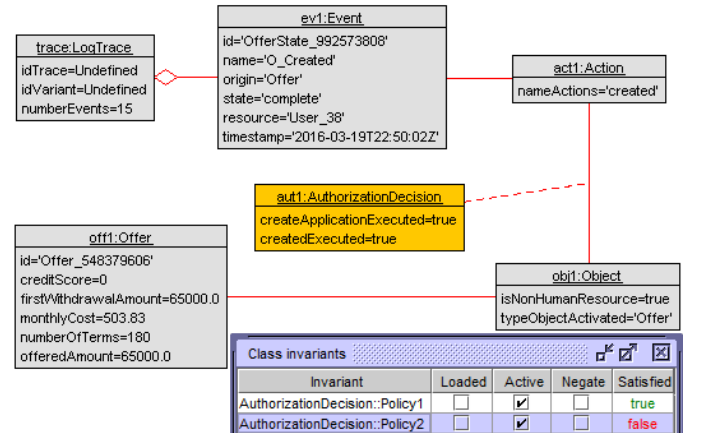


Fig. 8. Part of the invalid snapshot

V. DISCUSSION

Applicability. Thanks to the wide availability of open-source tools and existing model transformation methods, the workflow can leverage a process mining tool, a model transformation method, and a UML/OCL model validation tool without obstacles. Besides, the event log volume of the loan application process is sufficiently large and complex to demonstrate the validity of the workflow.

Effectiveness. As shown in Table I, the features offered by the workflow demonstrate the workflow's effectiveness. In addition, the proposed ABAC metamodel and its IPs enhance the consistency of the formalized ABAC policies.

TABLE I
RELEVANT RESEARCH ON ABAC POLICY SPECIFICATIONS USING THE
PROPOSED AC METAMODELS

Features	[10]	[11]	[12]	This paper
Determines the validity of performed user actions in multi-user environments			✓	✓
Formalizes ABAC policies in a formal language	–	✓		✓
Expresses contextual constraint types	✓	–		✓
Formalizes ABAC policies using patterns/templates				✓
Detects and models ABAC implementation errors visually			–	✓
Automatically checks the ABAC conformance				✓

The symbol ✓ indicates that the feature is available,
The symbol – indicates that the feature is limited.

Maintainability. Thanks to the *UML* specifications of the proposed ABAC metamodel, along with the well-defined IPs, ABAC policies obtained are highly maintainable, especially for policy engineers not involved in the formalization.

Limitations. The fact that unchecked traces may contain potential ABAC conformance issues (critical non-fitting behaviors). Besides, high-level requirement specification documents may be unavailable, making it challenging to adopt the workflow. Additionally, task 5 in the workflow requires considerable manual effort to accomplish.

VI. CONCLUSION

This paper presents a highly automated model-driven workflow for formalizing and conformance testing ABAC policies in process logs. The ABAC metamodel has been introduced to support this formalization and conformance testing. To reduce the effort and time spent on ABAC policy formalization and enhance the consistency of formalized ABAC policies, the first ABAC-driven IPs of the ABAC metamodel have been built. The approach utilizes the existing model-driven engineering means (specifically *Petri nets*, *BPMN*, *UML*, *OCL*, the *TMBC* transformation method, and metamodel) and the *ProM* and *USE* tools. The workflow assists policy engineers and security administrators in formalizing ABAC policies, and detecting ABAC policy violations in process logs.

In recent years, much research has been conducted on utilizing large language models (LLMs) like GPT-4 for analyzing sequential data, including logs. Applications to process logs have also begun to be explored [26]. Therefore, utilizing LLMs for task 5 in the workflow is a promising area for future work.

REFERENCES

- [1] R. S. Sandhu and P. Samarati, "Access control: principle and practice," *IEEE communications magazine*, vol. 32, no. 9, pp. 40–48, 1994.
- [2] E. Yuan and J. Tong, "Attributed based access control (ABAC) for web services," in *IEEE Int. Conf. on Web Services (ICWS'05)*, 2005.
- [3] V. C. Hu, D. Ferraiolo, R. Kuhn, A. R. Friedman, A. J. Lang, M. M. Cogdell, ... and K. Scarfone, "Guide to attribute based access control (ABAC) definition and considerations (draft)," NIST special publication, vol. 800, no. 162, pp. 1–54, 2013.
- [4] R. S. Sandhu, "Role-based access control," *Advances in computers*, vol. 46, pp. 237–286, 1998.
- [5] A. Leshob, "Towards a business-pattern approach for UML models derivation from business process models," in *2016 IEEE 13th Int. Conf. on e-Business Engineering (ICEBE)*, 2016, pp. 244–249.
- [6] A.K.Y.S. Mohamed, D. Auer, D. Hofer, and J. Kung, "A systematic literature review for authorization and access control: definitions, strategies and models," *Int. J. Web Inf. Syst.*, vol. 18(2/3), pp. 156–180, 2022.
- [7] N. Kashmar, M. Adda, M. Atieh, and H. Ibrahim, "A review of access control metamodels," *Procedia Computer Science* 184, 2021, 445–452.
- [8] M. Korman, R. Lagerstrom, and M. Ekstedt, "Modeling enterprise authorization: a unified metamodel and initial validation," *Complex Systems Informatics and Modeling Quarterly* 7, 2016, pp. 1–24.
- [9] D. H. Nguyen, Y. Sei, Y. Tahara, and A. Ohsuga, "A model-based approach for designing and validating ABAC policies," in *Software Engineering and Management: Theory and Applications*, vol. 17, R. Lee, Ed. Cham: Springer, 2025, pp. 19–36.
- [10] N. Kashmar, M. Adda, M. Atieh, and H. Ibrahim, "Access control metamodel for policy specification and enforcement: from conception to formalization," *Procedia Computer Science* 184, 2021, pp. 887–892.
- [11] M. Fernandez, I. Mackie, and B. Thuraisingham, "Specification and analysis of ABAC policies via the category-based metamodel," in *Proc. of the 9th ACM Conf. on Data and Application Security and Privacy*, 2019, pp. 173–184.
- [12] M. Zdilár, "Attribute based access control metamodel for spreadsheet programs," in *35th Int. Scientific Conf. CECIS 2024*, 2024, 409–416.
- [13] H. F. Boesing, S. N. Matos, V. C. Andrade, and E. Gueiber, "Using MDA to transform business processes from analysis classes," in *Proc. of the 16th Brazilian Symposium on Software Components, Architectures, and Reuse*, 2022, pp. 50–59.
- [14] B. F. Van Dongen, A. K. A. de Medeiros, H. M. Verbeek, A. J. M. M. Weijters, and W. M. van der Aalst, "The ProM framework: a new era in process mining tool support," in *Applications and Theory of Petri Nets 2005: 26th Int. Conf., ICATPN 2005, Proc.* 26, 2005, pp. 444–454.
- [15] M. Narouei, H. Khanpour, H. Takabi, N. Parde, and R. Nielsen, "Towards a top-down policy engineering framework for attribute-based access control," in *Proc. of the 22nd ACM on Symposium on Access Control Models and Technologies*, 2017, pp. 103–114.
- [16] W. M. Van der Aalst, "Process mining: a 360 degree overview," in *Process Mining Handbook*, W. M. P. Van der Aalst and J. Carmona, Eds. Cham: Springer, 2022, pp. 3–34.
- [17] F. Buttner and M. Gogolla, "Modular embedding of the object constraint language into a programming language," in *Formal Methods, Foundations and Applications: 14th Brazilian Symposium, SBMF 2011, Revised Selected Papers 14*, 2011, pp. 124–139.
- [18] G. Booch, I. Jacobson, and J. Rumbaugh, "The unified modeling language," *Unix Review*, vol. 14, no. 13, 5, 1996.
- [19] W. Reisig, *Petri nets: an introduction*, vol. 4. Springer Science & Business Media, 2012.
- [20] M. Chinosi and A. Trombetta, "BPMN: an introduction to the standard," *Computer Standards Interfaces*, vol. 34, no. 1, pp. 124–134, 2012.
- [21] M. Richters and M. Gogolla, "On formalizing the UML object constraint language OCL," *Int. Conf. on conceptual modeling*, 1998, pp. 449–464.
- [22] M. Gogolla, F. Buttner, and M. Richters, "USE: A UML-based specification environment for validating UML and OCL," *Science of Computer Programming*, vol. 69, no. 1-3, pp. 27–34, 2007.
- [23] D. H. Nguyen, Y. Sei, Y. Tahara, and A. Ohsuga, "Toward a pattern-based comprehensive framework using process mining for RBAC conformance checks," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 35, no. 2, pp. 157–194, 2025.
- [24] D. H. Nguyen, Y. Sei, Y. Tahara, and A. Ohsuga, "Data-driven OCL invariant patterns-based process model exploration for process mining," in *Networking and Parallel/Distributed Computing Systems*, vol. 18, R. Lee, Ed. Cham: Springer, 2024, pp. 117–135.
- [25] B. Van Dongen, *BPI Challenge 2017*, Eindhoven University of Technology, 2017. <https://doi.org/10.4121/UUID:5F3067DF-F10B-45DA-B98B-86AE4C7A310B>
- [26] A. Berti, H. Kourani, and W. M. van der Aalst, "PM-LLM-Benchmark: Evaluating large language models on process mining tasks," in *Int. Conf. on Process Mining*, 2024, pp. 610–623.

Group learning Integrating Behaviors and Emotions recognition to Assess Student Engagement in Classroom Context

Hung Q. Nguyen¹, Dat Ly¹, Nha Tran^{2,3,1*}, Tri Le¹, Hung Nguyen¹, and Hien D. Nguyen^{2,3}

¹Faculty of Information Technology, Ho Chi Minh City University of Education, Ho Chi Minh city, Vietnam

²Faculty of Computer Science, University of Information Technology, Ho Chi Minh City, Vietnam

³Vietnam National University, Ho Chi Minh city, Vietnam

Email: 4801104054@student.hcmue.edu.vn, 4901103017@student.hcmue.edu.vn, nhatt@hcmue.edu.vn, trilecs6102@gmail.com, hungnv@hcmue.edu.vn, hiennd@uit.edu.vn

* Corresponding author: Nha Tran (nhatt@hcmue.edu.vn)

Abstract—Group-based learning has become a cornerstone of modern education, fostering collaboration, critical thinking, and enhanced problem-solving skills among students. In this study, a deep learning framework, called *MoGL-Net*, is proposed. This framework integrates skeleton-based posture, sequential action features, facial emotion and classroom context to assess student engagement. Besides, a fusion strategy employing Self-Attention and Cross-Attention mechanisms is utilized to capture interdependencies between them. Additionally, this study also introduces a real-world dataset, *HCMUE-SEGL*, consisting of 1,663 video segments collected from 61 students participating in small-group learning activities. Experiments results show that *MoGL-Net* model achieves an accuracy of 90.70% on *HCMUE-SEGL*, highlighting the potential of multimodal learning for developing intelligent systems to monitor and objective classroom analytics in group learning environments.

Index Terms—Group learning, behavior recognition, Facial expression recognition, Classroom context

I. INTRODUCTION

In the current educational context, the teaching goal is to promote and comprehensively develop the qualities and competencies of learners. Accordingly, active teaching methods are encouraged to foster the initiative and engagement of students in learning tasks [1], [2]. Among these methods, group-based learning has been widely applied due to its ease of organization and superior effectiveness in enhancing educational quality. For these collaborative learning activities to be successful, the active and positive participation of students is essential [3], [4]. Therefore, teachers need to observe and assess the level of student participation in collaborative learning activities to provide timely adjustments and support. However, traditional methods such as teacher observations or self-assessments are often subjective and time-consuming. Thus, there is a need for an automated, objective evaluation framework that can analyze non-verbal signals such as behavior and facial expressions to provide timely and effective feedback [5], [6].

With the development of deep learning, many studies have applied it to assess student engagement in group learning. These models typically provide assessments based on the analysis of behavioral signals [7]–[9] or facial expressions [10], [11]. However, these methods have not fully utilized the relationship and interaction between behavior and emotions,

leading to an incomplete or inaccurate reflection of the actual engagement levels of learners [12]. To address this, multimodal methods are potential. They combine multiple information streams, including both behavioral and emotional features [13], [14]. However, most of these approaches still focus on analyzing individual behavior-emotion in traditional classroom settings and do not account for the contextual features of interactions between group members in group learning—an essential factor for reflecting group-level engagement.

This study proposes a multimodal deep learning model to assess individual student engagement in collaborative group learning environments, called *MoGL-Net*. The model integrates skeletal pose graphs, full-body RGB, facial expressions, and group context, using a two-stage attention mechanism to classify engagement into four levels: Disengaged, Normal, Engaged, and Very Engaged. Besides, a real-world multimodal dataset, the *HCMUE Student Engagement in Group Learning Dataset (HCMUE-SEGL)*, comprising 61 students with 1,663 annotated samples was constructed, capturing natural group learning interactions with synchronized pose, RGB, face, and context data. Experimental results demonstrated that integrating multimodal features significantly improves engagement classification accuracy, validating the effectiveness of the proposed approach.

II. RELATED WORKS

The assessment of learner engagement through artificial intelligence applications has become a significant area of research in recent years, especially within the growing context of digital classrooms and collaborative learning. One of the most common approaches is the analysis of behavioral cues through body posture and physical movements, observable via deep learning methods. The study in [8] proposed fine-tuning the C3D (3D CNN) model to learn spatio-temporal features of micro-body gestures to evaluate engagement levels in online learning environments. Despite achieving promising results, the absence of emotional analysis leads to an incomplete assessment of engagement. This is because a learner might appear engaged physically while actually feeling bored or anxious, which can negatively impact learning.

Therefore, facial expression analysis is emerging as a parallel research direction for inferring emotional states associated with engagement [15], [16]. The study in al. [16] developed a system to evaluate learner engagement in a classroom setting by analyzing facial expressions and head posture. Nevertheless, much like approaches that rely solely on behavior, a singular focus on facial expressions may fail to exploit the connection between behavior and emotion, which is essential for assessing the learner’s true engagement level.

Moreover, some studies have proposed multimodal approaches for assessing engagement. The studies in [17], [18] introduce models for real-time engagement monitoring in traditional classrooms by combining three modalities: facial expressions, head posture, and head movement. However, these methods exhibit shortcomings as they are designed exclusively for traditional classroom settings where learners work individually and thus ignore group-level interactions. Recent advances in deep learning for image analysis also emphasize architecture optimization and attention-based decoding strategies [19], but these remain confined to static image-level prediction and cannot capture the temporal, affective, and social dimensions that characterize real classroom interactions. In contrast, our method employs a hierarchical self- and cross-attention framework that fuses individual signals with classroom context. The context branch enables the model to capture interaction dynamics within the group, rather than treating engagement as an isolated individual phenomenon. By grounding individual signals in group-level context, our approach overcomes the limitations of unimodal or static methods and establishes a comprehensive foundation for engagement recognition in collaborative learning environments.

Although initial studies have contributed data on learner engagement [20], [21], these contributions are often unimodal, focusing primarily on a single aspect like behavior or emotion. To address this limitation, subsequent research has focused on constructing multimodal datasets [22], [23]. However, these datasets have predominantly concentrated on online learning environments. In this study, a new multimodal dataset for learner engagement in group activities is also constructed in both behavioral and emotional cues.

III. DATASET CONSTRUCTION

This section presents the building of a new multimodal dataset for assessing student engagement in group learning, called HCMUE-SEGL. This dataset is categorized into four levels: Very Engaged, Engaged, Normal, and Disengaged. These levels are constructed based on pairs of behavioral and emotional expressions as detailed in Table I.

The dataset includes nine types of behaviors (Standing, Raising hand, Discussing, Writing, Using Phone, Reading, Sleeping, Looking Around, Boring) and five basic emotions in a classroom environment (Surprise, Angry, Happy, Neutral, Sad). In which, the process of defining and grouping these behavior-emotion pairs into corresponding engagement levels was based on a survey conducted with 61 volunteers, the majority of whom were experienced high school teachers and university lecturers familiar with applying group-based learning methods. The volunteers were asked to select the most appropriate engagement level for each pair of expressions, and the level with the highest number of votes was chosen as the official label.

TABLE I: Description of Behavior and Emotion by Level of Engagement

Level of Engagement	Description of Behavior and Emotion
Very Engaged	Standing with an angry face, Standing with a happy face, Discussing with a surprised face, Discussing with an angry face, Discussing with a happy face, Raising hand with an angry face, Raising hand with a happy face, Writing with an angry face.
Engaged	Standing with a sad face, Standing with a neutral face, Raising hand with a sad face, Raising hand with a neutral face, Discussing with a sad face, Discussing with a neutral face, Writing with a happy face, Writing with a neutral face, Reading with an angry face, Reading with a happy face, Reading with a surprised face, Using phone with a happy face, Using phone with a surprised face, Using phone with an angry face.
Normal	Using phone with a neutral face, Using phone with a sad face, Reading with a sad face, Reading with a neutral face, Writing with a sad face.
Disengaged	Sleeping with a neutral face, Looking around with a neutral face, Boring with a neutral face.

The data collection process was conducted in various classrooms and at different times to increase the contextual diversity of the dataset. Recording over multiple days helped the learners’ emotions to be expressed more naturally, thereby enhancing the authenticity and objectivity of the data collected in the study. The dataset features the participation of 61 volunteers, with an average age of 21, as shown in Table II.

TABLE II: Demographic information of participants

Quantity	Age	Gender	Education level	Nationality
3	19	2 Male, 1 Female	Undergrad student	Vietnam
11	20	6 Male, 5 Female	Undergrad student	Vietnam
35	21	30 Male, 5 Female	Undergrad student	Vietnam
5	22	4 Male, 1 Female	Undergrad student	Vietnam
6	23	3 Male, 3 Female	Graduate student	Vietnam
1	26	1 Male	Undergrad student	Vietnam
Total: 61				

To protect privacy and personal information, all volunteers were thoroughly briefed on the procedures and proactively signed consent forms permitting the use of their images for research purposes. The cameras used for recording have a resolution of 720p and a frame rate of 30fps. The cameras were placed at the four corners of the table, as illustrated in Figure 1. This setup helps to clearly capture the learners’ behaviors and emotions from different perspectives.

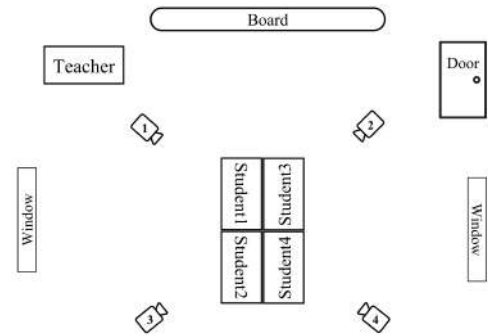


Fig. 1: Simulated diagram of volunteer seating arrangement and camera positions during filming.

The collected raw videos are first segmented into short

5-second clips. Next, these clips are processed through YOLOv11 [24] combined with the BoT-SORT algorithm to track each individual. From the frame sequence of each tracked subject, we manually filter and select 15 equally spaced frames. In cases where the subject's face is not clearly visible in a selected frame, we replace it with the nearest frame where the face is clear. The labeling process was carried out by 11 volunteers, with each data sample independently annotated by five individuals with teaching experience. Based on the analysis of the behavioral-emotional pairs in the 15-frame sequence, the annotators assigned an engagement level to each sample. The final label was determined through majority voting, which helps mitigate individual subjectivity. In addition, samples with inconsistent votes were re-checked and resolved through consensus, thereby enhancing annotation reliability and consistency.

Subsequently, this 15-frame sequence continues to be processed for feature extraction. The YOLO-Pose [25] framework is used to extract skeleton data, and RetinaFace [26] is used to detect and crop the facial region. Besides, the middle frame of the 15-frame sequence is extracted to represent the classroom context at that moment. After processing, the total number of collected samples is 1663. Detailed statistics of the samples are presented in Table III.

TABLE III: Statistics of the number of samples in the HCMUE-SEGL dataset

Category	Very Engaged	Engaged	Normal	Disengaged	Total
Number of samples	398	428	331	506	1663

IV. A MULTIMODAL NETWORK FOR GROUP LEARNING ENGAGEMENT ASSESSMENT - MOGLE-NET

This section presents MoGLE-Net, a deep learning model for estimating student engagement in group learning by integrating behavioral, emotional, and contextual features, as illustrated in Figure 2. The model processes 15-frame video sequences, where each branch contributes unique information: the Pose Skeleton branch, using SKE-A3TGCN [27], captures body posture and motion to reflect physical engagement; the Action RGB branch, based on ResNet50-LSTM, analyzes full-body actions to track active or passive participation; the Facial Emotion branch, using ViViT [28], extracts emotional expressions to gauge emotional engagement; and the Group Context branch, using ViT [29], captures spatial dynamics to incorporate group-level interaction.

A. Multimodal Feature Extraction

1) Pose-Based Behavioral Representation

The skeleton sequence $\mathcal{G} = \{G_t \in \mathbb{R}^{13 \times 2} \mid t = 1, \dots, T\}$, with $T = 15$, represents a 13-joint skeleton (Nose, Left Eye, Right Eye, Left Ear, Right Ear, Left Shoulder, Right Shoulder, Left Elbow, Right Elbow, Left Wrist, Right Wrist, Left Hip, Right Hip) per frame. Each frame $G_t = (V_t, E_t)$ is a graph where nodes $v \in V_t$ contain 2D coordinates for joint position and edges $e \in E_t$ encode distance and angles, consistent with the vertex and edge feature representations in the original paper [27]. The SKE-A3TGCN model processes $\mathcal{G}$ via spatial graph convolution:

$$H_t = \sigma \left(\hat{D}^{-1/2} \cdot \hat{A} \cdot \hat{D}^{-1/2} \cdot X_t \cdot W \right) \quad (1)$$

where X_t is the node feature matrix at time t , $\hat{D}^{-1/2}$ is the inverse square root of the degree matrix $\hat{D}$, used to normalize the adjacency matrix $\hat{A}$, ensuring balanced influence from nodes with different degrees, and W is a learnable weight matrix. The nonlinear activation function $\sigma(\cdot)$, typically ReLU, is applied after matrix multiplication to improve the model's ability to capture complex patterns. The resulting output, H_t , is the feature matrix of the graph convolution layer, containing the spatial features aggregated from neighboring nodes, representing the spatial relationships between joints at time step t . Temporal dynamics are captured via attention:

$$\alpha_t = \text{softmax}(q^\top \tanh(W_h H_t + b_h)) \quad (2)$$

$$\mathbf{F}_G = \sum_{t=1}^T \alpha_t H_t \in \mathbb{R}^{512} \quad (3)$$

where W_h , b_h are learnable parameters, q is a query vector, and α_t weights frame contributions, following the temporal attention mechanism. The embedding $\mathbf{F}_G$ captures motion and pose trends for engagement prediction, leveraging vertex and edge features.

2) Full-Body Action Encoding

The RGB action sequence $\mathcal{I} = \{I_t \in \mathbb{R}^{224 \times 224 \times 3} \mid t = 1, \dots, T\}$ is processed using a ResNet50-LSTM architecture to capture spatial and temporal motion patterns. Each frame I_t is passed through a ResNet50 backbone, extracting spatial features $f_t \in \mathbb{R}^{2048}$. The sequence $\{f_1, \dots, f_T\}$ is fed into a uni-directional Long Short-Term Memory (LSTM) network to model temporal dependencies:

$$h_t = \text{LSTM}(f_t, h_{t-1}) \quad (4)$$

where h_t is the hidden state at time step t . The output of the full-body action encoding branch is $\mathbf{F}_B \in \mathbb{R}^{512}$, which encapsulates dynamic full-body behavioral cues, leveraging spatio-temporal visual feature. This encoding enhances engagement recognition by integrating appearance-based motion information, thus providing a comprehensive understanding of the subject's physical involvement in the group learning activity.

3) Spatial-Temporal Facial Emotion Encoding

The facial expression sequence $\mathcal{F} = \{F_t \in \mathbb{R}^{224 \times 224 \times 3} \mid t = 1, \dots, T\}$, with $T = 15$ frame, is processed using the ViViT model to capture spatial-temporal facial features. Each frame F_t is cropped using RetinaFace, and the resulting face images are split into 3D tubelets, with each tubelet consisting of two consecutive frames and 16x16 pixel patches. These tubelets are projected into embeddings and passed through multiple Transformer layers.

The model applies Spatial Attention to capture relationships between patches within each frame and Temporal Attention to model expression dynamics over time. The embeddings are then aggregated via mean pooling across both spatial and temporal dimensions to produce the final facial emotion embedding $\mathbf{F}_E \in \mathbb{R}^{512}$. This embedding captures emotional intensity and variation, contributing to the engagement prediction by integrating dynamic facial expression information.

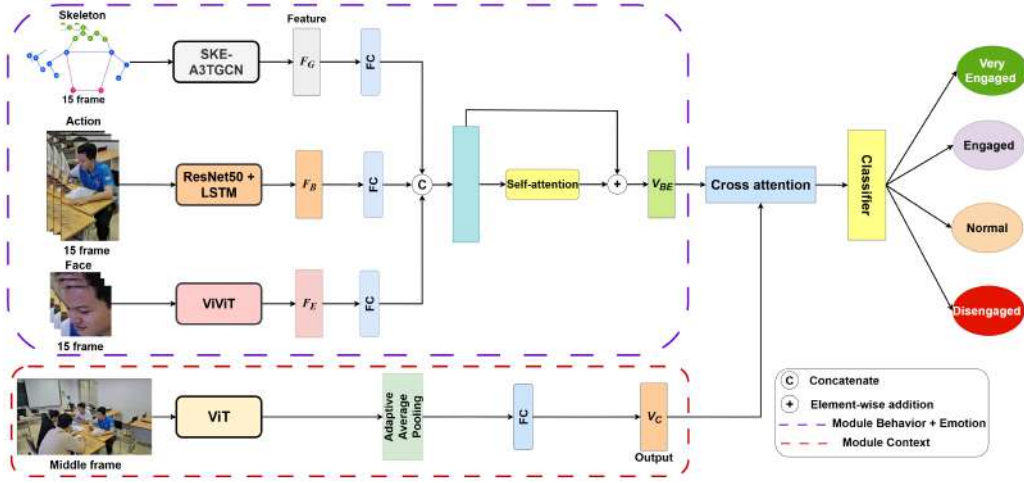


Fig. 2: The architecture of MoGLE-Net model.

4) Contextual feature

The context module processes a middle frame $C \in \mathbb{R}^{224 \times 224 \times 3}$ using the pre-trained ViT backbone. Unlike CNNs, ViT uses self-attention to capture global relationships across the entire frame, allowing it to better model group-level interactions and spatial context. This ability to focus on relevant areas, regardless of their spatial distance, enhances the understanding of group interactions in classroom settings.

The final context embedding $F_C \in \mathbb{R}^{512}$ is fused with behavioral and emotional features through cross-attention to improve engagement prediction.

B. Hierarchical Attention-Based Multimodal Fusion

To integrate unimodal embeddings from skeleton (F_G), action (F_B), and facial emotion (F_E) branches, we employ a hierarchical attention mechanism for engagement prediction. First, the unimodal embeddings are concatenated to form the behavior-emotion token set:

$$\mathbf{T}_{BE} = [F_G; F_B; F_E] \quad (5)$$

$$\mathbf{A}_{BE} = \text{SelfAttention}(\mathbf{T}_{BE}) \quad (6)$$

The self-attention mechanism is particularly effective when applied to features from the same data source, as it allows the model to learn internal dependencies between them. In this case, the pose, full-body, and facial emotion features are all expressions of the same individual's engagement, so their relationships are best learned within the same context. By applying self-attention, the model dynamically adjusts the importance of each modality based on its relevance at any given time, allowing for a more accurate representation of the person's engagement. After applying self-attention, the raw features are combined with the attended features through element-wise addition, followed by mean pooling across the tokens:

$$\mathbf{v}_{BE} = \text{MeanPool}(\mathbf{T}_{BE} + \mathbf{A}_{BE}) \quad (7)$$

This fusion process ensures that both raw and attended features contribute to the final behavior-emotion vector $\mathbf{v}_{BE}$,

which effectively captures the subject's behavioral and emotional state.

Group context is incorporated via cross-attention, with $\mathbf{v}_{BE}$ as the Query (Q) and the context sequence F_C as Key/Value (K/V):

$$\mathbf{v}_{fused} = \text{CrossAttention}(Q = \mathbf{v}_{BE}, K/V = F_C) \quad (8)$$

In this configuration, $\mathbf{v}_{BE}$ queries the context F_C to align individual behavior and emotion with group dynamics, such as interaction with classmate. By using cross-attention, the model adjusts the behavior-emotion features based on group context, enhancing context-aware predictions. The resulting fused vector combines individual and group-level cues, leading to more accurate and context-sensitive engagement estimation.

V. EXPERIMENTS AND RESULTS

A. Experiments setup

All experiments were carried out on the Google Colab Pro platform using NVIDIA A100 and L4 GPUs (22.5 GB VRAM each). The proposed model and all comparative baselines are trained for 30 epochs with a batch size of 4, optimizing via AdamW with an initial learning rate of $1e-5$.

B. Results and Analysis

To evaluate the MoGLE-Net architecture on the HCMUE-SEGL dataset, we conducted two experimental groups. The first, an ablation study, tested unimodal configurations including RGB stream (Full-body action), Pose stream (Pose-based behavioral), Facial stream (Spatial-temporal facial emotion) to assess individual stream performance. The second compared multimodal strategies, including dual-stream (RGB+Facial, Pose+Facial) and MoGLE-Net (RGB + Pose + Facial + Context), to demonstrate the benefits of integrated signals for engagement estimation. Results are summarized in Table IV, with confusion matrices in Figure 3.

The RGB and Facial streams achieved accuracies of 86.00% and 73.00%, with F1-scores of 83.00% and 72.00%, respectively (Table IV). Their confusion matrices (Figures 3a and 3b) show misclassifications between "Normal" and "Engaged," and between "Engaged" and "Very Engaged" because some

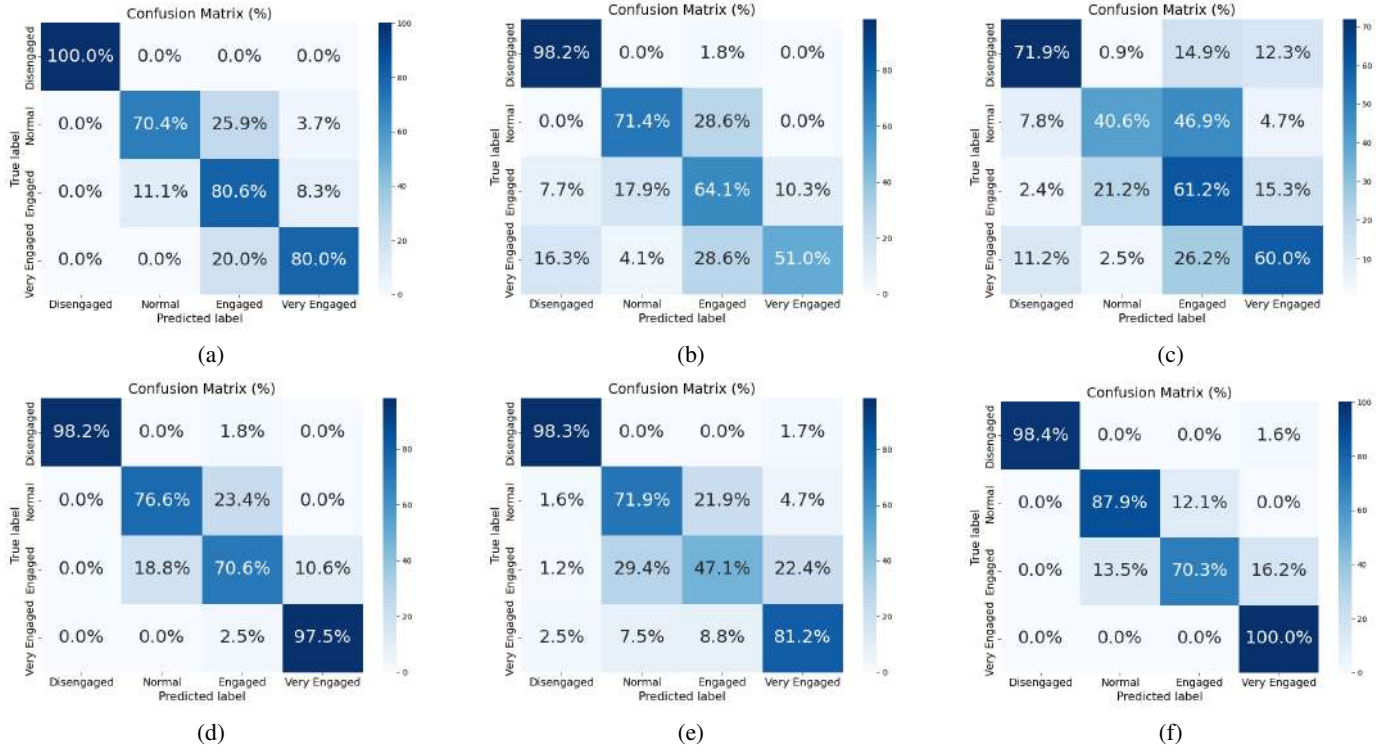


Fig. 3: Confusion matrices: (a) RGB Stream; (b) Facial Stream; (c) Pose Stream; (d) Action RGB + Facial Emotion; (e) Pose skeleton + Facial Emotion; (f) Action RGB + Pose skeleton + Facial Emotion + Context

behaviors and expressions appear in multiple engagement levels, making single-stream analysis insufficient. The Pose stream, with 60.64% accuracy and 58.92% F1-score (Figure 3c), struggles due to similar skeletal patterns in actions and lack of context.

TABLE IV: Performance comparison of different branches

Branch	F1	Prec.	Rec.	Acc.
RGB Stream	83.00%	84.00%	83.00%	86.00%
Facial Stream	72.00%	75.00%	72.00%	73.00%
Pose Stream	58.92%	60.97%	58.43%	60.64%
RGB Stream + Facial Stream	87.00%	87.00%	87.13%	87.13%
Pose Stream + Facial Stream	73.59%	73.73%	74.61%	76.74%
Action RGB + Pose Skeleton + Facial Emotion + Context	90.44%	90.83%	90.70%	90.70%

The two-modal configurations, Pose+Facial and RGB+Facial, improved performance over unimodal streams. Pose+Facial achieved 76.74% accuracy and 73.59% F1-score, surpassing Pose (60.64%, 58.92%) and Facial (73.00%, 72.00%). Similarly, RGB+Facial reached 87.13% accuracy and 87.00% F1-score, outperforming RGB (86.00%, 83.00%) and Facial alone. By combining skeletal or full-body motion with facial expressions, dual-modal models capture a holistic view of engagement. The synergy of behavioral cues and emotional signals effectively distinguishes subtle differences between levels like “Normal” and “Engaged,” unlike unimodal streams, which struggle with overlapping patterns. However, the confu-

sion matrix (Figure 3d and Figure 3e) shows misclassification between “Engaged” and “Normal” or “Very Engaged” due to the model’s focus on individual signals without capturing group interaction context, such as misclassifying a “happy” reader as “Engaged” instead of “Very Engaged” in a group discussion.

The full MoGLE-Net model (RGB+Pose+Facial+Context) achieved the highest performance with 90.70% accuracy and 90.44% F1-score, significantly reducing misclassification rates (Figure 3f). By integrating all modalities, it captures comprehensive behavioral, emotional, and contextual cues, effectively resolving ambiguities in engagement levels across unimodal and two-modal setups.

VI. CONCLUSION

This study proposed MoGLE-Net, a novel multimodal deep learning framework for assessing student engagement in group learning environments. By integrating behavioral, emotional, and contextual cues from various modalities, the model achieves an accuracy of 90.70%, demonstrating its effectiveness in predicting engagement levels. A HCMUE-SEGL dataset of 1,663 annotated video segments was constructed, enabling the model to process natural group learning interactions. The results highlight the advantages of using multimodal data, showing significant improvements over unimodal and two-modal approaches.

Despite these promising results, challenges remain when applying the model to real classroom data. Environmental factors such as lighting variation, occlusions, and spontaneous behaviors introduce noise and labeling ambiguity. In addition, generalization is limited: cultural differences shape how en-

gagement is expressed, and diverse classroom layouts alter behavioral and contextual cues, reducing robustness across educational settings.

In the future, the model will be expanded in applicability to diverse learning formats and integrate more granular facial feature analysis to improve sensitivity to subtle emotional states. Moreover, based on existing frameworks, future work should prioritize enhancing behavior recognition by moving beyond broad categories to capture micro-level behaviors and their social context. This involves developing models that can detect subtle actions like specific gestures, shifts in gaze direction, and purposeful fidgeting, which are more indicative of a student's cognitive state. Furthermore, it is crucial to analyze inter-student social dynamics to understand how behaviors influence one another, thereby capturing the essence of collaborative learning [30].

ACKNOWLEDGMENT

This research was supported by The Ho Chi Minh University of Education's Scientific Research Support Fund.

REFERENCES

- [1] Ministry of Education and Training, "Circular no. 32/2018/tt-bgddt on promulgating the general education program," Circular, December 26, 2018. [Online]. Available: <https://thuvienphapluat.vn/van-ban/Giao-duc/Thong-tu-32-2018-TT-BGDDT-Chuong-trinh-giao-duc-pho-thong-403454.aspx>, Dec. 2018, [Accessed: July 24, 2025].
- [2] S. Hoang, B. Nguyen, N. Nguyen *et al.*, "Enhanced task-based knowledge for lexicon-based approach in Vietnamese hate speech detection," in *Proceedings of 14th IEEE International Conference on Knowledge and Systems Engineering (KSE 2022)*, 2022, pp. 1–5.
- [3] B.-H. Duong and J. DeJaeghere, "From student-centered to competency-based reform: Exploring teachers' perspective of meaningful participation," *RISE working paper Series*, no. 22/089, 2022.
- [4] S. Feng, H. Chen, X. Ren *et al.*, "Collaborative group learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 7431–7438.
- [5] J. Zhao, H. Zhu, and L. Niu, "Bitnet: A lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, p. 101670, 2023.
- [6] S. Mandia, K. Singh, and R. Mitharwal, "Recognition of student engagement in classroom from affective states," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 18, 2023.
- [7] J. Liu, X. Mu, Z. Liu, and H. Li, "Human skeleton behavior recognition model based on multi-object pose estimation with spatiotemporal semantics," *Machine vision and applications*, vol. 34, no. 3, p. 44, 2023.
- [8] S. Khenkar and S. Jarraya, "Engagement detection based on analyzing micro body gestures using 3D CNN," *Computers, Materials & Continua*, vol. 70, no. 2, 2022.
- [9] Q. Tran, H. Nguyen, T. Huynh *et al.*, "Measuring the influence and amplification of users on social network with unsupervised behaviors learning and efficient interaction-based knowledge graph," *Journal of Combinatorial Optimization*, vol. 44, no. 4, pp. 2919–2945, 2022.
- [10] Y. Chen, J. Zhou, Q. Gao *et al.*, "MDNN: Predicting student engagement via gaze direction and facial expression in collaborative learning," *Computer Modeling in Engineering & Sciences (CMES)*, vol. 136, no. 1, 2023.
- [11] Z. Guo and R. Barmaki, "Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction," *Australasian Journal of Educational Technology*, vol. 36, no. 6, pp. 53–71, 2020.
- [12] A. Ben-Eliyahu *et al.*, "Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts," *Contemporary Educational Psychology*, vol. 53, pp. 87–105, 2018.
- [13] M. Hu, Y. Wei, M. Li *et al.*, "Bimodal learning engagement recognition from videos in the classroom," *Sensors*, vol. 22, no. 16, p. 5932, 2022.
- [14] H. Nguyen *et al.*, "Empowering classroom behavior recognition through hybrid spatial-temporal feature fusion," *Applied Intelligence*, vol. 55, no. 12, p. 863, 2025.
- [15] X. Tang, Y. Gong, Y. Xiao *et al.*, "Facial expression recognition for probing students' emotional engagement in science learning," *Journal of Science Education and Technology*, vol. 34, no. 1, pp. 13–30, 2025.
- [16] Ö. Sümer, P. Goldberg, S. D'Mello *et al.*, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1012–1027, 2021.
- [17] C. Pabba and P. Kumar, "A vision-based multi-cues approach for individual students' and overall class engagement monitoring in smart classroom environments," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 52 621–52 652, 2024.
- [18] A. Sabuncuoglu and T. M. Sezgin, "Developing a multimodal classroom engagement analysis dashboard for higher-education," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. EICS, pp. 1–23, 2023.
- [19] T. B. Nguyen-Tat, H.-A. Vo, and P.-S. Dang, "QMaxViT-Unet+: A query-based maxvit-unet with edge enhancement for scribble-supervised segmentation of medical images," *Computers in Biology and Medicine*, vol. 187, p. 109762, 2025.
- [20] W. Lu, Y. Yang, R. Song *et al.*, "A video dataset for classroom group engagement recognition," *Scientific Data*, vol. 12, no. 1, p. 644, 2025.
- [21] H. Nguyen, N. Tran, H. D. Nguyen *et al.*, "KTFEv2: Multimodal facial emotion database and its analysis," *IEEE Access*, vol. 11, pp. 17 811–17 822, 2023.
- [22] M. Singh, X. Hoque, D. Zeng *et al.*, "Do i have your attention: A large scale engagement prediction dataset and baselines," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 174–182.
- [23] C.-H. Wu, S.-y. Liu, X. Huang *et al.*, "Cmose: Comprehensive multi-modality online student engagement dataset with high-quality labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4636–4645.
- [24] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.
- [25] D. Maji *et al.*, "Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2637–2646.
- [26] J. Deng, J. Guo, E. Ververas *et al.*, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.
- [27] N. Tran, H. Nguyen, D. Ly *et al.*, "Advancing violence detection with graph-based skeleton motion analysis," *SN Computer Science*, vol. 6, no. 6, p. 595, 2025.
- [28] A. Arnab, M. Dehghani, G. Heigold *et al.*, "Vivit: A video vision transformer," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6816–6826, 2021.
- [29] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of The 2021 International Conference on Learning Representations (ICLR 2021)*, 2021.
- [30] P. Ta, N. Tran, H. Nguyen, and H. D. Nguyen, "Detecting signs of depression on social media: A machine learning analysis and evaluation," *Sustainable Futures*, p. 100827, 2025.

Vector-based Similarity for Personalized Recommendation Systems

Tran-Diem-Hanh Nguyen
College of Engineering and Technology
Tra Vinh University, Vinh Long Province, Vietnam
Email: diemhanh_tvu@tvu.edu.vn

Abstract—Personalized Recommendation Systems (RS systems) have been gaining popularity among researchers in data science as the systems help users search for products that satisfy their preferences. Likewise, the systems also provide benefits to companies who desire their products to be known to more future customers. Although a wide ranges of methods and techniques have been proposed to boost RS systems performances recently, the significant contributions of this kind of systems are always needed in this information age when more e-commerce platforms are available. In this article, two groups of methods including (1) basic prediction methods and (2) vector-based similarity methods combined with KNN are compared in this study. In particular, the basic prediction method employed Min, Max and Average of the user's ratings to predict the missing values. On the other hand, the vector-based methods employed the combination of KNN and vector-based similarity including Pearson and Cosine to predict the missing ratings. The experiments were carried out using the datasets MovieLens 100K. The evaluation metrics includes Mean Absolute Error (MAE), Cohen's Kappa metric (KAPPA), and Root Mean Square Error (RMSE) to compared the performances of the models. The vector-based similarity using Pearson similarity combines with KNN provided better performance than the other models for the metrics including RMSE and MAE.

Index Terms—Users-based Collaborative RS, personalized Recommender systems, e-commerce suggestion systems, Intelligent systems for E-commerce.

I. INTRODUCTION

RS systems are software programs that record user's preferences and search in the stored products to find items that are mostly satisfy with the users preferences. Such personalized RS systems are important because their significant contribution to users in their purchasing decisions from a wide range of products, ranging from news, foods, musics and clothes. In reality, the current increase in the number of online shoppers and products leads to the problem of searching for products that match with customer's preference both in computation and quality. Therefore, there is a need for more effective algorithms that can help improve the personalized RS performance. A RS model that can retrieve products that are preferred by users without using lots of memory and computation time. Generally, two main components, users' profiles and items' profiles should be modeled first for a typical recommendation system as they are main cores for the RS system. User's profiles are a set of user's information, which can be collected explicitly or implicitly, and items' profiles are a set of products. In

RS systems, modeling user's profiles and items profiles is important.

Once the two components are modeled, the user based approach is determined to develop RS systems. For collaborative-based approach CFbRS, user's profile information over the bought or viewed items is then used to recommend to new users. These preferences can be provided in either text or numeric ratings. On the other hand, Content-based RS systems (CbRS) employs a basic technique that utilizes item's attributes information and information about users preference over the items' attributes for recommending most suitable items to the corresponding users. For hybrid-based RS systems, the combination of two CFbRS and CbRS is employed to recommend most preferred items. Research in RS systems is continuing for different domains of information and good algorithms to enhance the recommendation performances in terms of more accurate rating predictions are needed to deal with large number of missing ratings.

This work compares the contribution of vector-based similarity methods including Cosine similarity and Pearson Similarity, and basic rating predictions to predict missing ratings. For doing this, two groups of RS models are proposed including (1) Basic rating prediction (2) Vector-based similarity based prediction for RS systems. Group (1) includes the models {Max_CFbRS, Min_CFbRS, Avg_CFbRS} that use *Max*, *Min*, *Avg* of available ratings respectively to predict the missing ones. Group (2) includes models that combine vector-based similarity including Cosine similarity and Pearson Similarity for missing rating predictions. The experiment results show that the model PS_CFbRS provides better results than the model CS_CFbRS and the basic rating prediction models including Min_CFbRS, Max_CFbRS and Avg_CFbRS according to the three main evaluation metrics. In other words, vector-based similarity combined with KNN in RS systems provides higher performance than the basic rating prediction methods.

The general information about this work is provided in Section I before Section II provides summaries of some existing works. The proposed models of personalized RS systems are proposed in Section III. Then, experiments are presented in Section IV, followed by the discussion about the assessment of recommendation performances. Section V presents the conclusions and future works.

II. RELATED WORKS

In short, RS systems can be divided into three main types: content-based, collaborative-based and hybrid-based RS systems. The collaborative-based RS systems utilize previous user's ratings over the items to predict missing ratings which might be not given by the users. Several works on collaborative-based RS systems can be listed as follows. To begin with, this collaborative-based method made use of word-of-mouth in advertisement, which was said to provide better suggestions than content-based systems. GroupLens was designed to recommend Usenet news to readers, aiming to obtain real ratings over a set of news from users and predict news that might satisfy the users' interest. Subsequently, a RS system for musics and movies is provided in [1]. This result originated from using information of items subjectively in recommendation phases through items and users representation.

In CFbRS systems, modeling user's interest and item's profiles is important to determine the success of the corresponding RS systems. For large-scale RS systems with a large number of items, the number of ratings given by users should be large enough for the systems to predict more accurate ratings for the missing ones. However, most users are not able to provide ratings over the large number of items. This leads to the problem of large number of items not obtaining ratings and items without any ratings. The cold-start problem is the term used for the case that items get no ratings from users. In other words, the cold-start problem occurs when users show no preferences over any products in the given database but these users need useful recommendations from the system. The term 'sparseness' is used to refer the case when items receive much less ratings from users. The problem of sparseness can be solved by employing matrix factorization. Similarly, the work in [2] added contextual information to the proposed model for boosting the performances of RS systems. The short review study in [3] which provides general information about the three types of RS systems. In particular, the work in [3] provided information about how to represent user's profiles and items profiles of a RS system; and how to generate predictions for missing ratings; hence solving the sparseness problem; and how to evaluate the RS systems.

III. THE PROPOSED MODEL

A. Vector-based representation of user's profiles

The user profiles are important in any RS system because they help to determine similar users and provide most satisfied recommendations. The information for building user's profiles can be collected implicitly or explicitly. Vectors of user's profiles can be organized in terms of a rating matrix $R(I_n, U_m)$ to describe the ratings of U_m on the item I_n . Similarities between user's interest provided in the rating matrix can be measured based on vector as follows.

B. Vector-based similarity

Definition 1: Cosine Similarity

Let $P = [p_1, p_2, \dots, p_i, \dots, p_n]$ and $Q = [q_1, q_2, \dots, q_i, \dots, q_n]$, $p_i, q_i \in [1, \dots, 5]$ be vectors of user's ratings over the set of n

items, the Cosine Similarity between the two vectors P and Q is denoted as $\mathbb{S}_{C,n}(P, Q)$ and measured as follows:

$$\mathbb{S}_{C,n}(P, Q) = \frac{\sum_{i=1}^n (p_i \times q_i)}{\sqrt{\sum_{i=1}^n (p_i)^2 \sum_{i=1}^n (q_i)^2}} \quad (1)$$

Where $p_i = P[i]$; $q_i = Q[i]$; and $\mathbb{S}_{C,n} \in R$

Definition 2: Pearson Similarity

Let $P = [p_1, p_2, \dots, p_i, \dots, p_n]$ and $Q = [q_1, q_2, \dots, q_i, \dots, q_n]$, $p_i, q_i \in [1, \dots, 5]$ be vectors of user's ratings over the set of n items, the Pearson Similarity between the two vectors P and Q is denoted as $\mathbb{S}_{P,n}(P, Q)$ and measured as follows:

$$\mathbb{S}_{P,n}(P, Q) = \frac{\sum_{i=1}^n (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^n (p_i - \bar{p})^2 \sum_{i=1}^n (q_i - \bar{q})^2}} \quad (2)$$

Where:

$p_i = P[i]$; $q_i = Q[i]$ and $\mathbb{S}_{P,n} \in R$
 $\bar{p} = \frac{1}{n} \sum_{i=1}^n P[i]$ and $\bar{q} = \frac{1}{n} \sum_{i=1}^n Q[i]$

C. Basic rating prediction methods

Definition 3: Min-rating prediction: Let $u_t = [x_1, x_2, \dots, x_i, \dots, x_n]$ be a vector comprising of the ratings of the user u_t over n items. $x_i = -1.0$ if x_i is the item without rating given by user u_t . The minimum predicted value for missing ratings for the item x_i is $r_{u_t}^m = \text{Min}\{x_1, x_2, \dots, x_i, \dots, x_n\}$; $1.0 \leq r_{u_t}^m \leq 5.0$.

Definition 4: Max-rating prediction: Let $u_t = [x_1, x_2, \dots, x_i, \dots, x_n]$ be a vector comprising of the ratings of the user t over n items. $x_i = -1.0$ if x_i is the missing value. The predicted maximum value for the item x_i in the vector u_t is denoted as $r_{u_t}^{m*}$ and calculated as $r_{u_t}^{m*} = \text{Max}\{x_1, x_2, \dots, x_i, \dots, x_m\}$; $1.0 \leq r_{u_t}^{m*} \leq 5.0$.

Definition 5: Average-rating prediction: Let $u_t = [x_1, x_2, \dots, x_i, \dots, x_n]$ be a vector comprising of the ratings of the user u_t over n items. $x_i = -1.0$ if x_i is the missing value. The predicted value for missing ratings for the item x_i is denoted as $r_{x_i}^a$ and calculated as $r_{x_i}^a = \frac{1}{n} \sum_{i=1}^n x_i \cdot I(p(x_i))$; $1.0 \leq r_{u_t}^a \leq 5.0, p(x_i) \in [1..5]$.

D. K-Nearest neighbor (KNN) for Collaborative RS systems

KNN method is common in collaborative-based RS systems for its easy calculations compared to other conventional methods. In particular, the problem of sparseness in user's rating can be solved by employing KNN. In this work, KNN was used in combination with vector-based similarity methods or basic-rating prediction methods to calculate the most similar users which are then used to predict the missing ratings values. Fig.1 illustrate 3 points which are most similar with the user u_{10} , u_m for general. This means that those three points can be used to recommend similar items for the user u_{10} .

E. Evaluation

Several conventional metrics of assessing a RS performance include Root Mean Square Error (RMSE), Cohen's Kappa and Mean Absolute Error (MAE). The detailed information of these metrics are provided as below.

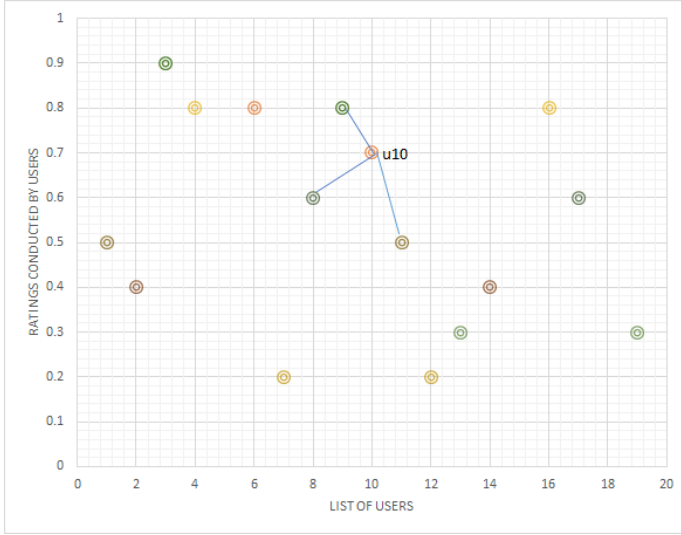


Fig. 1. Three most preference similarities with the user u_{10}

1) *Root Mean Square Error (RMSE)*: RMSE is a preferred metric to measure performances of RS systems, especially for the collaborative-based RS systems. The main principle of RMSE is to determine the errors between predicted rating and its real rating over the same item. Several RS systems use this metric to evaluate their performances as given in [4, 5].

$$RMSE = \sqrt{\frac{\sum_{j=1}^n (p_j - r_j)^2}{n}} \quad (3)$$

Where the predicted rating for item j^{th} is p_j ; The rating on item j^{th} is r_j ; and n is the number of test items

2) *Cohen's Kappa (KAPPA)*: Kappa metric for a set of n test samples is denoted as $\mathbb{K}$ and measured using this following equation.

$$\mathbb{K} = \frac{1}{n} \sum_{j=1}^n \frac{r_j - p_j}{1 - p_j} \quad (4)$$

Where p_j is the predicted rating for item j^{th} ; r_j is the real rating value of the item j^{th} ; n is the number of test items.

The Kappa revision is proposed as $\mathbb{K}^*$ to avoid the zero-division cases in Eq.(4).

$$\mathbb{K}^* = \frac{1}{n} \sum_{j=1}^n \frac{r_j - p_j + 6}{1 - p_j + 6} \quad (5)$$

3) *Mean Absolute Errors (MAE)*: The MAE metric determines the sum of errors between the predicted ratings and the real ratings. The MAE value over n test samples can be measured by using the following equation.

$$MAE = \frac{1}{n} \sum_{i=1}^n |p_j - r_j| \quad (6)$$

Where the predicted rating for item j^{th} is p_j ; the actual rating on item j^{th} is r_j .

In brief, the proposed model has provided some definitions for similarity measurement method based on vectors of user's profiles. The K-Nearest Neighbor was provided to calculate most similar users for a given user's profile. This section also gave information about how to evaluate performance of

RS systems. The next section is the experiments for comparing among methods that employ vector-based similarity and methods that employ basic rating predictions on the datasets MovieLens 100K.

IV. EXPERIMENTS

A. Experiment Models

The main aim of this section is to provide general information of the experiment models that employ KNN and the models that utilize the basic rating prediction methods as described in III-C. These models are presented as follows.

- **CS_CfBRS Model**: This model determines missing ratings using KNN and Cosine Similarity provided in Definition 1.
- **PS_CfBRS Model**: This model determines missing ratings using the combination of KNN and Pearson Similarity provided in Definition 2.
- **Min_CfBRS Model**: This model predicts missing ratings by replacing them with minimum of all provided ratings, which is provided in Definition 3.
- **Max_CfBRS Model**: This model predicts missing ratings by replacing them with maximum of all given ratings, which is provided in Definition 4.
- **Avg_CfBRS Model**: This model predicts missing ratings by replacing them with average value of all provided ratings, which is provided in Definition 5.

B. Experiment data and evaluation

The MovieLens datasets was first created in 1997 as in [6] and it has been upgraded since then. This datasets comprises of a list of users and a list of movies. Users ratings on the movies following five-scale ratings, from 1 to 5. MovieLens 100K including 100000 ratings from 943 users over 1682 movies is used in the following experiments.

Three major metrics were used including RMSE, MAE and KAPPA to compare performances of the experiment models.

C. Experiment Results

Rating statistics per item: The entire MovieLens datasets about users, items and user's ratings can be displayed in figures Fig.2, Fig.3, Fig. 4, and Fig.5. In each figure, the horizontal axis is the movie ID and the vertical axis is the number of ratings over the corresponding movie.

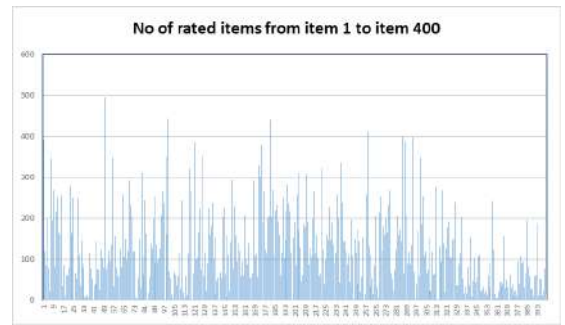


Fig. 2. Ratings on items from #1 to #400

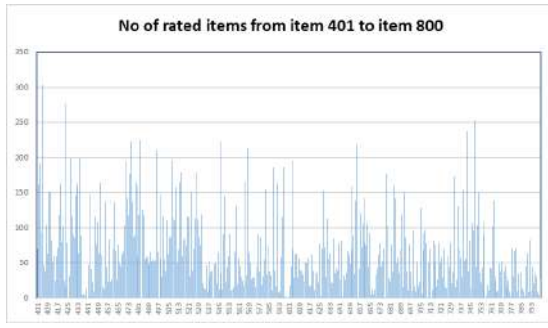


Fig. 3. Ratings on items from #401 to #800

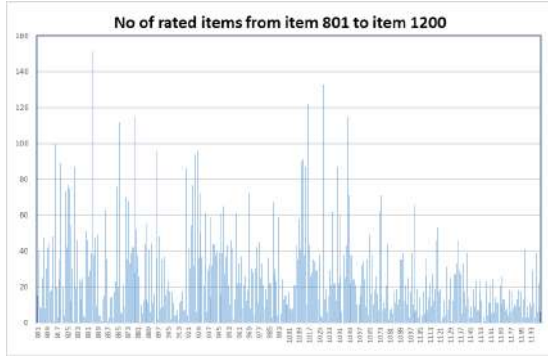


Fig. 4. Ratings on items from #801 to #1200

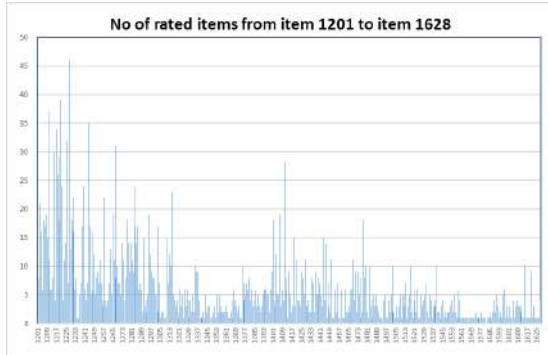


Fig. 5. Ratings on items from #1201 to #1628

Rating statistics per user: The entire MoivesLen datasets about users, items and user's ratings can be displayed in figures Fig.6, Fig.7. The horizontal axis is the user ID and the vertical axis is the number of ratings correspondingly. From the figures, it is obvious that users gave different rates over the items in the movie list in all two figures.

Figure Fig.8 provides overall statistics of users who gave ratings over items in the experiment datasets. According to the figure, a high proportion of users gave less ratings over the items.

Experimental results: Experiment results of the two groups of models including basic rating prediction methods with Min_CfBRS, Max_CfBRS, and Avg_CfBRS and vector-based similarity methods with CS_CfBRS and PS_CfBRS are pre-

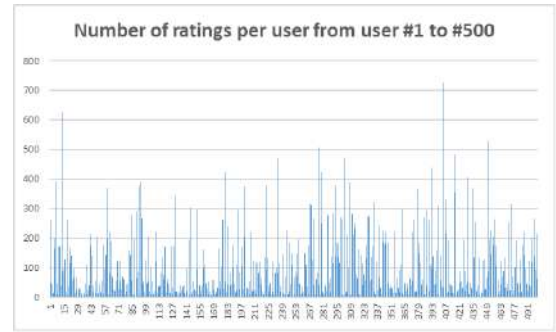


Fig. 6. Ratings given by users from #01 to #500

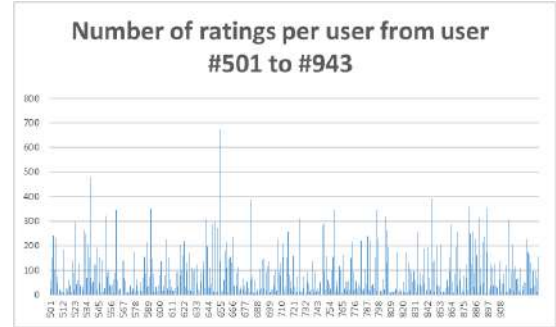


Fig. 7. Ratings given by users from #501 to #943

sented in Table I. The performance evaluation metrics includes RMSE, KAPPA and MAE.

TABLE I
COMPARISON OF EXPERIMENT RESULTS OF THE PERSONALIZED RECOMMENDATION SYSTEMS

Models	RMSE	KAPPA	MAE
PS_CfBRS	0.8668	1.3293	0.6924
CS_CfBRS	0.9781	1.3631	0.7510
Min_CfBRS	2.6556	1.1980	2.4002
Max_CfBRS	1.8516	1.5050	1.4701
Avg_CfBRS	1.2441	1.2777	1.0017

Table I shows that the model PS_CfBRS obtained better results than the other models including CS_CfBRS in RMSE. These experiment results indicate Pearson similarity is better than Cosine Similarity in similar rating measurements, with 0.8668 and 0.9781 respectively. Similar to the RMSE metric, the model PS_CfBRS provided better performance than other models in MAE, with 0.6924 in PS_CfBRS and 0.7510 in CS_CfBRS. It is noted that the model Min_CfBRS provided better performance than other models in KAPPA metric, with 1.1980 and the worst model is CS_CfBRS with 1.3631 in KAPPA.

D. Discussions:

- The sparseness in user's ratings should be discussed here first. As mentioned in the introduction section, this work is a user-based RS system, so the number of user's ratings over the provided items mainly determine performances of the RS

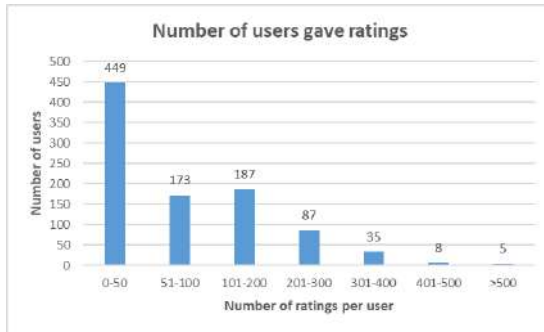


Fig. 8. Statistics of users who gave ratings

systems. As can be seen in the figure, Fig.8, the statistical number of user's ratings and the number of items are provided with totally 943 users, who conducted ratings over 1682 items. Nearly half of the users conducted less than 50 ratings, about 3.0% of the total number of items. This information indicates that the ratings for the entire items given by users are really sparse.

- The methods of missing-rating predictions can be discussed as follows. The first remark for RS systems performances is that more biased prediction values decrease performance of RS systems. For minimum based prediction, the missing values are predicted biased because the predicted values are applied for all missing values while the real missing values can be higher than the prediction values. This issue leads to a decrease in the system performance as can be seen in the Table I. Likewise for maximum based prediction, the biased problem significantly decrease the performance. Similarly, the performance of Avg_CFBRS is better than Min_CFBRS and Max_CFBRS owing to the average based prediction model which provides lesser biased prediction values.

- Finally, the third issue is the normalized rating prediction with vector-based similarity methods. In the model CS_CFBRS, cosine similarity method was used to measure how difference among user's preferences is. This model is different to the basic rating prediction methods because of the removal of biased factor in the cosine similarity method. Similarly, the model PS_CFBRS employed Pearson Similarity approach to measure the user's similarities. This model provided higher performance than the basic rating prediction methods as can be seen in the Table I due to the use of removing biased factors.

V. CONCLUSIONS AND FUTURE WORKS

Personalized RS systems are definitely beneficial for both companies and customers with the e-commerce platforms boom, a wide ranges of products, and an increasing number of online shoppers. In fact, this kind of systems is significantly contributing to users decision on purchasing preferred products and help more companies not only to save advertising budgets but also to get more prospective customers. This paper has provided some benefits of the collaborative-based approach in RS systems and general aspects for building a collaborative

RS system based on users ratings. In addition, this work has also provided several methods of rating predictions including basic ratings predictions and the combination of KNN and vector-based similarity methods to deal with the issue of missing ratings for providing suitable suggestions. Finally, the experiments were conducted to compare performances of both basic rating predictions and the combination of KNN and vector-based similarity methods, using the datasets MovieLens 100K. The results of above experiments shows that the vector-based similarity methods used for missing rating predictions provide higher performances than the basic rating prediction methods.

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest

ACKNOWLEDGMENT

This work is not under a funded project. The author is grateful for the supports of registration fee and travel expense by Tra Vinh University. Finally, the author appreciates reviewer's and anonymous reader's comments.

REFERENCES

- [1] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.
- [2] Hiep Xuan Huynh, Nghia Quoc Phan, Nghi Mong Pham, Van-Huy Pham, Le Hoang Son, Mohamed Abdel-Basset, and Mahmoud Ismail. Context-similarity collaborative filtering recommendation. *IEEE Access*, 8:33342–33351, 2020. doi: 10.1109/ACCESS.2020.2973755.
- [3] Tran-Diem-Hanh Nguyen. Recommendation systems: A short review. *International Journal of Innovative Research in Computer and Communication Engineering*, 12 (10.15680/IJIRCCE.2024.1205004), 2023.
- [4] Joseph A Konstan, Bradley N Miller, David Maltz, Jonathan L Herlocker, Lee R Gordon, and John Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.
- [5] Rishabh Ahuja, Arun Solanki, and Anand Nayyar. Movie recommender system using k-means clustering and k-nearest neighbor. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 263–268. IEEE, 2019.
- [6] F Maxwell Harper and Joseph A Konstan. The movie-lens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Flame Reavers@ALQAC 2025: Integrating Learned Rankers and LLM Reasoning in a Dynamic Hybrid Architecture for Legal Retrieval

Huy Trieu^{1,2}, Dang-Phuong-Nam Doan^{1,2}, Anh-Kiet Nguyen^{1,2},
Thanh-Thai Nguyen^{1,2}, Thanh-Nghia Vo^{1,2}, Tung Le^{1,2}, Huy Tien Nguyen^{1,2,*}

¹Faculty of Information Technology, University of Science, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

{tghuy22, ddpnam22, nakiet22, ntthai22, vtnggia22}@clc.fitus.edu.vn

{littung, ntienhuy}@fit.hcmus.edu.vn

*Corresponding Author: Huy Tien Nguyen (ntienhuy@fit.hcmus.edu.vn)

Abstract—Legal document retrieval remains a fundamental challenge in legal information systems due to the syntactic complexity and domain-specific semantics of legal texts. In this work, we introduce a novel hybrid retrieval architecture that attained promising performance in the 2025 Automated Legal Question Answering Competition (ALQAC). Our method reframes hybrid retrieval optimization as a pairwise learning-to-rank task, effectively eliminating the need for manual weight tuning by leveraging supervised training on relevance-labelled data. Furthermore, we introduce a reason-and-extract prompting scheme for large language model (LLM)-based reranking, which mimics hierarchical legal reasoning to extract justifiable answer candidates. To further enhance robustness, we incorporate a dynamic reranking controller that adaptively selects between a cross-encoder and the LLM-based reasoning module based on confidence thresholds. Experiments on the ALQAC benchmark demonstrate that our approach surpasses existing baselines, highlighting the promise of combining probabilistic retrieval with structured LLM reasoning to advance legal document retrieval.

Index Terms—Information Retrieval, Legal Case Retrieval, Learning-to-rank, Prompting Strategy.

I. INTRODUCTION

The rapid expansion of Vietnamese legal corpora, including statutes, codes, and case law, poses major challenges for information retrieval due to complex language, formal syntax, and dense cross-references. Manually identifying relevant articles is labor-intensive and inconsistent, limiting both scalability and reliability. These challenges highlight the urgent need for intelligent legal information systems to automate tasks such as document retrieval and question answering, thereby improving efficiency and accuracy.

In response to this demand, the Automated Legal Question Answering Competition (ALQAC) [1] has emerged as a key benchmark to advance research in legal AI for Vietnamese law. The 2025 edition introduces a newly curated dataset comprising manually annotated statutory texts, with the goal of fostering the development of accurate, generalizable, and reproducible legal information systems.

ALQAC 2025 features two primary subtasks: Legal Document Retrieval (Task 1) and Legal Question Answering

(Task 2). Task 1 focuses on retrieving all legal articles that are relevant to a given question, where relevance is defined as the ability of an article to provide sufficient information to answer that question. Task 2 then requires systems to generate answers in dynamic evaluation test-True/False, multiple choice, or free-text-based on the articles identified in Task 1. Performance is evaluated using the macro-average F2-score for Task 1 and accuracy for Task 2.

In this paper, we introduce our winning solution to both subtasks of the ALQAC 2025 benchmark. Our method is aimed to handle the issues of legal document retrieval and question answering in Vietnamese legal texts by merging traditional retrieval models, rich semantic representations, and big language model reasoning inside a unified hybrid architecture. For Task 1 (Legal Document Retrieval), we integrate sparse methods (BM25L [2] and TF-IDF [3]) with dense Vietnamese embeddings in a hybrid retrieval pipeline. Fusion weights are optimized via a pairwise learning-to-rank framework, replacing manual tuning. This backbone supports a two-stage reranking process: first, a fine-tuned Cross-Encoder with hard negative mining; second, a Qwen/Qwen3-8B [4] LLM reranker using structured multi-step reasoning and title-aware prompting. A hybrid reranking strategy dynamically selects between the Cross-Encoder and LLM outputs based on calibrated confidence thresholds to optimize efficiency and accuracy.

For Task 2 (Legal Question Answering), we construct a prompt-based answering module that also utilizes Qwen/Qwen3-8B [4]. The module is carefully engineered to handle different question formats, including true/false classification, multiple-choice selection, and open-ended generation. Each prompt includes strict output format constraints and explicit instructions for grounding responses in the retrieved legal articles. This design minimizes hallucination and enforces legal consistency across diverse question types.

Our team, Flame Reavers, achieved the top ranking in both subtasks of ALQAC 2025, with an F2-score of 0.9435 for legal document retrieval (top 1) and an accuracy of 0.9634

for legal question answering (top 5). These findings highlight the usefulness of our hybrid, reasoning-aware architecture in solving challenging legal AI problems.

II. RELATED WORKS

A. Legal Document Retrieval

The field of Legal Information Retrieval (LIR) has evolved significantly, moving from traditional lexical models to sophisticated neural architectures. Foundational baselines in LIR have long been established by lexical models such as BM25 and TF-IDF, which depend on phrase frequency. Many successful systems in past competitions, the NeuralMind [5], demonstrated the power of a well-implemented BM25, sometimes augmented with heuristic-based ranking. Similarly, the TUWBR [6] affirmed that traditional IR methods provide highly competitive results, which they augmented with external domain knowledge.

The paradigm shifted with the advent of Transformers [7], leading to a dominant two-stage retrieve-and-rerank architecture. In this setup, an efficient lexical model generates a candidate set, which is then re-ordered by a more computationally intensive neural model. The DoSSIER [8] investigated the effectiveness of BM25 as a first-stage retriever followed by BERT-based rerankers like BERT-PLI [9]. This approach of using paragraph-level interactions within BERT was a notable step, as seen in models like BERT-PLI, though it faced challenges with long-distance dependencies. Others, like the NOWJ [10] explored combining BM25 scores directly with outputs from models designed for longer texts, such as Longformer [11].

To improve upon this architecture, researchers have explored more advanced integration strategies. Multi-stage pipelines, such as the one developed by the JNLP [12] team, employ sequential filtering and re-ranking phases to refine the result set. Critically, the concept of systematically combining features from different retrieval methods using Learning-to-Rank has proven highly effective. The winning TQM team [13] at COLIEE 2024 [14] leveraged an LTR method on a feature set combining lexical matching and dense vector retrieval, highlighting the power of a data-driven fusion approach. Our work builds directly on this philosophy, but we formulate the weight optimization as a distinct pairwise ranking problem to fine-tune the synergy between sparse and dense signals.

More recently, Large Language Models (LLMs) have been integrated into LIR pipelines. Their application ranges from summarizing case documents to simplify retrieval, as explored CAPTAIN [15], to being used as end-to-end rerankers. Our approach distinguishes itself by creating a dynamic, hybrid reranking system that uses a traditional Cross-Encoder’s confidence score to selectively delegate difficult queries to an LLM, leveraging the LLM’s advanced reasoning capabilities in a targeted and efficient manner.

B. Legal Question Answering

Legal Question Answering (LQA) aims to generate accurate answers from legal texts based on user queries, typically

formulated as a binary classification or Natural Language Inference (NLI) task. While early methods relied on rule-based systems, recent approaches are dominated by neural models, especially BERT-based architectures. Top-performing systems combine powerful pretrained models (e.g., DeBERTa, RoBERTa) with data-centric techniques such as augmentation, entity inference, and relevant-span selection. For example, the KIS team enhanced BERT with rule-based logic and entity-aware augmentation [16], while HUKB employed selective article span extraction and augmentation strategies [17]. These efforts highlight that model performance in LQA depends not only on architecture, but also on careful data design and context grounding.

For Multiple-Choice Question Answering, another sub-task in LQA, research has emphasized enhancing reasoning capabilities and leveraging contextual evidence. Some approaches utilize structured representations or retrieval-augmented pipelines to compare candidate answers systematically, while others apply attention-based strategies to dynamically prioritize key segments of legal text, improving the model’s ability to distinguish among closely related options.

We propose a structured prompting framework that guides a pre-trained language model (Qwen3-8B) for accurate, context-aware legal question answering. Task-specific prompts with strict output formats ensure consistency across true/false, multiple-choice, and free-text responses. To reduce hallucinations, the model is instructed to ground its reasoning in retrieved legal articles. This strategy delivers scalable, high-precision LQA without domain-specific fine-tuning, offering a practical and economical solution for legal AI systems.

III. PRELIMINARY

A. Task Definition

Task 1 is a legal document retrieval task. Given a legal question q and a corpus of n legal articles $C = \{a_1, a_2, \dots, a_n\}$, the objective is to retrieve a set of relevant articles $A^* = \{a_i \mid a_i \in C \wedge \text{relevant}(a_i, q)\}$. The predicate $\text{relevant}(a_i, q)$ holds true if and only if the query q can be answered using knowledge given inside article a_i .

Task 2 subsequently requires participants to provide a direct answer to the question q , using the ground-truth relevant articles provided by organizers after the completion of Task 1.

B. Dataset

The dataset for ALQAC 2025 is a manually annotated collection derived from Vietnamese statute laws. It is structured to support both retrieval and question-answering tasks. The dataset is split into a public training set and a private test set. Both sets contain three distinct question formats designed to evaluate different aspects of legal reasoning: True/False, Multiple Choice, and free-text Essay questions. The statistical distribution of these question types is detailed in Table I.

IV. METHODS

The overall architecture of our system consists of four main components: (i) a hybrid retriever that integrates sparse

TABLE I
DISTRIBUTION OF QUESTION TYPES IN THE ALQAC 2025 DATASET

Question Type	Train Set	Test Set
True/False	387	11
Multiple Choice	286	60
Essay (Free-text)	56	11
Total	729	82

and dense retrieval methods, (ii) a fine-tuned Cross-Encoder reranker with hard negative mining, (iii) an LLM-based reranker with structured reasoning, and (iv) a dynamic hybrid controller that adaptively switches between the two rerankers. This general architecture is illustrated in Figure 1, which provides a high-level overview of how these components interact. In the following subsections, we describe each component in detail.

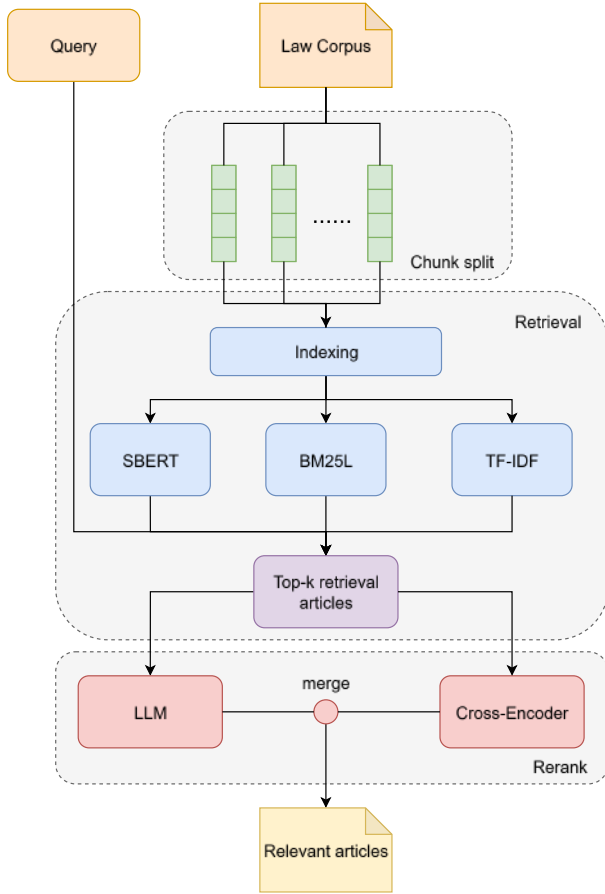


Fig. 1. An overview of the proposed hybrid retrieval and reranking architecture for Task 1.

A. Legal Document Retrieval

Our approach begins with the systematic ingestion of legal documents into a multi-modal retrieval system. The legal corpus consists of Vietnamese legal documents structured as a collection of laws, each containing multiple articles with associated textual content.

Given the substantial length of legal articles, we employ a fixed-size chunking strategy to optimize retrieval granularity while maintaining semantic coherence. Each legal document is systematically divided into overlapping chunks to ensure comprehensive coverage while preserving contextual relationships across chunk boundaries.

For the retrieval phase, our retrieval system constructs three complementary indices to capture different aspects of textual similarity. We employ BM25L [2] as our primary sparse retrieval mechanism. The BM25L score for a document d given query q is computed as:

$$\text{BM25L}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{\left(\frac{f(t, d)}{1 - b + b \cdot \frac{|d|}{\text{avgdl}}} + \delta \right) \cdot (k_1 + 1)}{k_1 + \frac{f(t, d)}{1 - b + b \cdot \frac{|d|}{\text{avgdl}}} + \delta} \quad (1)$$

where $f(t, d)$ is the term frequency of term t in document d , $|d|$ is the document length, avgdl is the average document length, and k_1 , b , and δ are tuning parameters.

For vector representation, we first process the Vietnamese text using the `underthesea`¹ tokenizer with lowercase normalization. Subsequently, we employ the `AITeamVN/Vietnamese_Embedding` model [18, 19] to encode each document chunk into a fixed-dimensional vector space:

$$\mathbf{e}_i = \text{Encoder}(d_i.\text{content}) \in \mathbb{R}^{1024} \quad (2)$$

The embedding model is truncated to 1024 dimensions to balance computational efficiency with representational capacity. Semantic similarity between query q and document d_i is computed using cosine similarity:

$$\cos\_sim(q, d_i) = \frac{\mathbf{e}_q \cdot \mathbf{e}_i}{|\mathbf{e}_q| \cdot |\mathbf{e}_i|} \quad (3)$$

To enhance the effectiveness of both sparse and dense retrieval components, we additionally construct a TF-IDF matrix [3], using the same tokenizer as applied in BM25L to ensure consistency in term segmentation. The TF-IDF score for term t in document d within corpus $\mathcal{D}$ is computed as:

$$\text{TF-IDF}(t, d, \mathcal{D}) = \text{TF}(t, d) \cdot \text{IDF}(t, \mathcal{D}) \quad (4)$$

where:

$$\text{TF}(t, d) = \frac{f(t, d)}{\sum_{t' \in d} f(t', d)} \quad (5)$$

$$\text{IDF}(t, \mathcal{D}) = \log \frac{|\mathcal{D}|}{|\{d \in \mathcal{D} : t \in d\}|} \quad (6)$$

Our hybrid retrieval combines the three retrieval modalities through weighted score fusion. For a given query q , the final relevance score for document d_i is:

$$\begin{aligned} \text{Score}(q, d_i) = & \alpha \cdot \text{BM25L}_{\text{norm}}(q, d_i) \\ & + \beta \cdot \text{TF-IDF}_{\text{norm}}(q, d_i) \\ & + \gamma \cdot \text{Embedding}(q, d_i) \end{aligned} \quad (7)$$

¹<https://github.com/undertheseanlp/underthesea>

where α , β , and γ are weight parameters constrained to sum to one, and the subscript “norm” indicates min-max normalized scores:

$$\text{Score}_{\text{norm}} = \frac{\text{Score} - \text{Score}_{\min}}{\text{Score}_{\max} - \text{Score}_{\min}} \quad (8)$$

Hybrid Retriever Weight Optimization. To determine the optimal parameters (α, β, γ) for the score fusion model in Equation 7, we developed a systematic, data-driven optimization procedure. The core principle of our optimization is to find a weight vector $W = (\alpha, \beta, \gamma)$ that, for each query, maximizes the cumulative score of relevant documents while simultaneously minimizing the score of non-relevant documents.

Our objective is to configure the weights such that the final computed score of a relevant document is consistently higher than that of any non-relevant document. Furthermore, the framework aims to keep the scores of documents that are both relevant to the same query clustered closely, preventing large, arbitrary score disparities between them.

To achieve this, we employ a grid search over the parameter space. For each potential combination of weights, we evaluate its effectiveness on the training dataset by measuring how well it satisfies our ranking objective. The weight combination that results in the most effective separation between relevant and non-relevant documents across all queries is selected as the optimal set of parameters for our hybrid retriever.

B. Traditional reranking

Our reranking approach is based on fine-tuning a pre-trained Cross-Encoder model using a domain-specific dataset. The process involves two main steps: data preparation through hard negative mining and model training.

To construct a robust training dataset, we first implemented a hard negative mining strategy. We employed a hard negative mining technique utilizing the AITeamVN/Vietnamese_Embedding bi-encoder [18, 19] to compute semantic embeddings for all legal articles in our corpus and for each training query. For every query, we identified the top 20 most semantically similar articles using cosine similarity (Equation 3). After removing known positive articles from this list, the remaining ones were designated as “hard negatives.” This technique provides challenging examples that force the model to learn nuanced differences between relevant and merely similar documents.

In the second stage, we fine-tuned the AITeamVN/Vietnamese_Reranker model [18, 19]. The model was trained on pairs of (query, article) where positive pairs (relevant articles) were labeled 1.0 and negative pairs (hard negatives) were labeled 0.0. For the fine-tuning process, we set the learning rate to 1e-5 and trained the model for a total of 3 epochs. During training, we monitored its performance on a held-out validation set, measuring Top-1 Accuracy—the percentage of queries where the highest-scoring document was a true positive. The model checkpoint that achieved the best Top-1 Accuracy was selected as our final reranker.

At inference time, the fine-tuned reranker operates as the second stage of a two-stage retrieval pipeline. For any given query, an initial fast retrieval model first identifies a set of top-10 candidate articles. Our Cross-Encoder then ingests these candidates, individually scoring each (query, article) pair. The final ranked list is generated by re-sorting these 10 articles in descending order based on the computed relevance scores, thereby leveraging the reranker’s high-precision capabilities.

C. Large Language Models reranking

To achieve a higher degree of precision, we incorporated a sophisticated reranking pipeline driven by a Large Language Model, specifically Qwen/Qwen3-8B [4]. In this stage, the LLM is tasked with analyzing up to 10 of the top candidate articles provided by our initial retrieval stage. The model performs a deep, multi-step reasoning process to evaluate the relevance of each candidate article relative to the user’s query.

This process goes beyond simple keyword matching to assess the core legal concepts, semantic alignment, and contextual fit of each article. The objective is to have the LLM distinguish truly relevant articles from those that are merely topically similar, identifying a definitive subset of articles that are most vitally important for answering the question. Following this reasoning phase, the LLM then extracts the identifiers of the selected articles and formats them into a strict JSON object for reliable downstream processing.

D. Hybrid Reranking Strategy

Our final system employs a hybrid reranking strategy to synthesize the outputs of both pipelines. We use the relevance score from the traditional Cross-Encoder as a confidence metric. For any query where the top-ranked article’s score falls below an empirical threshold of 0.8, we substitute the entire result with the high-precision output from our LLM pipeline. The threshold of 0.8 was selected based on our observation of the rerank scores across validation queries, where this value consistently provided the best trade-off between efficiency and robustness. This approach leverages the Cross-Encoder’s efficiency for high-confidence predictions while delegating more ambiguous cases to the LLM’s superior reasoning capabilities.

E. Legal Question Answering

Our approach to the question-answering task leverages tailored prompting strategies designed for each of the three distinct question types, namely: true/false, multiple-choice, and open-ended. For this task, we also employed the Qwen/Qwen3-8B large language model [4]. The core of our method involves constructing a specific system prompt and user prompt for each format. In all cases, the model is provided with the question and the full text of the relevant legal articles retrieved in the previous stage. This ensures the model’s reasoning is grounded strictly within the provided legal context, preventing hallucinations and improving factual accuracy.

The prompt design for each question type is highly specialized to constrain the model’s output and guide its reasoning.

For true/false questions, the system prompt instructs the model to act as an intelligent assistant whose sole task is to determine if a statement is True or False and to output only that single word. For multiple-choice questions, the prompt is similarly restrictive, demanding a response consisting of only the single character (A, B, C, or D) that corresponds to the correct answer, explicitly forbidding any additional explanation. In contrast, the prompt for open-ended questions instructs the model to synthesize a concise, short answer based on the provided legal articles, encouraging it to extract and formulate a direct response rather than just classifying information. This structured prompting methodology was crucial for generating accurate and correctly formatted answers, aligning the model’s capabilities with the specific requirements of each sub-task.

V. EXPERIMENTS AND RESULTS

A. Legal Document Retrieval

We evaluated our methods on the competition dataset using standard Information Retrieval (IR) metrics. Our primary evaluation metric is F2-macro, which is favored for placing greater weight on Recall, a critical aspect in the legal domain to avoid overlooking relevant documents [1]. It is calculated as the average F2-score across all queries, as defined in Equation 9.

$$\text{F2-macro} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \left(\frac{5 \times \text{Precision}_i \times \text{Recall}_i}{4 \times \text{Precision}_i + \text{Recall}_i} \right) \quad (9)$$

For a more comprehensive assessment, we also report Mean Average Precision (mAP) and R-Precision. Their formulations are provided in Equations 10 and 11.

$$\text{mAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}_q \quad (10)$$

$$\text{R-Precision}_i = \frac{r_i}{R_i} \quad (11)$$

In these equations, $|Q|$ denotes the total number of queries. For a given query i , Precision_i and Recall_i are the typical performance measurements based on the number of retrieved and relevant documents, as described in [1]. AP_q is the Average Precision for a single query q , r_i is the number of relevant documents retrieved within the top R_i results, and R_i is the total count of known relevant documents for that query.

TABLE II
TASK 1: EVALUATION ON THE PRIVATE TEST

Team	Methods	F2-macro	mAP	R-Precision
Ours	Our-LLM reranker	0.9435	0.9512	0.9329
	AITeamVN-LLM reranker	0.9388	0.9390	0.9085
	LLM reranker	0.9081	0.9390	0.9207
Others	vund	0.8970	-	-
	UIT legalGrep	0.8970	-	-
	NOWJ	0.8766	-	-
	UTEHY-NLU	0.8482	-	-

We evaluated our methods on the competition’s private test set, with comprehensive results presented in Table II. Our

final hybrid system with fine-tuned reranker established a new state-of-the-art, achieving an F2-macro of 0.9435. This score significantly surpasses not only the baseline models but also its constituent components when used in isolation. The results strongly validate our prompt engineering strategy. Our fine-tuned Cross-Encoder proved superior to the established baseline AITeamVN reranker, demonstrating the effectiveness of our hard negative mining technique.

B. Legal Question Answering

The principal evaluation metric for this work is accuracy, computed using the formula provided below. It is important to note that responses to essay questions undergo manual evaluation by experts to assess their correctness.

$$\text{Accuracy} = \frac{(\text{number of correct predictions})}{(\text{total number of instances})} \quad (12)$$

TABLE III
BASELINE PERFORMANCE ON THE TRAINING SET

Type	Accuracy	Count
True/False	0.8966	387
Multiple Choice	0.9301	286
Essay	1.0000	56
Weighted Average	0.9177	729

For Task 2, our approach leverages a large language model (LLM) with tailored prompting, which does not require a fine-tuning phase on the provided dataset. This training-free nature allowed us to use the training set directly as a development testbed to establish a performance baseline. The results, detailed in Table III, show a strong initial accuracy across all question types.

TABLE IV
TASK 2: EVALUATION ON THE PRIVATE TEST

Team	Accuracy	Rank
NOWJ	0.9878	1
Hallucineers	0.9878	1
UIT-Darkcow	0.9878	1
UTEHY-NLU	0.9756	4
FlameReavers (Ours)	0.9634	5
vund	0.9634	5
BAMBOOVN	0.9512	7
ABCD	0.9390	8

Table IV presents the official results on the private test set. Our system correctly answered 79 out of 82 questions. This performance corresponds to a final accuracy of 96.34%, securing 5th place in the competition standings for this task.

VI. CONCLUSION

In this study, we introduced a dynamic hybrid architecture intended to meet the difficult issues of legal document retrieval and question answering. Our approach introduces a data-driven framework to optimize an initial hybrid retrieval by

synergizing sparse and dense methods. This foundation supports a novel, two-stage reranking pipeline that dynamically switches between a fine-tuned Cross-Encoder for efficient, high-confidence predictions and a Large Language Model, guided by a “reason-and-extract” prompting strategy, for deep, hierarchical reasoning on more ambiguous cases.

The exceptional performance of our system in both competition tasks validates this hybrid strategy, proving that a principled architecture combining probabilistic retrieval with advanced LLM reasoning can substantially advance the state-of-the-art. This work establishes a powerful blueprint for future legal AI systems, with promising directions including the refinement of the dynamic switching mechanism and the exploration of end-to-end optimization for the entire retrieval and reasoning pipeline.

ACKNOWLEDGMENTS

This research is supported by research funding from the Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

REFERENCES

- [1] N. H. Thanh, B. M. Quan, C. Nguyen, T. Le, N. M. Phuong, D. T. Binh, V. T. H. Yen, T. Racharak, N. L. Minh, T. D. Vu, P. V. Anh, N. T. Son, H. T. Nguyen, B. Butr-indr, P. Vateekul, and P. Boonkwan, “A summary of the alqac 2021 competition,” *arXiv preprint arXiv:2204.10717*, 2022, summary of ALQAC 2021 competition.
- [2] A. Trotman, A. Puurula, and B. Burgess, “Improvements to bm25 and language models examined,” in *Proceedings of the 19th Australasian Document Computing Symposium*, 2014, pp. 58–65.
- [3] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.
- [4] Q. Team, “Qwen3 technical report,” *arXiv preprint arXiv:2505.09388*, 2025, latest version of Qwen family models.
- [5] G. M. Rosa, R. C. Rodrigues, R. Lotufo, and R. Nogueira, “Yes, bm25 is a strong baseline for legal case retrieval,” 2021. [Online]. Available: <https://arxiv.org/abs/2105.05686>
- [6] T. Fink, G. Recski, W. Kusa, and A. Hanbury, “Statute-enhanced lexical retrieval of court cases for coliee 2022,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.08188>
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [8] A. Abolghasemi, S. Althammer, A. Hanbury, and S. Verberne, “Dossier@ coliee2022: dense retrieval and neural re-ranking for legal case retrieval,” in *Sixteenth international workshop on Juris-informatics (JURISIN)*, 2022.
- [9] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, and S. Ma, “Bert-pli: Modeling paragraph-level interactions for legal case retrieval,” in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 3501–3507, main track. [Online]. Available: <https://doi.org/10.24963/ijcai.2020/484>
- [10] T.-M. Nguyen, H.-L. Nguyen, D.-Q. Nguyen, H.-T. Nguyen, T.-H.-Y. Vuong, and H.-T. Nguyen, “Nowj@coliee 2024: Leveraging advanced deep learning techniques for efficient and effective legal information processing,” 2024, pp. 183–199. [Online]. Available: https://doi.org/10.1007/978-981-97-3076-6_13
- [11] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” 2020. [Online]. Available: <https://arxiv.org/abs/2004.05150>
- [12] C. Nguyen, T. Tran, K. Le, H. Nguyen, T. Do, T. Pham, S. T. Luu, T. Vo, and L.-M. Nguyen, “Pushing the boundaries of legal information processing with integration of large language models,” in *JSAI-isAI*, 2024, pp. 167–182. [Online]. Available: https://doi.org/10.1007/978-981-97-3076-6_12
- [13] H. Li, Y. Chen, Z. Ge, Q. Ai, Y. Liu, Q. Zhou, and S. Huo, “Towards an in-depth comprehension of case relevance for better legal retrieval,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.00947>
- [14] J. Rabelo, M. Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, and K. Satoh, “Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021,” *Artificial Intelligence and Law*, pp. 111–133, 2022.
- [15] P. Nguyen, C. Nguyen, H. Nguyen, M. Nguyen, A. Trieu, D. Nguyen, and L.-M. Nguyen, “Captain at coliee 2024: Large language model for legal text retrieval and entailment,” in *New Frontiers in Artificial Intelligence*, T. Suzumura and M. Bono, Eds. Singapore: Springer Nature Singapore, 2024, pp. 125–139.
- [16] M. Fujita, T. Onaga, A. Ueyama, and Y. Kano, “Legal textual entailment using ensemble of rule-based and bert-based method with data augmentation by related article generation,” in *JSAI-isAI Workshops*, 2022, pp. 138–153. [Online]. Available: https://doi.org/10.1007/978-3-031-29168-5_10
- [17] M. Yoshioka, Y. Suzuki, and Y. Aoki, “Hukb at the coliee 2022 statute law task.” Berlin, Heidelberg: Springer-Verlag, 2022, p. 109–124. [Online]. Available: https://doi.org/10.1007/978-3-031-29168-5_8
- [18] N. V. H. Nguyen Nho Trung, Nguyen Nhat Quang, “Vietnamese_embedding: Embedding model in vietnamese language.” 2025.
- [19] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation,” 2024.

Local Focus, Global Coverage: A Hybrid Retrieval Architecture for Legal Text

Hieu M. Huynh^{1,2}, Lam M. H. Nguyen^{1,2}, Thang Q. Pham^{1,2}, Khang Huy Mai^{1,2}, and Kiet Van Nguyen^{1,2}

¹Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

Email: {22520434, 23520834, 23521431, 23520692}@gm.uit.edu.vn, kietnv@uit.edu.vn

Abstract—This paper presents the UIT legalGrep Team’s solution for the 5th Automated Legal Question Answering Competition (ALQAC 2025), which focuses on two main tasks: Legal Document Retrieval and Legal Question Answering. For the Legal Document Retrieval task, we proposed a framework that combines a Local Retrieval Module and a Global Retrieval Module. The Local Retrieval Module leverages a category classifier to select relevant legal domains, followed by specialized retrieval within each selected category. In parallel, the Global Retrieval Module integrates both lexical retrieval and semantic retrieval methods to capture relevant documents from the entire collection. The outputs from both modules are merged via a rank fusion strategy and subsequently reranked to produce the final set of relevant legal documents. For the Legal Question Answering task, large language models (LLMs) are employed to generate precise answers based on the relevant legal context. Our approach achieved 2nd place in the Legal Document Retrieval task (F2-score: 89.70%) and 12th place in the Legal Question Answering task (accuracy: 75.61%), demonstrating the effectiveness of our modular retrieval strategy.

Index Terms—ALQAC2025, Legal Document Retrieval, Legal Question Answering, Large Language Model

I. INTRODUCTION

The proliferation of legal texts, such as statutes and case law, presents a significant challenge for legal professionals who need to access and reason with relevant information efficiently. The complexity of legal language - characterized by intricate logical structures and domain-specific terminology - makes manual research a labor-intensive process. To address this, the field has seen a rise in automated systems for Legal Information Retrieval (IR) and Legal Question Answering (QA). Legal IR focuses on retrieving relevant statutes for a query, while Legal QA aims to generate precise answers from those texts.

To advance research in this area, the 5th Automated Legal Question Answering Competition (ALQAC 2025) was organized, featuring two primary tasks: Legal Document Retrieval and Legal Question Answering. This paper presents the UIT legalGrep team’s solution, which secured Rank 2 in the Document Retrieval task. We introduce a robust hybrid retrieval framework and an LLM-based QA pipeline, detailing the architectural decisions and empirical results that led to our competitive performance.

II. RELATED WORK

A. Legal Document Retrieval

Research in Legal IR has evolved from lexical to semantic retrieval methods. Early systems relied on lexical matching techniques like TF-IDF [1], [2] and BM25 [3], which, while effective, struggle to capture contextual meaning. To overcome this, modern approaches employ semantic retrieval using neural networks, with bi-encoder architectures being a common design.

A key trend is the use of domain-adapted language models, where transformers like BERT [4]–[6] are fine-tuned on legal corpora (e.g., VNLawBERT [7]), significantly boosting performance. Furthermore, hybrid systems that combine the robustness of lexical methods like BM25 with the semantic depth of neural rerankers (e.g., cross-encoders) have become state-of-the-art, a trend consistently validated in competitions like COLIEE and ALQAC. Our work builds on this hybrid approach, integrating specialized local and global retrieval strategies.

B. Legal Question Answering

Legal QA has similarly advanced with neural models. Early approaches adapted extractive QA models, fine-tuning BERT-based architectures to extract answer spans directly from retrieved legal texts. While effective for factoid questions, these models struggle when answers require synthesis across multiple clauses. To address this, subsequent research explored generative models like T5 [8] and BART [9] for more abstractive answering.

The current state-of-the-art is dominated by Large Language Models (LLMs) within a Retrieval-Augmented Generation (RAG) framework. In this paradigm, LLMs leverage their powerful reasoning capabilities to generate answers based on context retrieved from a legal corpus. Our QA system adopts this modern RAG approach, focusing on context optimization and tailored prompting strategies to enhance the accuracy of a Vietnamese-centric LLM.

III. METHODS

A. Datasets

Within the context of ALQAC2025, the primary data sources consist of a law set and a training set provided

directly by the organizers. The law set is composed of different Vietnamese laws, where each law is divided into multiple articles addressing a wide range of legal areas. The training set comprises labeled questions designed for legal document retrieval and legal question answering tasks. Three types of questions are present: multiple-choice, true/false, and free-text. Each question is mapped to the corresponding articles and laws. To provide a clear overview of the datasets used in this work, we summarize key statistics in Table I.

TABLE I
STATISTICAL SUMMARY OF THE LAW SET AND TRAINING SET

	Statistic	Quantity
Law set	Number of laws	28
	Number of articles	724
	Min articles per law	36
	Max articles per law	689
	Average articles per law	120.11
	Min token count per article	4
	Max token count per article	3566
	Avg. token count per article	269.66
Training set	Number of questions	729
	Min token count per question	6
	Max token count per question	98
	Avg. token count per question	31.03

The provided law set contains 28 distinct Vietnamese laws, but the training set only covers 16 and is highly imbalanced. For instance, the Civil Code contains 78 associated questions, while the Law on Youth has only 15. This imbalance and partial coverage pose significant challenges for Legal IR and QA systems.

B. Approach

Our proposed solution for ALQAC 2025 addresses both Legal Document Retrieval and Legal Question Answering tasks. The overall architecture is depicted in Figure 1. The system receives a legal query, processes it through a multi-stage retrieval pipeline, and then leverages Large Language Models (LLMs) to generate precise answers based on the retrieved context.

1) *Legal Document Retrieval*: Our legal document retrieval framework combines a Local Retrieval Module and a Global Retrieval Module, followed by a rank fusion and reranking pipeline to yield the final set of relevant documents.

a) *Local Retrieval Module*: This module aims to narrow down the search space by identifying the most relevant legal categories for a given query, inspired by the human intuition in classifying legal questions. This significantly reduces the document collection, making subsequent retrieval more efficient and precise.

With the category classifier, we frame category identification as a multi-class classification problem, utilizing PhoBERT-base [10], a BERT-based model fine-tuned for Vietnamese. We fixed top_k = 1 for category selection. This decision was based on the observation that most queries in the training data

map to a single legal domain and was made to maximize the efficiency of the subsequent retrieval step.

Once the top 1 category is identified, we perform a dedicated semantic-only retrieval within this specific legal category. For this, we leverage the *AITeamVN/Vietnamese_Embedding* model [11], which is based on the multilingual *BAAI/bge-m3* model [12]. We directly apply this model in a zero-shot manner. Document embeddings for each category are pre-computed and stored in separate collections. During retrieval, we load the corresponding collection and rank articles by computing the cosine similarity between the query embedding and document embeddings.

b) *Global Retrieval Module*: To mitigate potential misclassifications by the Local Retrieval Module and ensure comprehensive coverage, we implement a Global Retrieval Module that performs retrieval across the entire collection of legal articles. We employ BM25 as our global retrieval strategy. While our experiments show a hybrid semantic-lexical approach can perform better in isolation (Table II), BM25 was chosen for our final architecture due to its computational efficiency and its role as a strong, complementary signal to the semantic-only Local module. Default parameters for BM25 are used and, prior to indexing, all legal documents undergo word segmentation using the VNCORENLP toolkit [13], a critical preprocessing step for effective lexical matching in Vietnamese.

c) *Rank Fusion and Reranking*: The outputs from the Local and Global Retrieval Modules are then combined and refined through a two-stage process: rank fusion and reranking.

With rank fusion, we utilize the Reciprocal Rank Fusion (RRF) algorithm to consolidate the ranked lists from both modules. The RRF score for a document is calculated as

$$\text{score}(d) = \sum_{r \in R} \frac{1}{k + \text{rank}_d} \quad (1)$$

where d is the document, R is the set of retrievers, k is a constant set to 60, and rank_d is the rank of document d in retriever r . If a document appears at rank rank_d in any given list, it contributes to its total score; if not, it does not. The fused list is then truncated to the top 10 most relevant candidates.

For reranking, we employ *AITeamVN/Vietnamese_Reranker* [11], a cross-encoder based on *BAAI/bge-reranker-v2-m3* [12] fine-tuned on over a million Vietnamese (query, passage) pairs. It computes a relevance score for the top 10 candidates from the fusion step.

d) *Post-processing for Retrieval Augmentation*: Following the initial reranking, we implement a post-processing step to enhance recall, especially for cases where relevant information might be distributed across neighboring articles or for very long documents.

Neighboring Article Expansion: We augment the top 10 candidates by including adjacent articles. Specifically, for each candidate, we add a configurable number of articles to its left and right (empirically determined to be 3 to 5 articles on each side for the 1st article). This heuristic addresses potential missing ground truths due to truncation or

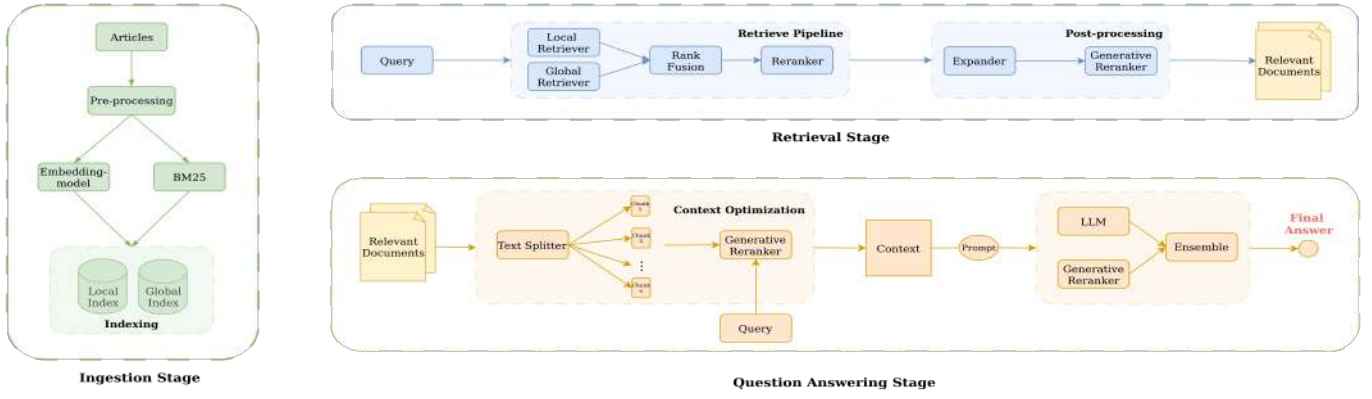


Fig. 1. Overall architecture of the proposed system.

the inherent contextual relationships between adjacent legal provisions. Duplicate articles are subsequently removed, with this expansion typically adding 4 to 7 new unique documents.

Generative Reranking: To overcome cross-encoder limits with very long documents and leverage advanced reasoning, we apply a generative reranking step using a causal Language Model. We use *Qwen/Qwen3-Reranker-8B* [14], known for handling long contexts (up to 8192 tokens). We then employ a Code-Switching Prompting strategy [15] with clear instructions and structured tags (e.g., Instruct, Query, Document) to guide reasoning. The model *Qwen/Qwen3-Reranker-8B*, has been trained for reranking, is applied directly with prompting. The model computes the log-softmax of “yes” and “no” tokens; the “yes” probability serves as the relevance score to re-rank the expanded legal documents, enabling richer semantic understanding and logical inference.

2) *Legal Question Answering:* The ALQAC 2025 competition is structured as sub-tasks, where the gold labels (relevant documents) from Task 1 are provided as input for Task 2. Our Legal Question Answering approach focuses on optimizing the context provided to the LLM and generating precise answers for different question types.

a) *Context Optimization:* We observed that retrieved legal articles often contain superfluous information not directly relevant to the query. To mitigate this “information overload” for the LLM, we implemented a context optimization strategy. Firstly, legal articles are chunked using *RecursiveCharacterTextSplitter*. This method recursively splits text based on a predefined list of delimiters, which is particularly effective for hierarchical legal documents, minimizing semantic loss and unnecessary overlaps. Then, for each query, the relevance between the query and individual chunks is calculated using a scoring method similar to the reranking step in the retrieval post-processing pipeline (employing *Qwen/Qwen3-Reranker-8B*). Chunks with a score below a fixed threshold are filtered out as irrelevant, thereby reducing the noise in the context provided to the LLM. Finally, the remaining relevant chunks are concatenated in order, with metadata about the originating article prepended, to form the optimized context for the LLM.

b) *Answer Generation:* We employ *Vistral 7B Chat* [16] as the core LLM for generating answers due to its strong Vietnamese language understanding capabilities. A generic system prompt establishes the LLM’s role as a legal AI assistant, emphasizing careful reasoning within `<thinking></thinking>` tags and precise answers within `<answer></answer>` tags. Specific instructions are provided for true/false, multiple-choice, and free-Text questions (concise, comprehensive, accurate). Prompts are tailored for each question type. For true/false questions, the prompt incorporates a “Re-Reading” [17] framework to enhance reasoning. For multiple-choice questions, the prompt guides the LLM to analyze options against the context. For free-text questions, the prompt emphasizes concise and complete answers. All prompt templates are designed to ensure structured responses amenable to parsing.

c) *Generative Reranking for Selected Answer Types:* For true/false and multiple-choice questions, we use a final reranking step to refine the answer. We use the *Qwen/Qwen3-Reranker-8B* model to score each possible choice (e.g., “True”, “False”, or options “A”, “B”, “C”, “D”) against the context. The final answer is determined by an ensemble rule: if the *Qwen-Reranker*’s top-scoring choice differs from the initial prediction made by *Vistral 7B Chat*, and its confidence (the log-softmax difference between the top-two choices) exceeds a pre-defined threshold (τ), we adopt the reranker’s answer. Otherwise, we default to the original *Vistral* prediction. For free-text questions, the *Vistral*-generated answer is always used directly.

C. Experimental Setup

All experiments were conducted on a system equipped with an RTX 3090 GPU with 24GB VRAM. The primary software frameworks and libraries utilized include PyTorch and Hugging Face Transformers.

IV. EXPERIMENTS AND RESULTS

A. Legal Document Retrieval

In this task, the official evaluation metrics, as defined by ALQAC 2025, are precision_i , recall_i and the macro-averaged

F2 – score. The formula is noted below, where precision_i and recall_i are the precision and recall of question i^{th} , respectively.

$$\text{F2}_{\text{macro}} = \text{average of } \left(\frac{5 \times \text{precision}_i \times \text{recall}_i}{4 \times \text{precision}_i + \text{recall}_i} \right) \quad (2)$$

Our experimental configurations are succinctly described using the following notation:

- $\mathcal{L}$: Local Retrieval Module enabled.
- $\mathcal{G}$: Global Retrieval Module enabled.
- $\mathcal{R}$: Reranking component enabled.
- $\mathcal{PP}$: Post-processing enabled.
- Methodologies: sem (semantic-based), lex (lexical-based), bm25 (BM25-only).

For example, $\mathcal{L}(\text{sem}) + \mathcal{G}(\text{bm25} + \text{sem}) + \mathcal{R}$ indicates a configuration employing Local (semantic), Global (BM25 and semantic hybrid), and Reranking components.

The following content will present the key findings from our experiments, highlighting the performance of different architectural components and retrieval strategies.

As shown in Table II, applying a cross-encoder reranker ($+\mathcal{R}$) is the single most impactful component, dramatically improving the F2-score across all base configurations (e.g., from 0.7249 to 0.8442 for the Global BM25 retriever).

An ablation of our modules shows that the Global retriever ($\mathcal{G}(\text{bm25}) + \mathcal{R}$) forms a stronger standalone baseline than the Local retriever ($\mathcal{L}(\text{sem}) + \mathcal{R}$), whose performance is capped by the classifier’s accuracy. Further analysis within the modules confirms that semantic search is superior for the focused Local task, while a hybrid bm25+sem approach is best for the comprehensive Global task.

TABLE II
ABLATION STUDY OF THE PROPOSED RETRIEVAL FRAMEWORK

Experiment	P@1	R@1	F2@1
<i>1. Baseline Modules (without Reranking)</i>			
$\mathcal{G}(\text{bm25})$	0.7298	0.7243	0.7249
$\mathcal{L}(\text{sem})$	0.7723	0.7695	0.7699
<i>2. Impact of Reranking on Individual Modules</i>			
$\mathcal{G}(\text{bm25}) + \mathcal{R}$	0.8491	0.8436	0.8442
$\mathcal{L}(\text{sem}) + \mathcal{R}$	0.8313	0.8278	0.8282
<i>3. Comparison of Retrieval Strategies (with Reranking)</i>			
$\mathcal{L}(\text{bm25}) + \mathcal{R}$	0.7956	0.7922	0.7926
$\mathcal{L}(\text{bm25} + \text{sem}) + \mathcal{R}$	0.8011	0.7977	0.7980
$\mathcal{G}(\text{sem}) + \mathcal{R}$	0.8242	0.8082	0.8114
$\mathcal{G}(\text{bm25} + \text{sem}) + \mathcal{R}$	0.8642	0.8594	0.8599
<i>4. Hybrid Framework Performance</i>			
$\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem})$	0.7462	0.7435	0.7438
$\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) + \mathcal{R}$	0.8930	0.8889	0.8893
<i>5. Impact of Post-Processing</i>			
$\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) + \mathcal{R} + \mathcal{PP}$	0.9163	0.9115	0.9121

Finally, the result confirms the synergy of our complete hybrid model $\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) + \mathcal{R}$, which at 0.8893 outperforms the best standalone Global module (0.8599). The addition of our final post-processing pipeline ($+\mathcal{PP}$)

Retrieval from a Long Document with Diluted Context

- **Question:** "What is the definition of 'childbirth by assisted reproduction techniques'?" (paraphrased)
- **Challenge:** The correct answer is a single sentence within a long article containing 25 distinct legal definitions.

Analysis of Pipeline Stages:

- 1) $\mathcal{L}(\text{sem}) \rightarrow$ Rank 8: The dense vector representation of the document was "diluted" by the large amount of irrelevant text, leading to a weak semantic similarity score.
- 2) $\mathcal{G}(\text{bm25}) \rightarrow$ Rank 3: This module succeeded due to strong, exact keyword matching ("*assisted reproduction*", etc.), proving its robustness against document length.
- 3) $\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) \rightarrow$ Rank 5: The RRF score was slightly degraded because the weak signal from the Local module pulled down the strong score from the Global module.
- 4) $\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) + \mathcal{R} + \mathcal{PP} \rightarrow$ Rank 1: The final Generative Reranker stage was decisive. The LLM’s ability to process long contexts allowed it to deeply understand the relationship between the query and the specific defining sentence, assigning it the highest relevance score.

Fig. 2. A case study demonstrating the effectiveness of the multi-stage pipeline on a long document retrieval task. The final Generative Reranker stage was critical in achieving the top rank.

provides another significant boost to 0.9121, highlighting the effectiveness of neighboring article expansion and generative reranking. This comprehensive approach culminated in our Rank 2 finish in the official competition (Table III).

To provide qualitative insight, we present a case study on a challenging query (Fig. 2). The task is to find a specific definition within a very long legal article, a classic needle-in-a-haystack problem that reveals the strengths and weaknesses of each component in our pipeline.

To further demonstrate the generalizability and efficacy of our framework, we conducted an additional evaluation on the Zalo-sponsored dataset provided within the same competition. This dataset comprises 3196 questions and 3271 law documents, which are further segmented into 61425 articles. The detailed results can be found in Table IV.

TABLE III
TOP 5 TEAMS BY F2 SCORE (TASK 1)

Team	F2 Score	Rank
Flame Reavers	0.9434765774	1
UIT legalGrep (Our)	0.89701897	2
vund	0.89701897	2
NOWJ	0.8766937667	4
UTEHY-NLU	0.8482384822	5

B. Legal Question Answering

Given that gold labels for Task 2’s private test set were not publicly released, our evaluation and analysis are conducted on a dedicated validation set, which was carefully separated from the organizers’ original training data. The primary evaluation metric for this task is Accuracy.

$$\text{accuracy} = \frac{\text{\#questions that were correctly answered}}{\text{\#questions}} \quad (3)$$

TABLE IV
EVALUATION RESULTS ON ZALO DATASET

Experiment	P@1	R@1	F2@1
$\mathcal{L}(\text{sem})$	0.5954	0.5867	0.5876
$\mathcal{G}(\text{bm25})$	0.3917	0.3860	0.3866
$\mathcal{G}(\text{sem})$	0.5879	0.5796	0.5805
$\mathcal{L}(\text{sem}) + \mathcal{R}$	0.6020	0.5943	0.5951
$\mathcal{G}(\text{bm25}) + \mathcal{R}$	0.5732	0.5657	0.5665
$\mathcal{G}(\text{sem}) + \mathcal{R}$	0.6070	0.5990	0.5998
$\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) + \mathcal{R}$	0.6652	0.6570	0.6579
$\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) + \mathcal{R} + \mathcal{PP}$	0.6815	0.6724	0.6733

1) *Context Optimization*: As shown in Table V, our context optimization strategy demonstrated a clear benefit for discriminative question types. Accuracy for true/false questions increased from 0.8376 to 0.8534, and for multiple-choice from 0.5814 to 0.6977. However, a notable decrease in Accuracy was observed for free-text questions, from 0.7647 to 0.7059. This highlights a nuanced impact on different question types and will be discussed in the next section.

TABLE V
PERFORMANCE WITH CONTEXT OPTIMIZATION

Experiment	True/False	Multiple-Choice	Free-text	Avg.
w/o Context Optimization	0.8376	0.5814	0.7647	0.7318
w/ Context Optimization	0.8534	0.6977	0.7059	0.7808

2) *Evaluation of Ensemble Strategy*: Our approach to Legal Question Answering leverages an ensemble strategy, combining the answer generation capabilities of *Vistral 7B Chat* with the generative reranking of *Qwen/Qwen3-Reranker-8B* for true/false and multiple-choice questions. We assess the contribution of each component and the overall ensemble’s performance on our validation set.

TABLE VI
PERFORMANCE OF ANSWER GENERATION STRATEGIES

Experiment	True/False	Multiple-Choice	Free-text	Avg.
Vistral Chat 7B Only	0.8534	0.6977	0.7059	0.7808
Qwen-Reranker Only	0.4767	0.5345	N/A	N/A
Ensemble Strategy	0.8276	0.5930	0.7059	0.7260

3) *ALQAC 2025 Ranking*: Despite the sophisticated approaches implemented for context optimization and ensemble answer generation, our UIT legalGrep Team achieved Rank 12 in Task 2 (Legal Question Answering) on the private test set, with an Accuracy of 0.7561. This result suggests that while our component-level optimizations were effective, further research is needed to bridge the gap with top-performing systems.

C. Error Analysis

To understand the limitations of our pipeline, we performed a detailed analysis of the failure cases. Firstly, we analyzed the rank of the first correct document for all 729 queries

and found the following failure distribution: 61 (8.4%), 41 (5.6%), 32 (4.4%) and 24 (3.3%) queries fail to place the correct document at rank 1, 2, 3 and 10, respectively. This observation indicates that while most relevant documents are retrieved, the reranking stages still struggle to consistently promote the correct one to the top position for a subset of challenging queries. A manual inspection of the 24 hard recall failures, questions in which the correct document was not in the top 10, reveals four primary challenges. The most frequent issue ($\approx 30\%$) is fine-grained semantic ambiguity, where the system correctly identifies the law but fails to disambiguate between highly similar articles (e.g., *The People’s Procuracy has the right to participate at any stage of the administrative case trial, true or false?*). Another significant portion of errors (25%) stems from query intent misinterpretation, where models latch onto secondary keywords and pursue an incorrect legal category from the outset (e.g., *What contents must a request for temporary emergency measures contain?*). Other failures are rooted in the queries themselves; approximately $\approx 20\%$ are due to inherent query ambiguity, where underspecified questions like *Which of the following acts is prohibited?* lack the necessary context for any retrieval method to succeed. Finally, narrative-based “situation questions” ($\approx 15\%$) cause a dual failure of our initial retrieval paths (e.g., *Mr. Nguyen Van C lived with his parents from birth until he was 12 years old. His parents lived in his father’s hometown in Hanoi. After his parents divorced, he returned to live with his mother in Tuyen Quang until he was 16 years old. Where is Mr. Nguyen Van C’s place of residence?*), as both the classifier and keyword-based modules fail, preventing the correct document from ever reaching the reranking stages.

For Legal QA, we identified two primary failure categories for free-text questions: Model’s Failure to Ground in Context, where the model hallucinates the absence of information despite having the perfect context, and Failure in Reasoning and Intent Matching, where it misunderstands a “how” query and provides a simple validation instead of the required summary.

V. DISCUSSION

A. Insights from the Hybrid Retrieval Framework

Our experiments, especially the superior F2-score of 0.8893 from the $\mathcal{G}(\text{bm25}) + \mathcal{L}(\text{sem}) + \mathcal{R}$ setup, validate the effectiveness of our hybrid retrieval design. Its success comes from the complementary strengths of each part: the Local Retrieval Module uses a category classifier to narrow the search space and boost precision, while the Global Retrieval Module acts as a safety net to avoid missing relevant documents. This lets semantic search excel in focused contexts and BM25 remains robust broadly. The following Reranking and Post-processing stages further refine results, leading to competitive ALQAC 2025 performance.

Despite strong results, the system has limits on complex legal datasets. It assumes each query maps cleanly to a few legal categories, but this fails when categories are many or overlap. Tasks needing multi-hop reasoning across articles also

challenge the initial keyword and semantic matching, making it hard to gather all needed documents. This weakness became evident in the Zalo-sponsored dataset, where larger, more complex laws lowered performance. To address this, potential future work includes a flexible classifier that can tag multiple categories simultaneously and graph-based representations of legal articles that encode their links to better support multi-hop reasoning.

B. Analysis of Legal Question Answering Performance

Our findings in the Legal Question Answering task reveal a complex relationship between our methodology and different types of questions. The context optimization strategy, which was a key to our QA pipeline, improved accuracy for true/false and multiple-choice by giving the LLM cleaner, focused context. But it reduced accuracy on free-text generation, likely because such tasks need broader context to build full answers; relevance filtering sometimes dropped helpful background clauses that the LLM needed to construct a coherent response.

The ensemble strategy for objective questions gave no significant gain over the single Vistral 7B Chat model. This suggests Vistral 7B Chat already excels at direct answering with the optimized context, and the Qwen-Reranker, while excellent for document ranking, may not add a sufficiently complementary signal for final answer selection.

VI. CONCLUSION

In this paper, we presented the UIT legalGrep Team's system for the ALQAC 2025 competition. Our proposed hybrid retrieval framework, which combines a category-based local semantic search with a robust global lexical search, followed by a multi-stage reranking pipeline, proved highly effective and secured Rank 2 in the Legal Document Retrieval task. Our analysis confirmed the value of each component and highlighted the significant impact of both cross-encoder and generative reranking. Future work will focus on addressing the challenges of multi-hop reasoning and the nuanced requirements of generative QA, potentially by exploring legal knowledge graphs and more advanced QA-specific models.

ACKNOWLEDGEMENTS

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under the grant number DS2025-26-01.

REFERENCES

- [1] T.-M. Nguyen, X.-H. Nguyen, N.-D. Mai, M.-Q. Hoang, V.-H. Nguyen, H.-V. Nguyen, H.-T. Nguyen, and T.-H.-Y. Vuong, "Nowj1@ alqac 2023: Enhancing legal task performance with classic statistical models and pre-trained language models," *arXiv preprint arXiv:2309.09070*, 2023.
- [2] H. N. Van, D. Nguyen, P. M. Nguyen, and M. Le Nguyen, "Miko team: Deep learning approach for legal question answering in alqac 2022," in *2022 14th international conference on knowledge and systems engineering (KSE)*. IEEE, 2022, pp. 1–5.
- [3] H. Q. Pham, Q. V. Nguyen, D. Q. Tran, T. K.-B. Nguyen, and K. V. Nguyen, "Top 2 at alqac 2024: Large language models (llms) for legal question answering," *International Journal of Asian Language Processing*, vol. 35, no. 01, p. 2450010, 2025.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [6] S. Ruder, A. Søgaard, and I. Vulić, "Unsupervised cross-lingual representation learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 31–38.
- [7] C.-N. Chau, T.-S. Nguyen, and L.-M. Nguyen, "Vnlawbert: A vietnamese legal answer selection approach using bert language model," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, 2020, pp. 298–301.
- [8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023. [Online]. Available: <https://arxiv.org/abs/1910.10683>
- [9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: <https://aclanthology.org/2020.acl-main.703/>
- [10] D. Q. Nguyen and A. Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.92/>
- [11] N. N. Q. Nguyen Nho Trung, "Vietnamese_embedding: Embedding model in vietnamese language." 2025.
- [12] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2318–2335. [Online]. Available: <https://aclanthology.org/2024.findings-acl.137/>
- [13] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Y. Liu, T. Paek, and M. Patwardhan, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 56–60. [Online]. Available: <https://aclanthology.org/N18-5012/>
- [14] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin *et al.*, "Qwen3 embedding: Advancing text embedding and reranking through foundation models," *arXiv preprint arXiv:2506.05176*, 2025.
- [15] Z. X. Yong, R. Zhang, J. Forde, S. Wang, A. Subramonian, H. Lovenia, S. Cahyawijaya, G. Winata, L. Sutawika, J. C. B. Cruz, Y. L. Tan, L. Phan, L. Phan, R. Garcia, T. Solorio, and A. F. Aji, "Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages," in *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, G. Winata, S. Kar, M. Zhukova, T. Solorio, M. Diab, S. Sitaram, M. Choudhury, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 43–63. [Online]. Available: <https://aclanthology.org/2023.calcs-1.5/>
- [16] Q. N. H. N. B. P. N. P. H. N. P. S. T. N. Chien Van Nguyen, Thuat Nguyen, "Vistral-7b-chat - towards a state-of-the-art large language model for vietnamese," 2023.
- [17] X. Xu, C. Tao, T. Shen, C. Xu, H. Xu, G. Long, J.-G. Lou, and S. Ma, "Re-reading improves reasoning in large language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 15 549–15 575. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.871/>

A Summary of the ALQAC 2025 Competition

Dinh-Truong Do^{1,*}, Son T. Luu^{1,2}, Trang Pham¹, Trung Vo¹, Nguyen-Hoang Chu¹, Quang-Huy Chu¹, Cuong Hoang¹, Tan-Minh Nguyen¹, An Trieu¹, Dat Nguyen¹, Hiep Nguyen¹, The-Hai Nguyen¹, Nguyen-Khang Le¹, Dieu-Hien Nguyen¹, Ngoc-Cam Le⁴, Thi-Thuy Le⁵, Hoang-Anh Dang¹, Ha-Thanh Nguyen³, Vu Tran¹, Le-Minh Nguyen¹

¹*Japan Advanced Institute of Science and Technology, Ishikawa, Japan*

²*VNU University of Information Technology, Ho Chi Minh City, Vietnam*

³*National Institute of Informatics, Tokyo, Japan*

⁴*Vietnam Judicial Academy, Hanoi, Vietnam*

⁵*Hanoi Law University, Hanoi, Vietnam*

Abstract—This paper provides an overview of the fifth edition of the Automated Legal Question Answering Competition (ALQAC 2025), aimed at advancing research in legal text processing for low-resource languages. This year’s competition focused on two core tasks involving Vietnamese legal statutes: Legal Document Retrieval and Legal Question Answering. To foster innovation, ALQAC 2025 introduced new challenges, including scenario-based problem settings and a combination of both expert-verified and non-expert-verified training data. We describe the design of the competition, datasets, and evaluation protocols, followed by an analysis of the diverse methodologies adopted by the participating teams. The results demonstrate continued progress in automated legal question answering and document retrieval, highlighting emerging trends and persistent challenges in this domain.

Index Terms—ALQAC 2025, Vietnamese legal texts, legal text processing, competition

I. INTRODUCTION

The rapid expansion of digital legal information has created a growing demand for intelligent systems that can efficiently retrieve legal documents and accurately answer complex legal questions. Such systems have the potential to transform legal research by streamlining access to relevant information and supporting informed decision-making. However, developing these technologies remains especially challenging for low-resource languages such as Vietnamese, where linguistic and legal resources are limited.

For the 2025 edition, ALQAC is organized in conjunction with the International Conference on Knowledge and Systems Engineering (KSE 2025) and centers on two main tasks: Legal Document Retrieval (Task 1) and Legal Question Answering (Task 2). This year’s competition features a newly curated legal question-answering dataset, carefully annotated using Vietnamese statutory law as the foundation. Both expert-verified and non-expert-verified training data were provided, encouraging participants to design robust systems that can handle data of varying quality. The overarching goal is to foster a vibrant research community dedicated to developing legal AI tools for languages with limited resources.

In Task 1, participants are challenged to design systems capable of retrieving pertinent legal articles in response to

specific queries. Evaluation is based on the Macro-F2 metric, supplemented by precision and recall. Task 2 requires participants to address a diverse set of legal questions, including True/False, multiple-choice, free-text, and scenario-based formats. To ensure the reliability of system outputs, answers are primarily assessed by legal experts, with accuracy serving as the principal performance criterion. A public leaderboard was maintained for preliminary testing, allowing teams to iteratively refine their systems before the final evaluation.

Motivated by the need to bridge this gap, the fifth edition of the Automated Legal Question Answering Competition (ALQAC 2025) was organized to stimulate research and collaboration in legal AI for low-resource settings. Inspired by prior initiatives such as COLIEE [1], ALQAC 2025 introduced novel aspects to address real-world legal challenges, including scenario-driven questions and heterogeneous training data quality. We have one team that achieved 1st place for Task 1, and three teams that achieved 1st place for Task 2 this year. Through the competition, we aim to catalyze the development of advanced legal support systems and foster a collaborative research community dedicated to addressing the unique challenges of legal text processing for under-resourced languages.

The paper is organized as follows: Section II provides an overview of the dataset. Sections III and IV delve into the specifics of the tasks, detailing the methodologies employed by participants and presenting the results for Task 1 and Task 2, respectively. Section V presents our discussion, while Section VI summarizes the conclusions drawn from the study.

II. DATASET

A. Data Annotation

For ALQAC 2025, the dataset was constructed through a refined multi-stage annotation protocol, building upon the foundations established in [2]. The process focused on generating a diverse range of legal questions, spanning True/False, multiple-choice, and free-text formats, all rooted in authentic Vietnamese legal contexts.

The annotation workflow consisted of four main phases:

- **Guideline development:** Comprehensive annotation instructions were created, outlining each question type

* Corresponding author: truongdo@jaist.ac.jp.

TABLE I
STATISTICS ON LAW DOCUMENTS USED IN ALQAC 2025

Law	Law in English	Number of articles
Luật Phòng, chống ma túy	LAW ON PREVENTION AND CONTROL OF NARCOTIC SUBSTANCES	55
Luật Bảo vệ môi trường	LAW ON ENVIRONMENTAL PROTECTION	171
Luật Viên chức	LAW ON PUBLIC EMPLOYEES	62
Luật Cư trú	RESIDENCE LAW	38
Luật Tiếp cận thông tin	LAW ACCESS TO INFORMATION	37
Luật Trồng trọt	LAW CROP PRODUCTION	85
Luật An ninh mạng	CYBERSECURITY LAW	43
Luật Chăn nuôi	LAW ON ANIMAL HUSBANDRY	83
Hiến pháp	THE CONSTITUTION OF THE SOCIALIST REPUBLIC OF VIETNAM	120
Luật Du lịch	LAW ON TOURISM	78
Luật Tổ tụng hành chính	LAW ON ADMINISTRATIVE PROCEDURES	372
Luật Hôn nhân và gia đình	LAW ON MARRIAGE AND FAMILY	133
Bộ luật dân sự	CIVIL CODE	689
Luật Điện ảnh	LAW ON CINEMATOGRAPHY OF VIETNAM	50
Luật Trọng tài thương mại	LAW ON COMMERCIAL ARBITRATION	82
Luật Thanh niên	YOUTH LAW	41
Luật Giáo dục	EDUCATION LAW	115
Luật Đường bộ	LAW ON ROAD	86
Luật Giá	LAW ON PRICES	75
Luật Căn cước	LAW IDENTIFICATION	46
Luật Khám bệnh, chữa bệnh	LAW ON MEDICAL EXAMINATION AND TREATMENT	121
Luật Phòng, chống tác hại của rượu, bia	LAW ON PREVENTION AND CONTROL OF HARMFUL EFFECTS OF ALCOHOLIC BEVERAGES	36
Luật Bảo vệ quyền lợi người tiêu dùng	LAW ON PROTECTION OF CONSUMER RIGHTS	80
Luật Giao dịch điện tử	LAW ON ELECTRONIC TRANSACTIONS	53
Luật Viễn thông	LAW ON TELECOMMUNICATIONS	73
Luật Tài nguyên nước	LAW ON WATER RESOURCES	86
Luật Đất đai	LAND LAW	260
Luật Nhà ở	LAW ON HOUSING	198

and providing illustrative examples to ensure consistency across annotators.

- **Initial annotation:** Annotators drafted candidate questions and answers by strictly adhering to the established guidelines.
- **Collaborative review:** A cross-validation phase is conducted among annotators, during which discussions are held to refine and finalize the annotations.
- **Expert validation:** Expert verification ensures quality control, as only questions and answers endorsed by at least one expert are included in the dataset.

This rigorous pipeline resulted in a high-quality resource tailored for the competition. During the review phase, validators provided feedback on any deviations or ambiguities, prompting the original annotator to revise and resubmit the question or answer as needed. Final approval required at least one expert’s endorsement. Notably, the dataset curation placed special emphasis on questions requiring numerical reasoning, temporal understanding, and contextually realistic scenarios, reflecting genuine complexities in Vietnamese legal practice. Following the cross-validation phase, we have 2 different people performing to validate the correctness of the answers and questions of the annotators. Each sample is marked as correct or incorrect by the validator, and the annotator must revise the incorrect samples. The accuracy between the two validators is approximately 81.66% in general, indicating the high agreement between the two validators in cross-checking. Overall, the annotation strategy ensured a reliable and challenging dataset for advancing legal AI research in low-resource settings. Moreover, to ensure the quality of the final private test dataset, the experts read and revise the validated questions and answers by the annotators to decide which sample is chosen

for the final private test.

B. Data Analysis

In ALQAC 2025, we supplement new law documents to enrich the law document database. In total, we have 28 law documents as shown in Table I to serve for the Legal Document Retrieval task.

TABLE II
STATISTICS ABOUT THE TRAINING AND TEST DATASETS IN ALQAC 2025

	Train	Public test	Private test
Yes/No	387	48	60
Multiple-choice	286	43	11
Free-text	56	9	11
<i>Total</i>	<i>729</i>	<i>100</i>	<i>82</i>
Max articles per samples	3	1	2
Min articles per samples	1	1	1
Average articles per samples	1.02	1.08	1.01

Besides, Table II shows the overall statistical information about the training, public, and private tests of ALQAC 2025. It can be seen that the Yes/No questions take a large proportion of the three sets, especially on the private test. Besides, the number of articles per question in the three sets is similar on average, indicating that most of the questions have one relevant article. On the other hand, Table III shows the length of questions and free-text answers computed at the token level¹. On average, the distribution of question length on the three sets: training, public, and private tests is nearly similar, where the length of the question is around 20 to 29 tokens. For the

¹We use the Pyvi <https://pypi.org/project/pyvi/> for tokenizing the text into tokens

free-text answers, the length in the training set is much longer than in the public and private test. In the private test, the free-text answer is completely verified and revised by the law experts to ensure the correctness, logic, and consistency of the legal context. Therefore, the free-text answer in the private test challenges the AI model in generating a correct and compact answer for a legal query.

TABLE III
STATISTICS ABOUT THE LENGTH OF QUESTIONS AND FREE-TEXT ANSWERS IN THE ALQAC 2025 DATASET

	Train	Public test	Private test
Max question length	66	115	76
Min question length	4	6	7
Average question length	21.81	29.86	21.52
Max free text answer length	80	21	14
Min free text answer length	1	2	2
Average free text answer length	17.42	7	7.81

Overall, ALQAC 2025 provided a manually annotated dataset with expert verification for the competition. To ensure fairness, the participant teams cannot use externally annotated datasets specifically created for legal question answering or legal entailment, such as labeled QA pairs or entailment examples.

III. METHODOLOGY OF PARTICIPANTS

A. Task 1: Legal Document Retrieval

Task Description: Given a legal question Q , the objective of Task 1 is to identify and retrieve a set of relevant legal articles $S_1, S_2, \dots, S_N$ from Vietnamese statutory law. Each selected article should directly relate to the question, ensuring that, together, the retrieved articles provide sufficient information to justify or answer Q accurately. Participants are expected to avoid including unnecessary or irrelevant articles, resulting in a concise and focused evidence set. The required format for Task 1 submissions is as follows:

```
{
  "question_id": "<qid>",
  "question_type": "<qtype>",
  "text": "<Question>",
  "relevant_articles": [
    {
      "law_id": "<law_id>",
      "article_id": "<article_id>"
    }
  ]
}
```

Participant Methods: There are 13 teams with 37 submission entries for the Legal Document Retrieval task. We summarize the method of 13 teams as below:

- **Flame Reavers:** The teams use Qwen3-8B to predict the final ranking by reasoning and reranking. Besides, the teams use the hybrid approach that uses two different embedding encoder models, including *AITeamVN/Vietnamese_Reranker* and *huyg1108/vi-law-reranker-mixed-dataset* to create the initial ranking score, then re-ranking by substituting LLM's output to scores that are below 0.8.
- **vund:** The team use *AITeamVN/Vietnamese_Embedding* for chunking and *AITeamVN/Vietnamese_Reranker* for re-ranking. Then, the team uses LGBMRanker and Qwen3-8B to select the top 1 and the top 3 candidates, respectively.
- **UIT legalGrep:** The teams use BM25 as a global retriever and *AITeamVN/_Embedding* with *vinai/phobert-base* as a local retriever to choose the top-k candidate, then re-ranking by *Qwen/Qwen3-Reranker-8B model*.
- **NOWJ:** The teams use *BGE-M3 (fine-tuned)* + *BM25* as a bi-encoder to extract relevant candidates, then apply a re-ranking method based on the ensembling of *bge-reranker-v2-m3 (fine-tuned)*, *AITeamVN/Vietnamese_Reranker*, *mBERT*, and *multilingual-e5-large* via LightGBM. Finally, the team chose top-k equal to 2 and 3 for their submissions.
- **UTEHY-NLU:** The teams construct a two-stage retrieval as follows: they first use dense embedding models like *AITeamVN/Vietnamese_Embedding*, *BAAI/bge-m3*, and *multilingual-e5-large* for retrieving top-k candidates, then apply cross-encoder models such as *AITeamVN/Vietnamese_Embedding* and *Qwen/Qwen3-4B* for re-ranking.
- **Hallucineers:** The author use BM25 with re-ranking models including *huyydangg/DEK21_hcmute_embedding* and *AITeamVN/Vietnamese_Reranker* fine-tuned on the provided dataset to find top-k candidates. Additionally, the team utilizes GridSearch to identify the most optimized hyperparameters.
- **D2:** The teams use the BM25 to score raw legal texts, then using *AITeamVN/Vietnamese-Reranker* to rerank candidates based on contextual relevance. To enhance the retrieval task, the team also employs *AITeamVN/GRPO-VI-Qwen2-7B-RAG* to rewrite the input queries.
- **UIT-Darkcow:** The team uses BM25 to find top-k relevant articles based on lexical similarity, and then combines the scores from BM25 with SentenceTransformer to re-rank the results.
- **ABCD:** The teams use a hybrid search strategy by leveraging two Vietnamese embedding models, including *truuro7/vn-law-embedding* and *dangvantuan/vietnamese-document-embedding*, to construct the dense vector embedding for combination with sparse keyword matching to perform retrieving relevant legal documents.
- **Warriors:** The teams use a Hybrid Candidate Retrieval strategy with BM25 as lexical matching, combined with a Vietnamese Bi-encoder for dense embedding by the Reciprocal Rank Fusion (RRF) algorithm. Finally, the team uses the *itdainb/PhoRanker* as a re-ranker model to find the top-k candidates.
- **LAD:** The team implements a multi-stage retrieval system that combines BM25 with PhoBERT-based reranking to improve semantic relevance for selecting the top-5 most relevant candidates. Besides, the team also fine-tuned the BGE model on the ALQAC 2025 dataset by pair-wise question-article sets for the retrieval task.
- **BAMBOOVN:** The team integrates lexical and semantic

searching strategies with topic classification and relevant grader models in an adaptive RAG system. To perform the system, the team used the Ollama library with *qwen2.5:7b* and **bge-m3** models.

Evaluating metric: We use the F2-score for measuring the performance of proposed systems by participants. The F2 score is computed via precision and recall as follows:

$$Precision_i = \frac{\#correctly\_retrieved\_articles\_question\_i^{th}}{\#retrieved\_articles\_of\_question\_i^{th}} \quad (1)$$

$$Recall_i = \frac{\#correctly\_retrieved\_articles\_question\_i^{th}}{\#relevant\_articles\_of\_question\_i^{th}} \quad (2)$$

$$F2_{macro} = \frac{1}{N} \sum_1^N \frac{5 * Precision_i * Recall_i}{4 * Precision_i + Recall_i} \quad (3)$$

Overall, most teams approach the Retrieval task in a two-stage retrieval: utilize a lexical-based approach like BM25, for scoring to find top-k candidates, and then rank them by a dense embedding or LLMs. This year has witnessed the appearance of many Vietnamese text embedding models for legal, such as *AITeamVN/Vietnamese_Embedding*² and *AITeamVN/Vietnamese_Reranker*³ - the two text embedding models based on the FlagEmbedding [3] for the Vietnamese language, and *dangvantuan/vietnamese-document-embedding*⁴, *huyydangg/DEK21_hcmute_embedding*⁵, and *truuro7/vn-law-embedding*⁶ - the three text embedding models based on the SentenceBERT architecture [4] for Vietnamese. Along with Vietnamese embedding, other state-of-the-art multilingual embeddings like mBERT [5] and mE5 [6] are also used by participants. In addition, the teams also used several advanced methods like LightGBM [7] and GridSearch to improve the performance of the model as well as vector databases like Weaviate⁷, Qdrant⁸, and FAISS [8] for boosting up similarity search in the retrieval stage.

B. Task 2: Legal Question Answering

Task Description: Task 2 requires participants to generate the correct answer for a given legal question Q , accompanied by a set of relevant articles $S_1, S_2, \dots, S_N$. The questions in this task are presented in three formats: True/False, multiple-choice, and free-text [2]. For free-text questions, participants must provide a detailed written response, which is subsequently evaluated by legal experts for accuracy and appropriateness. The expected input and output format for Task 2 is illustrated below:

```
{
  "question_id": "<qid>",
  "question_type": "<qtype>",
  "text": "<Question>",
```

²https://huggingface.co/AITeamVN/Vietnamese_Embedding

³https://huggingface.co/AITeamVN/Vietnamese_Reranker

⁴<https://huggingface.co/dangvantuan/vietnamese-document-embedding>

⁵https://huggingface.co/huyydangg/DEK21_hcmute_embedding

⁶<https://huggingface.co/truuro7/vn-law-embedding>

⁷<https://weaviate.io/>

⁸<https://qdrant.tech/>

```
"relevant_articles": [
  {
    "law_id": "<law_id>",
    "article_id": "<article_id>"
  },
  ...
],
"answer": "<Answer>"
}
```

Participant Methods: There are 13 teams with 30 submission entries for the Legal Document Retrieval task. We summarize the method of 13 teams as follows:

- **NOWJ:** The team ensembles five different LLMs, including DeepSeek-R1-0528-Qwen3-8B, Qwen2.5-7B-Instruct, Qwen 2-7B-Instruct, LLaMA3-8B-Instruct, Gemma2-9B-it, and Gemma3-4B-it via different few-shot prompting techniques to extract the correct answers by voting among models.
- **Hallucineers:** The team uses *Qwen3-8B* with Group Relative Policy Optimization (GRPO) to enhance the thinking ability of models in answering the legal questions.
- **UIT-Darkcow:** The team leverages SentenceTransformer for encoding the question and the relevant law paragraph and finding the answer based on cosine similarity for the Yes/No and Multiple-choice questions. For free-text questions, they use a seq2seq QA model (T5 or BART) for generating the answers.
- **UTEHY-NLU:** The authors use Chain-of-thought prompting with *Qwen/Qwen2.5-7B-Instruct* and *Qwen3-8B* models to generate answers.
- **FlameReavers:** The team use *Qwen3-8B* model as LLMs for generating the answer.
- **vund:** The team use *Qwen3-8B* with inference prompting techniques for generating the answer.
- **BAMBOOVN:** The team uses LLM with a prompting technique (3 prompts for three kinds of questions) for question answering.
- **ABCD:** The team use three LLMs including *Qwen2.5-7B-Instruct*, *Llama-3.1-8B-Instruct*, and *gemma2-9b-it* with Tree-of-Thought (ToT) prompting for question answering task.
- **D2:** The team use *AITeamVN/GRPO-VI-Qwen2-7B-RAG* with few-shot prompting for free-text questions. For Yes/No and Multiple-choice questions, the team ensembles five LLMs, including *AITeamVN/GRPO-Vistral-7B-Chat-RAG*, *Qwen/Qwen2.5-7B-Instruct*, *AITeamVN/GRPO-VI-Qwen2-7B-RAG*, *Qwen/Qwen3-8B*, and *AITeamVN/Vi-Qwen2-7B-RAG* with major voting to determine the final label as answer.
- **VnLegalLLM25:** The team use *Qwen3-4B* and *Qwen3-8B* with thinking inference by multi-temperature strategy to generate answers.
- **Warriors:** The author used *VinaLLaMA-7B*, *VinaLLaMA-2.7B*, and *PhoGPT-4B* with the Retrieve-Augmented-Generation (RAG) pipeline for generating a corresponding answer to a legal question based on retrieved articles.

- **UIT legalGrep**: The team first segments the long article into smaller chunks, then selects the most relevant chunks by *Qwen/Qwen3-Reranker-8B*. Next, the team concatenates the selected chunks and uses *Mistral/Vistral-7B-Chat* for generating the output answers.
- **LAD**: The team uses *AITeamVN/GRPO-VI-Qwen2-1.5B-RAG* to generate answers for legal questions according to their relevant articles.

Evaluating metric: We use accuracy to measure the proportion of correct answers across the entire test set. To ensure the reliability of the evaluation, free-text responses are carefully reviewed and validated by legal experts. The accuracy formula is shown below:

$$\text{Accuracy} = \frac{\# \text{correctly\_answered\_questions}}{\# \text{total\_questions}} \quad (4)$$

For the question answering task, all approaches proposed by the participant teams are large language models (LLMs), indicating the robustness of LLMs in Vietnamese legal text processing tasks. Several popular LLMs, such as Qwen3 [9], Qwen2.5 [10], Llama3.1 [11], and DeepSeek [12], are mostly used by participants' systems, including the variants of those models for the Vietnamese language. Besides, several monolingual LLMs for the Vietnamese language only are also employed by participant teams like VinaLLama, Vistral, [13], and PhoGPT [14]. To instruct the LLMs to generate the answers for legal questions based on relevant articles, the participants usually use prompting techniques as Chain-of-Thought (CoT) [15], Tree-of-Thought (ToT) [16], and Retrieval-augmented-Generation (RAG) [17].

IV. RESULTS FROM PARTICIPANTS

A. Main results

Table IV shows the ranking of participating teams on the Retrieval Task (Task 1) in ALQAC 2025. The **Flame Reavers** achieved 1st place with 94.35% by F2 score in this task, while **vund** and **UIT legalGrep** both attained 2nd place with the F2 score of 89.70%. Almost 13 teams have submitted 2 to 3 runs in the private test phases of this task. Overall, the Retrieval task in ALQAC 2025 saw a significant improvement in the F2 score with the introduction of a robust embedding model for Vietnamese text.

For the question-answering task (Task 2), there is a change in the ranking among teams compared to Task 1. As shown in Table V, the three teams **NOWJ**, **Hallucineers**, and **UIT-Darkcow** achieved the 1st place in the question answering task with an Accuracy of 98.78%. Especially, **UIT-Darkcow** obtained the top 1 ranking only with 1 run. In general, the LLMs show their robust ability for the question-answering task with an impressive performance.

B. Discussion

Table VI summarizes the proposed method by the top-1 teams from 2021 to 2025 for two tasks. On the Retrieval task, the BERT model and its variants are efficient in encoding the text, which achieved high performance in ALQAC 2021 and

TABLE IV
TASK 1 FINAL RANKING OF PARTICIPANTS

Rank	Team	F2	Submission entries
1	Flame Reavers	0.9435	3
2	vund	0.8970	2
2	UIT legalGrep	0.8970	3
4	NOWJ	0.8767	3
5	UTEHY-NLU	0.8482	3
6	Hallucineers	0.8279	3
7	D2	0.8238	3
8	UIT-Darkcow	0.7873	3
9	ABCD	0.7730	3
10	Warriors	0.6897	3
11	LAD	0.6589	2
12	BAMBOOVN	0.5556	3
13	VnLegalLLM25	0.5190	3

TABLE V
TASK 2 FINAL RANKING OF PARTICIPANTS

Rank	Team	Accuracy	Submission entries
1	NOWJ	0.9878	3
1	Hallucineers	0.9878	3
1	UIT-Darkcow	0.9878	1
4	UTEHY-NLU	0.9756	3
5	FlameReavers	0.9634	1
5	vund	0.9634	1
7	BAMBOOVN	0.9512	3
8	ABCD	0.9390	2
8	D2	0.9390	3
10	VnLegalLLM25	0.9024	3
11	Warriors	0.8171	2
12	UIT legalGrep	0.7561	3
13	LAD	0.3049	1
14	LAD	0.6341	2*

* unofficial submission

ALQAC 2022. Besides the retrieval task, BERT also showed impressive performance for the question answering task in ALQAC 2021 and ALQAC 2022, where BERT-based models for the Vietnamese language, as VNLawBERT [18] and PhoBERT [19], attain high results in the question answering task.

Beginning in ALQAC 2023, the two-stage retrieval paradigm for legal document retrieval is mostly chosen by participants to propose systems for retrieval tasks. The two-stage retrieval model begins with a scoring model like BM25 - a lexical matching method to assess the semantic similarity between the input question and a set of candidate documents, then re-rank the results with a language model such as mBERT (team NeCo [20] in ALQAC 2023) or even combined between a language model with LLM like ViT5+GPT-4 (team NOWJ [21] at ALQAC 2024). In AQLAC 2025, the Vietnamese text embedding models based on the BGE FlagEmbedding architecture [3] are mostly used by the participants and achieved the best performance (team Flame Reavers at AQLAC 2025, according to Table IV), indicating the efficiency and robustness of FlagEmbedding for text embedding in the retrieval task.

Finally, for the Question answering task, the Large-language model (LLMs), even commercial or open-source LLMs, show the state-of-the-art performance since ALQAC 2023.

TABLE VI
PROPOSED METHOD FOR ALQAC FROM 2021 TO 2025

Competition year	Task 1 - Legal Retrieval		Task 2 - Question Answering	
	Method	F2	Method	Acc
ALQAC 2021	VnLawBERT	0.8807	VNLawBERT	0.7102
ALQAC 2022	RoBERTa	0.9667	PhoBERT+BERT-QA	0.6333
ALQAC 2023	BM25+mBERT	0.8661	Flan-Alpaca-XXL	0.8637
ALQAC 2024	BM25+ViT5+GPT-4	0.8771	GPT-4	0.9800
ALQAC 2025	VietnameseBGE + Qwen3-8B	0.9435	Open source LLMs and LM	0.9878

V. CONCLUSION

This paper summarizes the fifth edition of the Automated Legal Question Answering Competition (ALQAC 2025), continuing the mission to advance automated legal text processing for low-resource languages, particularly Vietnamese. Building on the progress of previous years, ALQAC 2025 introduced additional innovations, such as an expanded focus on scenario-based reasoning and a more rigorous evaluation framework combining expert-verified and non-expert-verified data. The competition attracted increased participation, resulting in a diverse array of approaches to both Legal Document Retrieval and Legal Question Answering tasks. The performance of the submitted systems demonstrates not only ongoing improvements in model accuracy and generalization, but also the growing sophistication of methodologies tailored to real-world legal challenges. Collectively, the results of ALQAC 2025 underscore the continued momentum in the development of legal AI systems and reinforce the vital role of collaborative benchmarking in fostering research for under-resourced languages. The advancements achieved this year provide a strong foundation for future innovations, bringing us closer to practical, accessible legal support tools for diverse linguistic communities.

REFERENCES

- [1] R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka, "Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024," in *JSAI International Symposium on Artificial Intelligence*. Springer, 2024, pp. 109–124.
- [2] D.-T. Do, S. T. Luu, T. Pham, T. Vo, N.-H. Chu, Q.-H. Chu, C. Nguyen, M. Nguyen, A. Trieu, D. Nguyen *et al.*, "A summary of the alqac 2024 competition," in *2024 16th International Conference on Knowledge and System Engineering (KSE)*. IEEE, 2024, pp. 422–427.
- [3] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2318–2335.
- [4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [6] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," *arXiv preprint arXiv:2402.05672*, 2024.
- [7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
- [8] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," *arXiv preprint arXiv:2401.08281*, 2024.
- [9] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [10] Q. Team, "Qwen2.5 technical report," 2024.
- [11] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
- [12] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.
- [13] C. Van Nguyen, T. Nguyen, Q. Nguyen, H. Nguyen, B. Plüster, N. Pham, H. Nguyen, P. Schramowski, and T. Nguyen, "Vistral-7b-chat-towards a state-of-the-art large language model for vietnamese," 2023.
- [14] D. Nguyen, L. T. Nguyen, C. Tran, D. N. Nguyen, D. Phung, and H. Bui, "Phogpt: Generative pre-training for vietnamese," *arXiv preprint arXiv:2311.02945*, 2023.
- [15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [16] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023.
- [17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
- [18] C.-N. Chau, T.-S. Nguyen, and L.-M. Nguyen, "Vnlawbert: A vietnamese legal answer selection approach using bert language model," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, 2020, pp. 298–301.
- [19] D. Q. Nguyen and A. Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042.
- [20] H.-L. Nguyen, D.-Q. Nguyen, H.-T. Nguyen, T.-T. Pham, H.-D. Nguyen, T.-A. Nguyen, T.-H.-Y. Vuong, and H.-T. Nguyen, "Neco@alqac 2023: Legal domain knowledge acquisition for low-resource languages through data enrichment," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6.
- [21] N.-D. Mai, X.-B. Le, M.-T. Nguyen, H.-D. Nguyen, V.-H. Nguyen, H.-T. Nguyen, and T.-H.-Y. Vuong, "Nowj@alqac 2024: Leveraging document ranking methods and large language models for legal document retrieval and question answering," in *2024 16th International Conference on Knowledge and System Engineering (KSE)*, 2024, pp. 447–452.

H-OCR: A Hybrid OCR Approach to Vietnamese Scanned Administrative Text Recognition

Bui Cong Thanh, Nguyen The Hung, Nguyen Trung Dung, Nguyen Anh Tu,
Nguyen Van Anh, Le Thi Quynh, Le Xuan Duc

National Laboratory for Securing Information (NLSI), Vietnam Ministry of Defence, Hanoi, Vietnam
congthanhtmt@gmail.com, hungnguyenthe1211@gmail.com, ntdtoanud2011@gmail.com, nguyenanhthu789@gmail.com,
nguyenanhb14@gmail.com, lequynh.k51.ktqs@gmail.com, lexuanducvn@gmail.com

Abstract—This paper proposes an enhanced hybrid OCR (H-OCR Enhanced) that builds upon a foundational hybrid system (H-OCR Base), combining VietOCR’s Vietnamese-optimized recognition and PaddleOCR’s reliable text detection. For degraded Vietnamese administrative documents, H-OCR Enhanced incorporates thorough preprocessing and NLP-based post-processing. The proposed approach uses an integrated pipeline that combines picture enhancement, stamp removal, modular text detection-recognition, and domain-adapted language model correction to overcome major issues, including old paper quality, overlapping official stamps, and complex diacritical marks. H-OCR Enhanced achieves 6.82% Word Error Rate (WER) and 2.31% Character Error Rate (CER), significantly outperforming H-OCR Base (10.15% WER) and other comparative methods like Tesseract, EasyOCR, and popular commercial OCR services, according to experimental validation on a dataset of 300 full-text pages of gathered real-world Vietnamese administrative documents. By offering a strong framework that manages document deterioration, stamp interference, and linguistic complexity in a cohesive manner, the findings advance the automated digitization of Vietnamese administrative archives.

Index Terms—OCR, PaddleOCR, VietOCR, Hybrid OCR, LLM, Administrative Documents

I. INTRODUCTION

In Vietnam, modernizing government processes and improving public service accessibility depend heavily on the digitization of administrative documents. These records frequently include important legal and regulatory data that need to be precisely digitized to guarantee effective processing, retrieval, and storage. Because it facilitates automated data extraction, makes cross-referencing legal documents easier, and allows sophisticated analytics for policymaking and compliance monitoring, optical character recognition (OCR) technology is highly valued in this context [1]. However, in research [2], the authors state that due to overlapping official stamps, a combination of modern and old fonts, legal terminology, different scan quality, and inconsistent digitization standards, Vietnamese administrative documents provide special hurdles for OCR techniques. Recognition is further complicated by the intricate system of Vietnamese diacritical marks, which OCRs designed for other Latin-based languages struggle to accurately capture. Older documents and low-resolution scans

(such as 200 dots per inch (DPI)) make this problem worse. These factors result in blurrier images, lower signal-to-noise ratio (SNR), and less distinguishable text.

In recent years, OCRs have embraced modular architectures that provide them with more flexibility and adaptation to a variety of document types by separating text detection from recognition [3]. A notable change from previous end-to-end models to more reliable and adaptable pipelines is represented by this trend. Several OCR tools have been made available for Vietnamese text recognition, including Tesseract [4], EasyOCR [5], and VietOCR [6]. Although Tesseract and EasyOCR are popular and lightweight, they have issues with sensitivity to scan quality and incorrect recognition of diacritical marks [2]. VietOCR, which was created especially for Vietnamese, improves accuracy on printed documents but is less useful for end-to-end automation on low-quality inputs because it does not have integrated text identification and post-processing capabilities. These drawbacks emphasize the necessity of a more reliable integrated pipeline designed specifically for the examination of Vietnamese administrative documents.

Text detection is the first step in the modular OCR architecture and is used to locate textual areas in document pictures before recognition processing. Modern detection methods have evolved from conventional edge-based techniques to deep learning architectures capable of handling complex layouts and varied text orientations. PaddleOCR [7] has demonstrated strong performance for low-resource languages and multilingual settings, with its Differentiable Binarization (DB) detection approach achieving state-of-the-art results in speed and accuracy on benchmark datasets. The PaddleOCR system’s ongoing improvements, as detailed in technical publications like PaddleOCR 3.0 [8], also reinforce its standing as a top OCR toolkit, frequently exceeding other well-liked substitutes in practical applications.

Furthermore, specific methods for document pretreatment and post-processing have grown in significance for difficult administrative documents, going beyond conventional OCR components. Promising outcomes in maintaining underlying text while eliminating visual distractions have been demonstrated by deep learning algorithms for document-specific challenges, such as stamp removal utilizing YOLOv5 recogni-

tion and U-Net inpainting [9]. We use self-supervised learning to adapt the Qwen3-8B language model to the Vietnamese administrative domain for post-processing enhancement [10].

In this work, we propose an enhanced hybrid OCR (H-OCR Enhanced) method that integrates the power of VietOCR and PaddleOCR text identification. The pipeline incorporates comprehensive preprocessing (grayscale conversion, morphological denoising, binarization, deskewing), stamp removal, and NLP post-processing. By tackling the junction of document deterioration, stamp interference, and linguistic complexity, experimental validation on the dataset of gathered real-world Vietnamese administrative documents shows notable gains in recognition accuracy when compared to current solutions.

The remainder of the paper is organized in this manner. Section II of Related study looks at the latest developments in OCR technology and Vietnamese language processing, while Section III describes our suggested method that offers a general hybrid pipeline and algorithm for processing PDF documents and extracting Vietnamese text. Section IV of the experimental design provides baseline comparisons with current OCRs, whereas Section V highlights the results and discussion. Section VI displays the conclusion.

II. RELATED WORK

OCR technology has progressed from early rule-based and template-matching systems to statistical learning models like SVMs and HMMs, and then to contemporary deep-learning architectures [11]. Tesseract, which was first created by HP and then improved by Google with LSTM-based recognition capabilities, became one of the most popular OCR engines among these advancements [4]. This was expanded upon with the release of EasyOCR, a Python-based framework that makes use of CNN-RNN-CTC architectures. In multilingual text recognition tasks, it has shown competitive performance and supports more than 80 languages [5].

Vietnamese OCR remains challenging due to complex diacritical markings that generic algorithms often misrecognize. Early practitioners relied on general-purpose tools (Tesseract, EasyOCR), which struggled with Vietnamese diacritics in administrative documents [2], until Vietnamese-specific models like VietOCR emerged. Modern systems adopt modular architectures separating detection and recognition stages, improving adaptability to varied document types [3]. One well-known example is PaddleOCR, which combines cutting-edge recognition modules with a Differentiable Binarization (DB) detection methodology to get impressive results on documents with intricate layouts and multilingual content. Similar to this, YOLO (You Only Look Once) architectures have drawn interest in text recognition jobs because of their strong performance in localizing text regions inside intricate document layouts and their real-time processing capabilities [12]. Specialized frameworks such as PaddleOCR's DB approach frequently offer better accuracy for text-specific identification applications, even though YOLO is excellent at generic object detection [7].

Hybrid OCR systems, which mix deep learning and conventional methods to improve accuracy and resilience, have been studied recently. Text detection is frequently approached as instance segmentation using techniques such as Mask R-CNN and DBNet, or as object identification using models like YOLO. Many systems use a two-stage pipeline that combines CNN-based detectors with attention-based decoders or sequence models like RNN-CTC. Furthermore, rule-based or NLP-based post-processing is frequently used in hybrid pipelines to fix OCR problems, especially in documents that are domain-specific. [13] provides a detailed survey of these methods.

Although PaddleOCR's detection component is still quite competitive, its Vietnamese-specific recognition models are not. VietOCR, an OCR designed especially for Vietnamese, on the other hand, uses recognition architectures that are either transformer-based or CRNN-based and have been optimized for Vietnamese corpora. On Vietnamese papers, it routinely performs better than general-purpose recognizers, especially when it comes to managing tone marks and context-dependent errors.

While existing Vietnamese OCR research has made significant contributions, substantial gaps remain for administrative document processing. Vietnamese handwritten character recognition systems focus primarily on individual character classification with CNN architectures achieving 97.2% accuracy on clean handwriting datasets [14], but these approaches cannot address the complex challenges of full document processing. Recent hybrid OCR developments, such as the MC-OCR Challenge system combining CRAFT detection with VietOCR recognition for receipt processing [15], demonstrate the potential of modular architectures but remain limited to structured commercial documents with basic BERT fine-tuning for post-processing. While Vietnamese administrative documents exhibit the previously discussed challenges, they introduce a further critical requirement inadequately addressed by existing systems: maintaining temporal consistency across decade-long document collections. Current Vietnamese OCR solutions only partially address these challenges, and comprehensive approaches tailored to administrative contexts are still lacking. Furthermore, while recent hybrid approaches employ basic post-processing techniques, exploration of domain-adapted large language models for contextual error correction in Vietnamese governmental settings has so far been limited. Our H-OCR Enhanced addresses these critical gaps by introducing the first comprehensive hybrid OCR framework specifically designed for Vietnamese administrative document digitization, incorporating advanced preprocessing pipelines, intelligent stamp removal capabilities, and domain-adapted language model post-processing to achieve state-of-the-art performance on real-world governmental archives.

III. H-OCR ENHANCED: A PROPOSED HYBRID OCR APPROACH

In light of the unique features of Vietnamese administrative documents outlined earlier, we propose H-OCR Enhanced, a

comprehensive hybrid OCR system that combines VietOCR’s Vietnamese-optimized recognition with PaddleOCR’s robust text detection. Our approach incorporates comprehensive preprocessing, intelligent stamp removal, and domain-adapted NLP post-processing to effectively address document degradation, stamp interference, and linguistic complexity.

Figure 1 illustrates the complete architecture, while Algorithm 1 details the implementation steps. The pipeline operates through four sequential stages, each addressing specific challenges identified in baseline OCR systems.

Algorithm 1 H-OCR Enhanced: Hybrid OCR for Vietnamese Administrative Documents

Require: Document images $\mathbf{I} = \{I_1, \dots, I_N\}$ where $I_i \in \mathbb{R}^{H \times W \times 3}$

Ensure: Vietnamese text $T \in \mathbb{S}^*$

```

1: Stage 1: Preprocessing
2: for each  $I_i \in \mathbf{I}$  do
3:    $I_i \leftarrow \text{Grayscale} \rightarrow \text{MorphDenoise} \rightarrow \text{OtsuBinarize} \rightarrow \text{Deskew}(I_i)$ 
4: end for
5: Stage 2: Stamp Removal
6: for each  $I_i$  do
7:    $\mathbf{S}_i \leftarrow \text{YOLOv5.detect\_stamps}(I_i)$ 
8:   if  $\mathbf{S}_i \neq \emptyset$  then
9:      $I_i \leftarrow \text{UNet.inpaint}(I_i, \mathbf{S}_i)$ 
10:  end if
11: end for
12: Stage 3: Text Detection & Recognition
13: for each  $I_i$  do
14:    $\mathbf{R}_i \leftarrow \text{PaddleOCR.detect}(I_i)$  {DB method}
15:    $T_i \leftarrow \text{VietOCR.recognize}(\mathbf{R}_i)$  {Transformer}
16: end for
17: Stage 4: NLP Post-processing
18:  $T \leftarrow \text{Qwen3-8B.correct}(\text{SpatialSort}(\bigcup T_i))$ 
19: return  $T$ 

```

A. Stage 1: Preprocessing

For processing multi-page documents, H-OCR Enhanced begins with PDF-to-image conversion. The preprocessing stage (Lines 2-4 in Algorithm 1) implements a two-phase enhancement strategy. First, images undergo grayscale conversion to reduce computational complexity while preserving text features. Morphological denoising with 2×2 kernels removes salt-and-pepper noise common in aged documents, followed by Otsu binarization for adaptive thresholding that handles varying illumination conditions across document regions. Second, the deskewing step corrects skew angles up to ± 30 degrees using Hough Line Transform and affine transformation, ensuring proper text alignment for subsequent detection. This comprehensive preprocessing is critical for handling the degraded quality of Vietnamese administrative archives.

B. Stage 2: Stamp Removal

Official stamps represent a unique challenge, as they frequently overlap critical text content. Stage 2 (Lines 6-10) addresses this through a two-step process: YOLOv5-based stamp detection identifies stamp regions as bounding boxes, followed by U-Net inpainting that reconstructs the underlying text. Unlike simple masking approaches, the U-Net architecture learns to intelligently fill stamp regions by leveraging contextual information from surrounding pixels, effectively removing visual interference while preserving the legibility of overlapped characters. This stage significantly reduces recognition errors in stamp-affected regions.

C. Stage 3: Text Detection and Recognition

The core OCR functionality combines PaddleOCR’s detection capabilities with VietOCR’s specialized recognition (Lines 12-15). PaddleOCR’s Differentiable Binarization (DB) method adaptively determines binarization thresholds at the pixel level, achieving robust text localization even in complex layouts with varying text orientations and scales. For each detected region $\mathbf{R}_i$, VietOCR’s Transformer-based architecture performs character recognition optimized for Vietnamese diacritical marks. This hybrid approach leverages the strengths of both systems: PaddleOCR’s proven performance on multilingual text detection and VietOCR’s language-specific optimization for Vietnamese character recognition.

D. Stage 4: NLP Post-processing

The final stage (Lines 17-18) performs linguistic correction using a domain-adapted Qwen3-8B language model. Text regions are first spatially sorted by coordinates to reconstruct proper reading order across multi-column layouts. The Qwen3-8B model, fine-tuned through continued pre-training with masked language modeling on Vietnamese administrative corpora, corrects OCR errors, including diacritical mark mistakes, spacing inconsistencies, and character substitutions. By internalizing bureaucratic terminology and formal syntactic patterns characteristic of governmental documentation, the model achieves context-aware correction that simple rule-based methods struggle to match. This self-supervised adaptation enables the model to handle domain-specific vocabulary without requiring extensive labeled training data. In this work, we integrate Qwen3-8B, a strong large language model, into our pipeline to demonstrate the effectiveness of the proposed hybrid OCR framework when combined with domain-adapted language modeling. We intentionally focus on validating the pipeline design for Vietnamese administrative documents, whose challenges (degraded scans, official stamps, complex diacritics) are not directly reflected in existing Vision Language Model (VLM) benchmarks. Thus, rather than exhaustively benchmarking against multiple state-of-the-art vision–language models, we adopt Qwen3-8B as a representative backbone to highlight the improvements brought by our pipeline. A broader comparison with recent VLMs will be considered in future work.

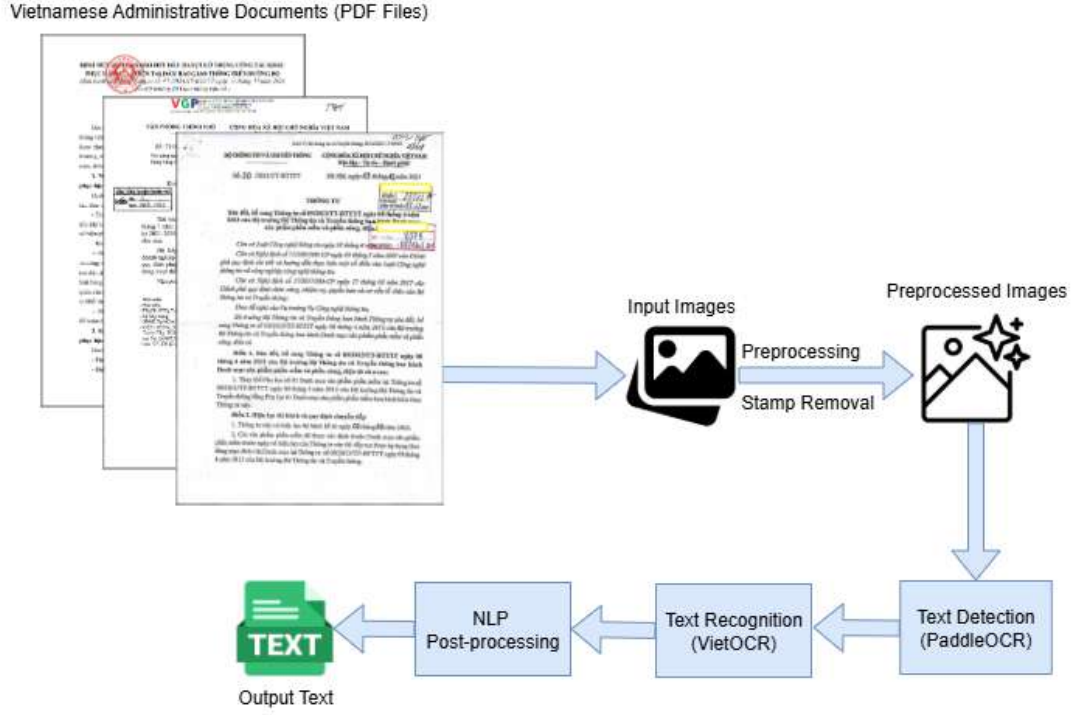


Figure 1: Proposed H-OCR Enhanced approach for Vietnamese administrative document OCR, incorporating preprocessing, text detection, recognition, and post-processing stages

Computational Complexity: $O(N \cdot W \cdot H + \sum_{i=1}^N K_i \cdot C_{OCR} + |T'| \cdot C_{NLP})$, where $N \cdot W \cdot H$ represents image preprocessing and detection across N pages of size $W \times H$, $\sum_{i=1}^N K_i \cdot C_{OCR}$ denotes text recognition for K_i regions per page with VietOCR complexity C_{OCR} , and $|T'| \cdot C_{NLP}$ corresponds to NLP post-processing on text sequence of length $|T'|$ with Qwen3-8B complexity C_{NLP} .

IV. EXPERIMENTAL DESIGN

A. Dataset Description

1) *Dataset for Training LLM:* We gathered a training corpus of 100,000 articles from official Vietnamese government portals and agency websites spanning 2015-2024. Sourcing from these authoritative government channels ensures the corpus reflects authentic administrative writing conventions, including formal vocabulary, bureaucratic terminology, and syntactic structures typical of Vietnamese governmental documentation. This enables the model to learn administrative language patterns through continued pre-training with Masked Language Modeling (MLM)¹, subsequently facilitating accurate OCR error correction without supervised annotation.

2) *Dataset for OCRs Validation:* Our validation dataset comprises 300 newly collected and manually annotated pages of Vietnamese administrative documents. The dataset was systematically constructed by selecting the first page from the

first 30 documents per year spanning 2015-2024, sourced from [16]. The systematic selection of the first 30 documents per year eliminates selection bias. This chronological sampling ensures temporal coverage while capturing the evolution of administrative language patterns, formatting standards, and digitization practices over the past decade. Documents include official correspondence, regulatory texts, legal notices, and bureaucratic forms typical of Vietnamese government operations. All annotations were performed by native Vietnamese speakers with administrative domain expertise to ensure reliable ground truth for OCR evaluation. This dataset addresses a significant gap in Vietnamese OCR research, as no comparable temporally distributed collection was previously available for systematic evaluation in this domain.

B. Experimental Setups

Several OCR settings are compared with H-OCR Base for the recognition of Vietnamese administrative text. The investigation aims to determine whether H-OCR Base can be used to develop a system that tackles the unique difficulties associated with administrative text processing.

As indicated in Table I, the experimental setup uses a two-stage pipeline in which PaddleOCR first identifies text sections as polygonal bounding boxes, which are subsequently sent to different recognition systems for character identification. These setups can methodically assess several configurations and determine the best setup for Vietnamese administrative papers thanks to this modular architecture.

¹Training hyperparameters: batch size 12, gradient accumulation 3, max sequence length 4096, learning rate 6e-6, warmup steps 500, weight decay 0.01, 2 epochs, masking ratio 0.15, cosine annealing scheduler.

Table I: OCR Pipelines with Recognition Models, Descriptions, and Recognition-Specific Parameters

Pipeline	Recognition Model	Description	Recognition Parameters
H-OCR Base	Transformer (VGG19 backbone)	CNN-based feature extractor (VGG19) + Transformer encoder-decoder (seq2seq), no pretrained weights	<code>config = Cfg.load_config_from_name('vgg_transformer'), pretrained=False, device='cpu'</code>
PaddleOCR + Tesseract	Tesseract (LSTM-based)	OCR engine using LSTM-based seq2seq model (OEM=1), supports multiple page segmentation modes	<code>lang='vie', psm=6, oem=1</code>
PaddleOCR + EasyOCR	CRNN with CTC decoder	Deep learning-based OCR using a CRNN backbone and CTC decoding	<code>lang_list=['vi'], decoder='CTC'</code>

Note: All pipelines use identical PaddleOCR detection settings: `use_doc_orientation_classify=False` (no rotation correction), `use_doc_unwarping=False` (no perspective correction), `use_textline_orientation=False` (no text line rotation), and `lang='vi'` (Vietnamese).

We also examined a number of well-known commercial OCR providers in order to guarantee impartial assessment and set performance benchmarks. Notably, Convertio OCR², 2PDF Vietnamese OCR³, and Google Docs' built-in OCR feature demonstrate robust accuracy and reliable diacritical mark recognition for general Vietnamese documents. These tools represent significant advances in Vietnamese OCR technology and have proven effective for everyday use cases. However, the specialized requirements of Vietnamese administrative documents necessitate tailored approaches beyond general-purpose commercial solutions.

C. Evaluation Metrics

We use standard OCR evaluation metrics:

1) Character Error Rate (CER):

$$CER = \frac{S + D + I}{N} \times 100\% \quad (1)$$

where S , D , I are character-level substitutions, deletions, and insertions, and N is the total number of characters.

2) Word Error Rate (WER):

$$WER = \frac{S_w + D_w + I_w}{N_w} \times 100\% \quad (2)$$

where subscript w denotes word-level operations.

3) *Processing Time*: We calculate how long it takes to process the complete 300-page test dataset. Since administrative documents are typically digitized in batches rather than in real-time, this statistic provides insight into the computing demands and scalability of the systems, even though it is not essential for our use case. Generally speaking, accuracy is more crucial than speed when digitizing ancient materials.

V. RESULTS AND DISCUSSION

The performance comparison of several OCRs on our administrative document dataset is shown in Table II. Table II shows that, although having the greatest processing time (1:19), H-OCR Base performs the best overall, with the lowest WER of 10.15% and CER of 4.88%. On the other hand, online providers like Google Docs, Convertio OCR, and 2PDF Vietnamese OCR have greater mistake rates (WERs ranging from 16.32% to 25.94%) but faster processing times (between 00:08 and 00:46). Notably, clean text sections are typically

well-known, whereas stamp-affected regions account for the majority of errors in online services, as shown in Table II.

Table II: Performance Comparison of OCRs on Vietnamese Administrative Documents

OCR	WER (%)	CER (%)	Average Time (mm:ss)
H-OCR Base	10.15	4.88	01:19
PaddleOCR + Tesseract	22.44	12.60	00:49
PaddleOCR + EasyOCR	48.69	36.05	00:30
Convertio OCR	25.94	10.57	*
2PDF Vietnamese OCR	17.89	8.02	*
Google Docs	16.32	7.69	*

* Commercial services excluded from processing time comparison due to infrastructure and network variability.

This implies that while the models can handle the particular difficulties of Vietnamese administrative papers, they are not strong enough. Tesseract and EasyOCR are unsuitable for this site because they have trouble reading Vietnamese text, attaining WERs of 22.44% and 48.69%, respectively. These findings lead us to choose H-OCR Base as the foundation of our improved pipeline due to its high accuracy and reliable Vietnamese text recognition capabilities.

The outcome of H-OCR Base, combining with preprocessing, stamp removal, and NLP post-processing, is displayed in Table III.

Table III: Performance of Enhanced Pipeline on Administrative Documents

Configuration	WER (%)	CER (%)	Average Time (mm:ss)
H-OCR_Base	10.15	4.88	01:19
PP-H-OCR_Base-SR	8.60	3.27	01:27
PP-H-OCR_Base-SR-NLP	7.45	2.68	01:52
PP-H-OCR_Base-SR-NLP_Retrained (H-OCR Enhanced)	6.82	2.31	01:58

Abbreviations: PP: Preprocessing; H-OCR_Base: our H-OCR Base; SR: Stamp Removal; NLP: Natural Language Processing.

With a WER of 6.82% and a CER of 2.31%, the configuration that combines preprocessing, H-OCR Base, and retrained NLP post-processing performs best for Vietnamese administrative document recognition, as indicated in Table III. In comparison to the H-OCR Base, this indicates a significant improvement of 32.8% in WER (from 10.15% to 6.82%) and 52.7% in CER (from 4.88% to 2.31%). The H-OCR Enhanced exhibits notable advancements in managing the difficulties

²<https://convertio.co/ocr/>

³<https://2pdf.com/vi/ocr/vietnamese/>

posed by substandard scans, stamps, and administrative jargon. The significant decrease in error rates makes this the best option for government document digitization projects where accuracy is more important than speed, even if it takes 39 more seconds to complete (from 1:19 to 1:58).

We also examine the performance of OCRs on Vietnamese administrative documents from 2015 to 2024, from year to year, monitoring changes in WER and CER.

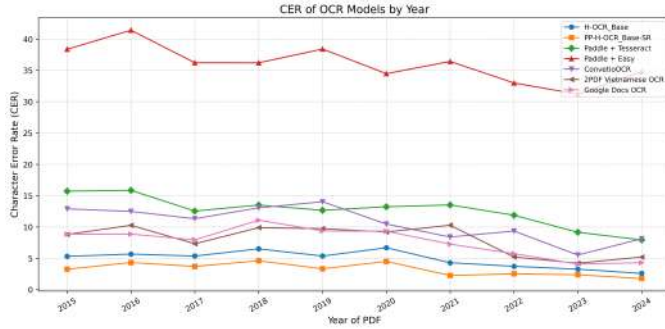


Figure 2: CER trends across different OCR configurations on Vietnamese administrative documents from 2015 to 2024

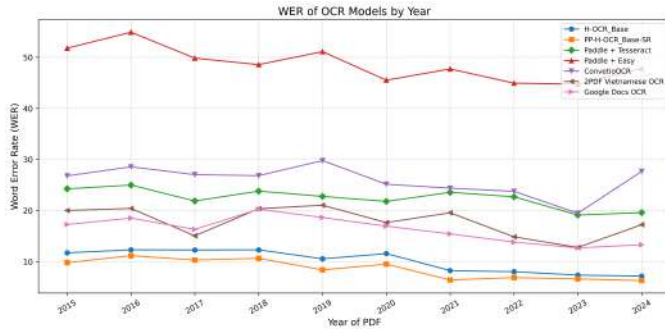


Figure 3: WER trends across different OCR configurations on Vietnamese administrative documents from 2015 to 2024

According to the data, H-OCR Base and its improved counterparts exhibit a discernible lower trend in WER and CER between 2020 and 2024, suggesting steady progress over time. Tesseract, EasyOCR, ConvertioOCR, and 2PDF Vietnamese OCR, on the other hand, exhibit inconsistent performance since they don't show a clear declining trend. Although Google Docs OCR's overall accuracy is still lower than H-OCR Base's, it has gradually improved, particularly around 2020. All things considered, our suggested configuration shows more consistent gains over time as well as reduced mistake rates.

The results highlight several key observations. With 10.15% WER and 4.88% CER on difficult administrative documents, H-OCR Base performs the best. By incorporating stamp removal, preprocessing, and NLP post-processing, H-OCR Enhanced reduces error rates to 6.82% WER and 2.31% CER, a notable improvement for real-world implementation.

VI. CONCLUSION

We present H-OCR Enhanced, a hybrid OCR pipeline that effectively addresses the unique challenges of Vietnamese administrative documents: diacritical complexity, official stamp interference, and document degradation. By combining PaddleOCR detection, VietOCR recognition, U-Net-based stamp removal, and domain-adapted language modeling with Qwen3-8B, our system achieves 6.82% WER and 2.31% CER—32.8% and 52.7% improvements over baseline methods. These results demonstrate the effectiveness of domain-specific pipeline design for specialized document processing. Future work will explore the interactions between pipeline components, investigate optimal configurations, and expand evaluation to additional vision-language models for diverse Vietnamese administrative document types.

References

- [1] S. V. Nguyen, D. A. Nguyen, and L. S. Q. Pham, "Digitalization of administrative documents a digital transformation step in practice," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, Hanoi, Vietnam, 2021, pp. 519–524.
- [2] D. A. Nguyen, S. V. Nguyen, and L. S. Q. Pham, "A survey on vietnamese document analysis and recognition: Challenges and future directions," *arXiv preprint arXiv:2506.05061*, 2025. [Online]. Available: <https://arxiv.org/abs/2506.05061>
- [3] Y. Liu, S. Zhou, S. Yang, X. Zhang, C. Zhang, and C. Yao, "Bridging the gap between end-to-end and two-step text spotting," *arXiv preprint arXiv:2404.04624*, 2024. [Online]. Available: <https://arxiv.org/abs/2404.04624>
- [4] T. O. contributors, "tesseract: Open source ocr engine," GitHub, 2025, accessed July 8, 2025. [Online]. Available: <https://github.com/tesseract-ocr/tesseract>
- [5] JaidedAI, "Easyocr: Ready-to-use ocr with 80+ supported languages," GitHub Repository, 2021. [Online]. Available: <https://github.com/JaidedAI/EasyOCR>
- [6] P. B. C. Quoc, "Vietocr – vietnamese text recognition using transformer and attentionocr," Blog, 2025, accessed: June 30, 2025. [Online]. Available: <https://pbccquoc.github.io/vietocr/>
- [7] M. Liao, X. Bai, W. Shen, Z. Yang, S. Wang, and W. Jia, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11 477–11 484.
- [8] B. Research, "Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system," *arXiv preprint arXiv:2206.03001v2*, 2022.
- [9] S.-A. Research, "stamp_processing," GitHub repository, 2025, accessed July 8, 2025. [Online]. Available: https://github.com/sun-asterisk-research/stamp_processing
- [10] J. Bai *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
- [11] M. A. Islam and M. M. R. Chowdhury, "A review on optical character recognition," in *2018 International Conference on Computing, Communication, and Automation (ICCCA)*, Greater Noida, India, 2018, pp. 1–6.
- [12] Y. He, K. Wang, and X. Bai, "Text detection in natural scenes with improved yolov4 network," *Sensors*, vol. 21, no. 3, p. 888, 2021. [Online]. Available: <https://doi.org/10.3390/s21030888>
- [13] K. Subramani, A. A. Mohammed, P. R. Saligram, and S. Setlur, "A survey of deep learning approaches for ocr and document understanding," *arXiv preprint arXiv:2011.13534*, 2020. [Online]. Available: <https://arxiv.org/abs/2011.13534>
- [14] T. Q. Vinh, L. H. Duy, and N. T. Nhan, "Vietnamese handwritten character recognition using convolutional neural network," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 276–283, 2020.
- [15] T. B. Hieu *et al.*, "Mc-ocr challenge 2021: A multi-modal approach for mobile-captured vietnamese receipts recognition," in *2021 RIVF*. IEEE, 2021.
- [16] Government of Vietnam. (2025) Official documents. [Accessed: Sep. 27, 2025]. [Online]. Available: <https://vanban.chinhphu.vn/>

Hallucineers@ALQAC 2025: Unified LLM Pipelines for Legal Document Filtering and Question Answering

Quan Minh Le^{1,2}, Phong Tuan Nguyen^{1,2}, Dat Tien Truong^{1,2}, Tung Le^{1,2}, Huy Tien Nguyen^{1,2,†}

¹Faculty of Information Technology, University of Science, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

{22120291, 22127326, 22120060}@student.hcmus.edu.vn

{lftung, ntienhuy}@fit.hcmus.edu.vn

[†]Corresponding Author: Huy Tien Nguyen (ntienhuy@fit.hcmus.edu.vn)

Abstract—The intersection of law and natural language processing presents significant opportunities to enhance legal services through computational methods. The Automated Legal Question Answering Challenge (ALQAC) provides a platform to advance research in legal text understanding. This paper presents our system for the ALQAC 2025 challenge, designed for legal document retrieval (Task 1) and question answering (Task 2) under strict model size limits. Our approach features a fine-tuned dense retriever for document retrieval and a compact large language model for question answering, enhanced by a two-stage instruction fine-tuning pipeline and Chain-of-Thought reasoning. Our results demonstrate the effectiveness of this resource-efficient design, achieving a first-place finish in question answering and a competitive sixth-place rank in document retrieval. This work highlights the potential of smaller, specialized models to achieve top-tier performance in complex legal domains.

Index Terms—Legal NLP, Question answering, Document Retrieval, Large Language Models, Prompt Engineering

I. Introduction

The application of automated systems in the legal field is crucial for enhancing the efficiency and accessibility of legal services, particularly for semantically complex languages like Vietnamese. To foster research in this domain, the annual Automated Legal Question Answering Competition (ALQAC) [1]–[4] serves as a key benchmark. This paper details the system developed by our team, **Hallucineers**, for the 5th ALQAC in 2025, which featured two tasks: Legal Document Retrieval and Legal Question Answering.

A defining challenge of ALQAC 2025 was the restriction to open-weight Large Language Models (LLMs) under 10 billion parameters, prohibiting the use of large proprietary models. This constraint shifted the focus towards methodological innovation rather than reliance on model scale. In response, we designed a multi-stage system. For document retrieval, we developed a pipeline centered on a fine-tuned dense retriever. For question answering, our approach leveraged a compact LLM enhanced by a two-stage instruction fine-tuning process and advanced prompt engineering techniques, including Chain-of-Thought reasoning.

Our system demonstrated remarkable effectiveness, culminating in a **first-place victory in the Legal Question Answering task**. Our main contributions are:

- We present a winning methodology for Task 2, achieving a **98.78%** accuracy. This was realized through a novel two-phase training strategy (Supervised Fine-Tuning followed by Group Relative Policy Optimization) on an LLM augmented with generated Chain-of-Thought reasoning steps.
- For Task 1, our system achieved a competitive sixth-place rank with an F2-score of **0.828**. Our approach combined a fine-tuned dense retriever with a multi-stage filtering and reranking pipeline, providing a robust foundation for the downstream QA task.
- We demonstrate that resource-efficient LLMs (<10B parameters), when coupled with sophisticated data augmentation and fine-tuning strategies, can achieve state-of-the-art performance in complex, specialized domains, offering valuable insights for low-resource legal NLP.

II. Related Works

A. Legal Document Retrieval

Effective legal document retrieval is a foundational step for any LQA system. Early approaches relied heavily on sparse retrieval methods like BM25 [5], which excel at keyword matching. However, recent advancements have gravitated towards Transformer-based [10] dense retrievers and cross-encoder rerankers, which better capture semantic nuances in legal texts [9].

A common pattern observed in recent legal NLP competitions is the use of hybrid or multi-stage systems. For instance, the winning solution of ALQAC 2023 employed an ensemble of BM25 and fine-tuned mBERT. Similarly, in COLIEE 2024, a top team utilized BM25 followed by reranking with multiple BERT-based models and monoT5 [25]. The NOWJ team, winners of ALQAC 2024, also integrated lexical models with semantic models (sBERT, ViMonoT5) and used a tree-based model (LightGBM) for ensembling [24]. Our work for Task

1 aligns with this trend, employing a multi-stage pipeline that includes fine-tuned dense retrievers and a cross-encoder reranker to progressively refine the set of candidate articles.

B. Legal Question Answering

The approaches to LQA have evolved significantly with the rise of LLMs, often in systems that couple a retriever with a generator, a paradigm known as retrieval-augmented generation (RAG) [11]. Previous ALQAC competitions saw successes with various methods. In ALQAC 2022, a winning strategy involved data augmentation and fine-tuning PhoBERT on the BoolQ dataset [23]. For ALQAC 2023, teams began leveraging LLMs with diverse prompting techniques [26], while others transformed multiple-choice questions into text-pair classification tasks for BERT-based models [7], [8]. The trend of using LLMs for complex reasoning was further solidified in the 2024 competition [6].

Our work for Task 2 builds directly upon the most recent advancements in LLM alignment and reasoning. While standard fine-tuning adapts a model to a specific domain, instruction fine-tuning has been shown to unlock more general instruction-following capabilities [12], and even more sophisticated techniques are required to elicit complex reasoning. Chain-of-Thought (CoT) prompting has been shown to significantly improve the reasoning abilities of LLMs by encouraging them to generate intermediate steps. Furthermore, recent alignment techniques like Direct Preference Optimization (DPO) and its variants, such as Group Relative Policy Optimization (GPRO) [15], have emerged as powerful methods to steer model behavior based on preference data, moving beyond simple supervised fine-tuning. Our winning approach integrates these cutting-edge techniques: we first enrich the training data with CoT explanations, then apply a two-phase training pipeline combining Supervised Fine-Tuning (SFT) with GPRO to create a highly specialized and accurate legal reasoning model.

III. Competition and Dataset

The ALQAC 2025 dataset is a collection of Vietnamese legal questions and their corresponding answers, designed to facilitate the development and evaluation of machine learning models for legal question answering (QA). The dataset is composed of two primary files: `alqac25_law.json`, which contains the legal corpus, and `alqac25_train.json`, which provides the training data for both tasks. An optional supplementary dataset, sponsored by Zalo, was also provided for Task 1.

Table I shows the statistics of the legal corpus provided for the competition. The dataset includes 28 distinct laws, comprising a total of 3,368 articles. A notable characteristic is the vast range in article length, from as few as 4 words to as many as 3,157 words, with an average length of approximately 240 words. This significant variance highlights the complexity and diverse nature of legal documents, posing a substantial challenge for retrieval models that must handle both concise and lengthy passages. The corpus vocabulary consists of

TABLE I
Statistics of the Legal Corpus (`alqac25_law.json`).

Statistic	Quantity
Number of Laws	28
Number of Articles	3,368
Min Article Length (words)	4
Max Article Length (words)	3,157
Average Article Length (words)	239.96
Vocabulary Size	2,487

2,487 unique words, reflecting the specialized and intricate terminology of the legal domain.

TABLE II
Statistics of the Training Questions (`alqac25_train.json`).

Statistic	Quantity
Total Questions	729
- True/False	387
- Multiple-choice	286
- Free-text	56
Min Question Length (words)	4
Max Question Length (words)	90
Average Question Length (words)	27.79

Table II summarizes the distribution of the 729 questions in the training set. The questions are divided into three types, with True/False questions being the most prevalent (387), followed by Multiple-choice (286) and Free-text (56) questions. This distribution requires participating systems to be versatile in handling different reasoning patterns and output formats. The question lengths range from 4 to 90 words, with an average of 27.79 words, indicating questions of moderate length and varied complexity.

The ALQAC 2025 competition imposed significant constraints to foster open research and create a level playing field. Participants were limited to open-weight models with fewer than 10 billion parameters, with a strict ban on closed-source models like GPT-4. Furthermore, using externally annotated legal datasets was disallowed. These rules directly guided our methodological choices toward optimizing smaller, accessible models for this specialized domain. These constraints significantly influenced our methodological choices, guiding us towards optimizing smaller, accessible models for a highly specialized domain.

IV. Proposed Method

A. Task 1: Legal Document Retrieval

We propose a three-stage legal document retrieval system tailored for the ALQAC-2025 challenge. Our method integrates LLM-based law tag filtering, sparse and dense retrieval, and a reranking mechanism. The system architecture is illustrated in Figure 2.

- 1) **LLM-Based Law Tag Filtering:** We leverage a large language model to predict relevant law article IDs from

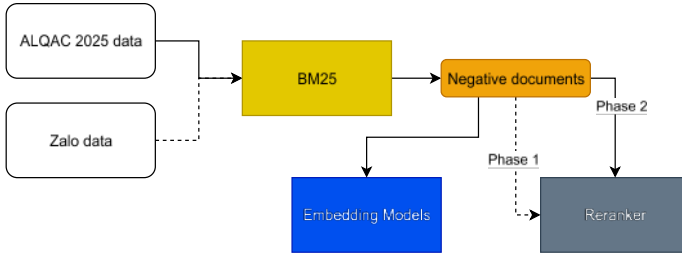


Fig. 1. Training process for 1

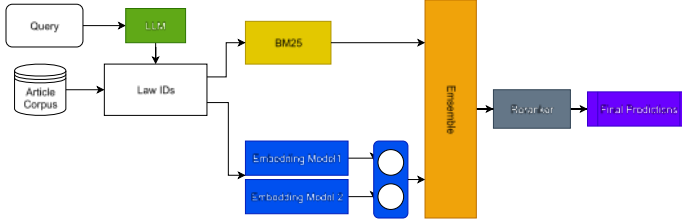


Fig. 2. Overview of our retrieval system for Task 1

the user query. This reduces the search space by restricting candidate documents to only those associated with the predicted tags.

- 2) **Hybrid Retrieval (Sparse + Dense):** For the filtered document set, we compute relevance scores using both sparse retrieval (BM25) and dense embeddings generated by two finetuned Vietnamese legal embedding models. The final score is a weighted combination of all three methods.
- 3) **Reranking:** The top-ranked candidates are passed to a reranker model, finetuned on ALQAC and Zalo datasets, to refine the ranking and improve precision. The final output is a list of highly relevant legal documents.

1) *Data Preprocessing and Pair Construction:* We begin by preprocessing all questions and legal documents to normalize formats, remove special characters, and tokenize Vietnamese text where appropriate. To train supervised models, we use BM25 to retrieve top-ranked candidate and construct positive and hard negative query-document pairs. Documents not labeled as relevant but ranked highly by BM25 are used as hard negatives, improving the discriminative capacity of dense models.

The full training data is split into 80% training and 20% validation sets for model development and tuning.

2) *Dense Retriever Fine-tuning:* We fine-tune two Vietnamese sentence embedding models to encode both queries and legal articles:

- AITeamVN/Vietnamese_Embedding_v2
- huyyandang/DEK21_hcmute_embedding

Each model is trained on the constructed triplet dataset using contrastive loss to maximize similarity between relevant query-document pairs and minimize similarity with hard negatives.

3) *Hybrid Retrieval Scoring:* At inference time, we encode both queries and legal documents using the two fine-tuned

embedding models. We also compute BM25 scores using the Pyserini toolkit. For each query-document pair, we compute:

- S_{bm25} : BM25 sparse retrieval score
- S_{emb_1} : cosine similarity from 2 sentence embedding models.

The scores are normalized and combined using a weighted sum:

$$S_{hybrid} = \alpha S_{bm25} + (1 - \alpha) * (0.5 * S_{emb_1} + 0.5 * S_{emb_2})$$

where α are weights determined via grid search on the validation set.

4) *Law ID Filtering with LLM:* To reduce noise and irrelevant articles, we use a large language model (LLM) to predict the likely law ID related to the input query. Candidate articles whose law IDs do not match the predicted ID are filtered out before reranking. This step significantly improves precision by enforcing domain-specific consistency.

5) *Threshold-based Filtering:* We apply a threshold θ_{retr} on the hybrid retrieval scores to further eliminate weak candidates. Only articles with $S_{hybrid} \geq \theta_{retr}$ are passed to the reranking stage.

6) *Cross-Encoder Reranking:* The final reranking step is performed using a fine-tuned cross-encoder model, AITeamVN/Vietnamese_Reranker. This model is trained in two stages: first on Zalo legal data, then on ALQAC-2025 training data. For each remaining candidate, the reranker outputs a relevance score $S_{rank}(q, d)$ based on the full text of the query and document.

We apply a second threshold θ_{rank} on the reranker outputs to determine the final list of retrieved articles:

$$\text{Final\_Docs} = \{d \mid S_{rank}(q, d) \geq \theta_{rank}\}$$

7) *Evaluation and Hyperparameter Tuning:* We evaluate the pipeline using Precision, Recall, and F2-score, following ALQAC-2025 official metrics. Both θ_{retr} and θ_{rank} are optimized using grid search on the validation set. All training and evaluation steps are conducted using open-source libraries, ensuring reproducibility.

B. Task 2: Legal Question Answering

Our approach for this task is a three-stage workflow centered on Chain-of-Thought (CoT) reasoning [13]. Chain-of-Thought reasoning has proven effective for complex question answering tasks [22]. The stages are: (1) augmenting the training data with CoT explanations, (2) a two-phase model training strategy involving Supervised Fine-Tuning (SFT) [14] and Group Relative Policy Optimization (GPRO) [15], and (3) ensembling multiple models for the final submission. Fig. 3 illustrates the detailed pipeline of our method.

1) *Data Preparation with CoT:* The original dataset lacked explicit reasoning steps. To address this, we employed a data enrichment strategy, a technique proven effective in low-resource legal domains [16]. Specifically, we generated a detailed reasoning process for each of the 729 training samples by prompting a powerful base model, DeepSeek-R1-0528-Qwen3-8B [17]–[19], to explain the logic connecting the legal context to the correct answer.

2) *Model Training*: We adopted a two-phase training process, detailed in Algorithm 1. Initially, we split the training data into 629 samples for training and 100 samples for validation to evaluate and select the best training configuration. After identifying the optimal config, we retrained the model on the entire 729 samples for the final submission.

Algorithm 1 Training Pipeline for Legal Question Answering

- 1: **Base Model**: DeepSeek-R1-0528-Qwen3-8B
- 2: **Phase 1 (SFT)**: Perform supervised fine-tuning on the 629 augmented training samples for 14 epochs. This yields the SFT model, denoted as **qwen_14e**.
- 3: **Phase 2 (GPRO)**: Use the **qwen_14e** model as a starting point for preference optimization. We train three separate GPRO models ($\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$) using three distinct reward function configurations.

3) *Ensemble Strategy for Final Submission*: To maximize performance and robustness, our final submission was generated using a majority-vote ensemble, a strategy also successfully employed in previous ALQAC competitions [20]. This strategy combines the outputs from all four of our trained models: the three distinct Group Relative Policy Optimization (GPRO) models ($\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$) and the primary Supervised Fine-Tuning (SFT) model (**qwen_14e**). For the free-text questions, the answer is chosen from ($\mathcal{G}_3$)

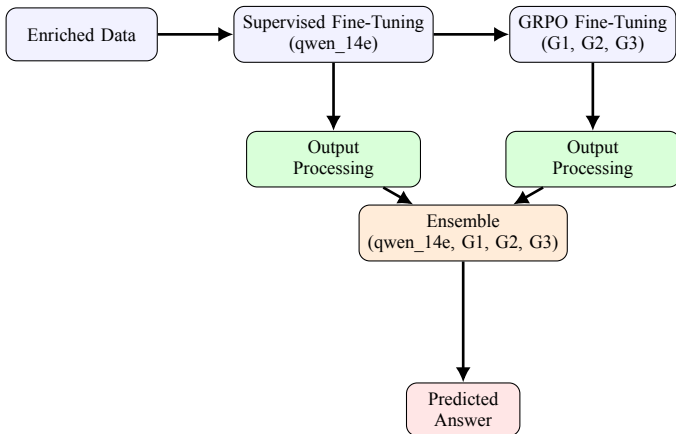


Fig. 3. Pipeline for Legal Question Answering System

During inference, we used the prompt structure to guide the model through a Chain-of-Thought reasoning process before it generated the final answer. Prompts are translated into English for illustration.

Prompt for True/False and Free-text questions

Based on the context below, please analyze carefully before answering the question.
 Question type: {question_type}
 Question: {question_text}
 Context: {article_content}
 Please generate the detailed reasoning part according to the template below inside the <think>...</think> tags.
 You need to write all the analysis, evidence, and reasoning steps fully before giving a concise answer.
 Do not answer immediately; you must reason fully first inside <think>.

1. Analysis of the question: [brief_analysis_of_question]
2. Evidence from context:
 - Please split each long paragraph in the context into several clear points.
 - Separate at least 3 to 5 points as evidence from the context.
 - Each point should clearly state the legal content, written briefly and understandably.
 - For example:
 - [point_1_from_legal_clause_A]
 - [point_2_from_legal_clause_A]
3. Step-by-step reasoning:
 - a) [reasoning_step_1_based_on_evidence]
 - b) [reasoning_step_2_next]
 - ...
4. Conclusion: [final_answer]

<think>
 [final_concise_answer_after_reasoning]

Prompt for Multiple-choice questions

Based on the context below, please analyze carefully before answering the question.
 Question type: {question_type}
 Question: {question_text}
 Context: {article_content}
 Four choices: A: {choice_A}; B: {choice_B}; C: {choice_C}; D: {choice_D}
 Please generate the detailed reasoning part according to the template below inside the <think>...</think> tags.
 You need to write all the analysis, evidence, and reasoning steps fully before giving a concise answer.
 Do not answer immediately; you must reason fully first inside <think>.

1. Analysis of the question: [brief_analysis_of_question]
2. Evidence from context:
 - Please split each long paragraph in the context into several clear points.
 - Separate at least 3 to 5 points as evidence from the context.
 - Each point should clearly state the legal content, written briefly and understandably.
 - For example:
 - [point_1_from_legal_clause_A]
 - [point_2_from_legal_clause_A]
3. Step-by-step reasoning:
 - a) [reasoning_step_1_based_on_evidence]
 - b) [reasoning_step_2_next]
 - ...
4. Conclusion: [final_answer]

<think>
 [final_concise_answer_after_reasoning]

Note: For the generation of training data, the prompt is prefaced with the following instruction to the model:

“Kết luận cuối cùng sau khi suy luận đã có. Vui lòng phân tích và làm rõ các bước lập luận để giải thích cho kết luận:

{answer}, đối với câu hỏi thuộc loại {question_type}.”
(Translation: “The final conclusion after reasoning is already determined. Please analyze and elucidate the reasoning steps that justify the conclusion: {answer}, for a question of type {question_type}.”)

V. Experiments and Results

A. Evaluation Metrics

1) *Task 1*: The metrics used are Precision, Recall, and F2, defined as follows:

$$\text{Precision}_i = \frac{\text{the number of correctly answered } i\text{th}}{\text{the number of question } i\text{th}} \quad (1)$$

$$\text{Recall}_i = \frac{\text{the number of correctly answered } i\text{th}}{\text{the number of question } i\text{th}} \quad (2)$$

$$F2_i = \frac{5 \times \text{Precision}_i \times \text{Recall}_i}{4 \times \text{Precision}_i + \text{Recall}_i} \quad (3)$$

The macro-average F2 is defined as:

$$F2 = \text{average}(F2_i) \quad (4)$$

In addition to the evaluation measures mentioned above, ordinal information retrieval metrics such as Mean Average Precision (MAP) and R-Precision can be employed to analyze the characteristics of the submission results. The macro-average F2 is the principal measure for Task 1.

2) *Task 2*: The principal evaluation measure is Accuracy, defined as follows:

$$\text{Accuracy} = \frac{\text{questions that were correctly answered}}{\text{the number of questions}} \quad (5)$$

For the free-text questions, the answer is free-text and is evaluated by human experts.

B. Experimental Configuration

1) *Task 1*: For the legal document retrieval task, we fine-tuned the system components using a multi-stage approach combining sparse and dense retrieval methods. All training processes were executed on a single NVIDIA T4 GPU with 16GB of VRAM, except for the reranker, which was trained using 2 GPUs. The overall training pipeline consisted of the following stages with their respective settings:

- **Stage 1 (Dense Encoder Fine-tuning)**: We fine-tuned two sentence transformer models using hard negative pairs generated by a BM25 retriever. Each model was trained for 3 epochs with a batch size of 16 and a learning rate of 2×10^{-4} . The input sequence length was set to 8192 tokens to accommodate long legal documents. The fine-tuning process took approximately 1 hour per model.
- **Stage 2 (Reranker Fine-tuning)**: A reranking model based on the FlagEmbedding architecture was further fine-tuned using top-k pairs from the BM25 retrieval stage. Training was conducted in two phases: first on additional Zalo legal data (4 epochs, 2 hours), followed by fine-tuning on the main ALQAC-2025 dataset (3

epochs, 1 hours), using a learning rate of 6×10^{-5} and a per-device batch size of 4.

This staged training regime allowed us to build a robust hybrid retrieval system, leveraging both traditional lexical matching and semantic embeddings, tailored specifically for Vietnamese legal texts.

2) *Task 2*: For the legal question answering task, we employed instruction fine-tuning on the LLM using the LoRA methodology [21]. The entire training process was conducted on a single NVIDIA T4 GPU with 16GB of VRAM. It consisted of two phases with identical LoRA settings (rank = 8, alpha = 16, and dropout = 0):

- **Phase 1 (SFT)**: Supervised fine-tuning was conducted for 14 epochs, taking approximately 12 hours to complete.
- **Phase 2 (GRPO)**: Group Relative Policy Optimization was performed for each of our three distinct model configurations. Each run consisted of 100 training steps and required approximately 7 hours, resulting in a total of 21 hours for this phase.

The fine-tuning utilized a batch size of 1, a maximum sequence length of 8192 tokens, and a learning rate of 2×10^{-4} .

C. Results

1) *Task 1*: Table III presents the official leaderboard for the Legal Document Retrieval task (Task 1) on the private test set. Our system, Hallucineers, achieved an F2-score of 0.828, securing 6th place among a competitive pool of high-performing teams. Leveraging a fine-tuned dense retriever combined with a hybrid ranking strategy, our approach demonstrated consistent robustness and competitiveness despite challenges posed by the specialized legal domain and resource constraints. Detailed rankings of the leading teams are provided in Table III.

These results validate the effectiveness of our retrieval pipeline in delivering reliable performance even under the strict limitations of open-weight models and computational resources.

TABLE III
Task 1: Top Teams on the Official Leaderboard on Private Test Set

Team	F2-score	Rank
Flame Reavers	0.9435	1
vund	0.8970	2
UIT legalGrep	0.8970	2
NOWJ	0.8767	4
UTEHY-NLU	0.8482	5
Hallucineers	0.8279	6
D2	0.8238	7
UIT-Darkcow	0.7873	8
ABCD	0.7730	9
Warriors	0.6897	10

2) *Task 2*: Table IV presents the official leaderboard for the Legal Question Answering task (Task 2) on the private test set. Our team, Hallucineers, secured the top position with an accuracy of 98.78%, tying with NOWJ and UIT-Darkcow. This outstanding result underscores the robustness

and effectiveness of our Chain-of-Thought-based instruction fine-tuning combined with an ensemble strategy, as detailed in Section IV.

TABLE IV
Task 2: Top Teams on the Official Leaderboard on Private Test Set

Team	Accuracy (%)	Rank
NOWJ	98.78	1
Hallucineers	98.78	1
UIT-Darkcow	98.78	1
UTEHY-NLU	97.56	4
FlameReavers	96.34	5
vund	96.34	5
BAMBOOVN	95.12	7
ABCD	93.90	8
D2	93.90	8
VnLegalLLM25	92.04	10

Our system’s top accuracy of 98.78% attests to the effectiveness of integrating Chain-of-Thought prompting, fine-tuned open-weight LLMs, and model ensembling. These strengths, combined with careful prompt engineering and strategic model selection, enable highly competitive performance under stringent resource constraints and challenging conditions.

VI. Conclusion

This paper presented an efficient and effective system for Vietnamese legal question answering in ALQAC 2025, combining a fine-tuned dense retriever with a compact large language model enhanced by two-stage instruction fine-tuning and Chain-of-Thought prompt engineering. Despite strict model size constraints, our approach achieved state-of-the-art results, securing first place in question answering and strong performance in document retrieval. These findings demonstrate the potential of carefully designed fine-tuning and prompt techniques to optimize smaller open-weight models for complex legal NLP tasks.

Acknowledgment

This research is supported by research funding from the Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City

References

- [1] N. H. Thanh *et al.*, “A Summary of the ALQAC 2021 Competition,” *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, Bangkok, Thailand, 2021, pp. 1–5, doi: 10.1109/KSE53942.2021.9648724.
- [2] C. Nguyen *et al.*, “ALQAC 2022: A Summary of the Competition,” *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, Nha Trang, Vietnam, 2022, pp. 1–5, doi: 10.1109/KSE56063.2022.9953764.
- [3] C. Nguyen *et al.*, “A Summary of the ALQAC 2023 Competition,” *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, Hanoi, Vietnam, 2023, pp. 1–6, doi: 10.1109/KSE59128.2023.10299527.
- [4] D.-T. Do *et al.*, “A Summary of the ALQAC 2024 Competition,” *2024 16th International Conference on Knowledge and Systems Engineering (KSE)*, Kuala Lumpur, Malaysia, 2024, pp. 422–427, doi: 10.1109/KSE63888.2024.11063484.
- [5] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework,” *Foundations and Trends in Information Retrieval*, 2009, doi: 10.1561/15000000019.
- [6] H. Q. Pham, Q. V. Nguyen, D. Q. Tran, T. N. K. Bao, and K. V. Nguyen, “Top 2 at ALQAC 2024: Large Language Models (LLMs) for Legal Question Answering,” *International Journal of Asian Language Processing*, 2025.
- [7] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- [9] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The muppets straight out of law school,” in *Findings of the Association for Computational Linguistics: EMNLP*, 2020.
- [10] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [11] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [12] J. Wei *et al.*, “Finetuned language models are zero-shot learners,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [13] J. Wei, X. Wang, D. Schuurmans, *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” *arXiv preprint arXiv:2201.11903*, 2022.
- [14] J. Wei, X. Wang, A. Schuurmans, *et al.*, “Finetuned language models are zero-shot learners,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [15] T. Xie, Z. Jiang, X. Wang, *et al.*, “Group Regularized Policy Optimization (GRPO) for Language Model Alignment,” *arXiv preprint arXiv:2402.03623*, 2024.
- [16] H.-L. Nguyen, *et al.*, “Neco@alqac 2023: Legal domain knowledge acquisition for low-resource languages through data enrichment,” *arXiv preprint arXiv:2309.13540*, 2023.
- [17] DeepSeek Team, “DeepSeek LLM: Vision, Architecture, and Training,” *arXiv preprint arXiv:2405.13237*, 2024.
- [18] Qwen Team, “Qwen: A Scalable and Advanced Large Language Model,” *arXiv preprint arXiv:2309.16609*, 2023.
- [19] C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister, “Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes,” *Findings of the Association for Computational Linguistics (ACL)*, 2023, *arXiv preprint arXiv:2305.02301*, 2023.
- [20] H. N. Trung and S. N. Truong, “Ensemble learning methods for legal processing tasks in alqac 2022,” in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2022, pp. 1–5.
- [21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Xia, and D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” *arXiv preprint arXiv:2201.11903*, 2022.
- [23] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, “Boolq: Exploring the surprising difficulty of natural yes/no questions,” *arXiv preprint arXiv:1905.10044*, 2019.
- [24] T. M. Nguyen, N. D. Nguyen, V. H. Tran, and H. T. T. Tran, “NOWJ at ALQAC 2024: A Hybrid Approach for Legal Question Answering,” in *2024 16th International Conference on Knowledge and Systems Engineering (KSE)*, Kuala Lumpur, Malaysia, 2024, pp. 440–445, doi: 10.1109/KSE63888.2024.11063594.
- [25] T. M. Nguyen, N. D. Nguyen, H. M. Tran, V. H. Tran, N. H. Doan, and H. T. T. Tran, “NOWJ@COLIEE 2024: Leveraging Advanced Deep Learning Techniques for Efficient and Effective Legal Information Processing,” in *JSAI International Symposia on AI*, 2024.
- [26] L. Hoang, T. Bui, C. Nguyen, and L.-M. Nguyen, “AIEPU at ALQAC 2023: Deep Learning Methods for Legal Information Retrieval and Question Answering,” in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, Hanoi, Vietnam, 2023, pp. 1–6.

Combining Large Language Model and Supportive Models for Legal Question Answering

1st Long Vu Van*

University of Engineering and Technology
Vietnam National University
Ha Noi, Viet Nam
23025015@vnu.edu.vn

2nd Hai Le Hong*

University of Engineering and Technology
Vietnam National University
Ha Noi, Viet Nam
hailh@vnu.edu.vn

Abstract—This paper details the methods proposed by the BAMBOOVN team during their participation in the 5th Automated Legal Question Answering Competition (ALQAC 2025). During the competition we have to resolve two main tasks: Legal Document Retrieval and Legal Question Answering. For the Legal Document Retrieval task we applied an ensemble method: semantic searching with embedding outputs from BGE-M3 transformer model and traditional ranking methods using word information: BM25, TF-IDF supported by a topic classification model. After pre-ranking step we used a relevant classification model to grade the retrieved document as relevant or not relevant. These supportive models (topic/relevant classification) were trained by combining both original data and LLM-augmented data for better accuracy. The final score is combined using tuned hyperparameters and will be compared with an ensemble threshold to get filtered. For the Legal Question Answering task, we have three types of questions: true/false questions, multiple-choices question and free-text question. Suppose that the retrieved document from previous stage is correct, we address this problem by using Large Language Model (LLM) with prompt engineering technique to attain the generalization of LLM on the question answering task. While our approach did not attain the highest scores in the competition, it proved that using LLM-synthesized data within an adaptive pipeline (integrating LLMs with auxiliary models) effectively mitigates data scarcity limitations.

Index Terms—Large Language Model, Legal Document Retrieval, Legal Question Answering, Retrieval-Augmented Generation

I. INTRODUCTION

The exponential growth of legal documentation and the increasing complexity of legal frameworks have created a pressing need for efficient and accurate Legal Question Answering (LQA) systems. Such systems promise to democratize access to justice, enhance legal research efficiency, and support practitioners in navigating vast repositories of statutes, case law, and regulations. However, automating LQA poses significant challenges, including the precise retrieval of relevant documents from large corpora and the nuanced understanding required to generate correct answers to diverse legal queries (e.g., fact verification, multiple-choice, and open-ended reasoning).

Traditional approaches to legal information retrieval (IR), such as keyword-based methods (e.g., BM25, TF-IDF), offer efficiency but often struggle with semantic matching and contextual relevance. Conversely, while modern neural retriev-

ers and Large Language Models (LLMs) excel at semantic understanding, they can be computationally intensive and may lack the precision needed for highly specialized legal domains without careful tuning. Similarly, in the Question Answering (QA) phase, LLMs demonstrate remarkable generative capabilities but are prone to hallucination and inconsistency when domain-specific knowledge is inadequately grounded.

This paper presents the methodology developed by the BAMBOOVN team for the 5th Automated Legal Question Answering Competition (ALQAC 2025), addressing these dual challenges through a hybrid pipeline. For **Legal Document Retrieval**, we propose an ensemble strategy combining:

- Semantic Search using embeddings from the state-of-the-art transformer model BGE-M3 [2].
- Traditional information retrieval methods (BM25, TF-IDF) for lexical matching.
- A Relevance Classification Model, trained on both original and LLM-augmented data, to refine the ensemble's output.

For **Legal Question Answering**, we leverage prompt-engineered LLMs to generate answers across three question types (true/false, multiple-choice, free-text), conditional on high-confidence document retrieval. Our core contribution lies in synergistically integrating robust retrieval (blending semantic, lexical, and learned relevance signals) with constrained, context-aware LLM inference to enhance accuracy and reliability.

The remainder of this paper is structured as follows: Section II reviews related work in legal information retrieval and question answering. Sections III details our methodology for document retrieval and question answering, respectively. Section IV describes experimental setups and results from ALQAC 2025. Section V discusses limitations and conclusions with future directions.

II. RELATED WORK

Automatic Legal Document Retrieval (ALDR) and Legal Question Answering (LQA) represent critical subfields within legal AI, drawing significant research attention due to their practical impact on judicial efficiency and accessibility. This section surveys key advancements and methodologies relevant to our approach.

Traditional IR techniques have long been applied to Vietnam legal corpora. Keyword-based methods like TF-IDF and BM25 remain popular for their simplicity, efficiency, and interpretability in early systems like UIT-DarkCow [5], [10], [15]. NOWJ-XOSA [9] proposed a solution that combines lexical strategy with semantic searching using strong base model to achieve top 2 on retrieval task. NOWJ [6] in ALQAC 2024 has proposed an ensemble algorithm with combination of lexical matching and state-of-the-art models with data engineering steps and a tree-based model. Not participated in ALQAC Long et al [8] proposed a pipeline using lexical matching (BM25), semantic search (mBERT) and LLM-supported score. Hybrid solution has been widely applied in COLIEE 2025 (International Competition on Legal Information Extraction and Entailment) such as [11], [12] and [13]. Se7ense [14] in ALQAC 2024 combined BM25 scores with attention scores between the corpus and query, ranking based on the combined scores. To overcome the problem of lacking training data, Nguyen et al in ALQAC 2023 [4] proposed a solution that used a crawled dataset from the internet to enrich their datasets. For legal retrieval, they used an ensemble technique that combines lexical and semantic matching.

Current approaches to legal question answering bifurcate into two principal methodologies: deep learning classification systems and generative answer architectures. For span-based questions requiring exact text extraction, contemporary systems implement answer extraction techniques that directly map queries to statutory segments, as demonstrated in Nguyen et al.'s Neco@ALQAC2023 [4] framework which addresses textual grounding through contextual alignment. To overcome the pervasive challenge of lengthy legal documents exceeding conventional model context windows, chunking techniques dissect articles into semantically coherent units, with retrieval subsequently performed through lexical matching methods like BM25—an approach validated across multiple ALQAC entries. For closed-domain question types (true/false, multiple-choice), methodological divergence emerges: Nguyen et al [4] employ text-pair classification to independently score question-choice relevance, while NOWJ (ALQAC 2023) [7] implements unified BERT-based classification treating true/false as binary decisions and multiple-choice as 4-class predictions. With the raising of LLM UIT-Darkcow [5] (top 2 in ALQAC 2024), se7ense at ALQAC 2024 [14] applied prompting techniques and finetuning LLM to achieve prominent result. AIEPU [15] proposed using LLM prompting technique with voting from top-k prompt results to achieve top 1 accuracy for task 2 in ALQAC 2023.

While hybrid retrieval and LLM-based QA are established, key gaps persist:

- Most ALDR ensembles use simple score averaging based on ranking algorithm and filter threshold, lacking adaptive relevance filtering and reflexion judgement from third model (LLM/classification).
- Most solutions lack preprocessing techniques such as topic classification, rewriting question to enhance the downstream task.

- LLM-augmented data is rarely leveraged to train lightweight relevance classifiers for retrieval refinement.

Our methodology addresses these by:

- Proposing a tuned ensemble with a trainable relevance classifier (using LLM-augmented data),
- Designing an adaptive RAG pipeline for diverse legal question types while conditioning on high-precision retrieval.

III. PROPOSED METHOD

Our integrated framework for Legal Question Answering (LQA) employs a multi-stage pipeline designed to address the dual challenges of precise document retrieval and contextually grounded answer generation. The system initiates with domain-aware preprocessing, progresses through a sophisticated retrieval ensemble with learned relevance filtering, and culminates in task-specialized LLM inference. This holistic approach ensures both computational efficiency and competitive accuracy across diverse legal question types. The overall architecture is illustrated in 1.

A. Topic classification

During our research, we found out that we can preprocess the initial question to get further metadata such as the topic of the question, important keywords, etc for enhancing the retrieval pipeline's performance. Topic classification serves as the critical first stage, where incoming legal questions are categorized into specialized narrow domains (28 domains) using a fine-tuned Legal-BERT classifier. This domain identification enables two strategic advantages: (1) dynamic corpus filtering that restricts downstream document search to topic-relevant subsets of the legal database, significantly reducing computational overhead and noise, and (2) contextual priming for subsequent LLM prompts.

B. Relevant classification

The relevant classification model constitutes our novel refinement mechanism for retrieved documents. Following initial retrieval, a binary classifier [1] processes concatenated question-document pairs to predict relevance probabilities. This model serves as a filter to eliminate documents scoring below the optimized threshold, thereby ensuring that only high-confidence passages proceed to subsequent answer generation stages while significantly reducing computational overhead in the LLM processing pipeline.

C. LLM rewrite question for full text search

To address the critical limitation of traditional lexical search methods (BM25/TF-IDF) where retrieval performance is frequently compromised by unimportant keywords in natural language queries, we implement an LLM-based query refinement stage. Legal questions often contain extraneous words—common verbs, filler phrases, or contextually redundant terms—that dilute search precision when using statistical text matching algorithms. These methods inherently struggle

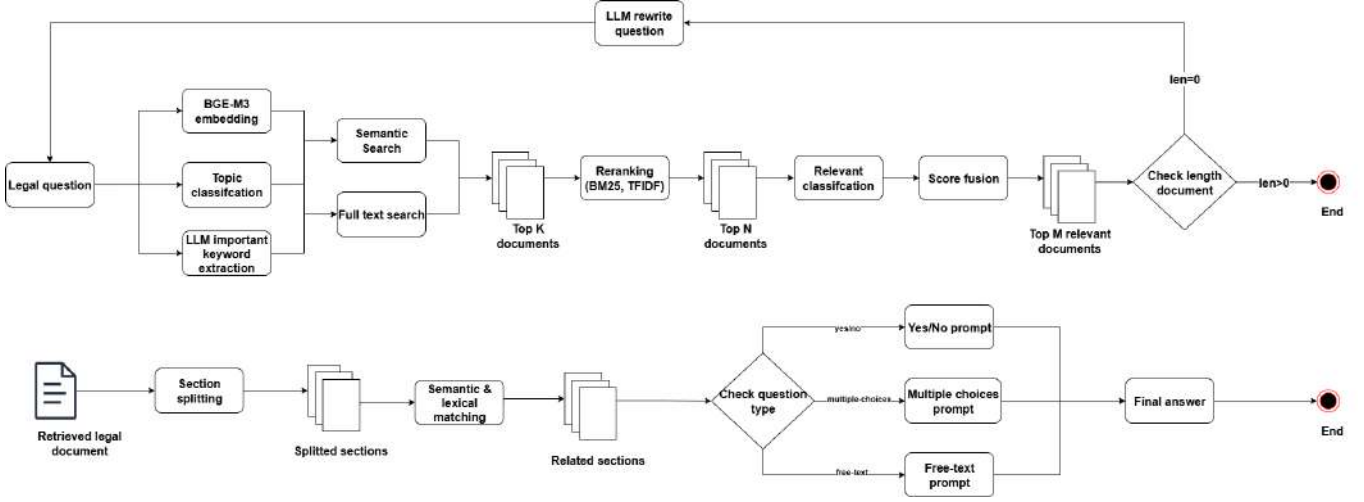


Figure 1. Overall ensemble pipeline: the upper part is the retrieval pipeline, the lower part is the question answering pipeline

Table I
LLM REWRITE QUESTION PROMPT

Type	Value
Original question	Quan hệ hôn nhân và gia đình có yếu tố nước ngoài là quan hệ hôn nhân và gia đình mà ít nhất một bên tham gia là người nước ngoài, người Việt Nam định cư ở nước ngoài, đúng hay sai?
System prompt	Bạn là một người viết lại câu hỏi, chuyển đổi câu hỏi đầu vào thành phiên bản tốt hơn được tối ưu hóa cho việc truy xuất vectorstore. Hãy xem xét đầu vào và cố gắng phân tích ý định ngữ nghĩa/ý nghĩa cơ bản.
User prompt	Câu hỏi ban đầu: {question} Chỉ trả về câu hỏi được viết lại, không giải thích gì thêm!
Rewritten question	Quan hệ hôn nhân và gia đình có yếu tố nước ngoài là gì?

to distinguish between semantically meaningful legal terminology and incidental linguistic elements, leading to irrelevant document recalls that degrade downstream performance.

D. Ensembling retrieval

Ensembling Retrieval synergizes multiple retrieval paradigms through a weighted fusion strategy. In the first stage, we execute parallel document searches: semantic retrieval via BGE-M3 embeddings [2] for semantic search, lexical matching through full text search using the important keywords extracted from step III-C. Each method returns K candidate documents. These documents are then re-ranked through an ensemble method including BM25 (optimizing term weighting), and TF-IDF (ensuring broad keyword coverage) these two methods score then fused by average. After the preranking, N highest scored documents are chosen to be scored with BERT model. The final ensemble formula score is described in formula III-D.

$$score = \alpha \times (\beta \times BM25 + \gamma \times TFIDF) + (1 - \alpha) \times BERT$$

Hyperparameters $K = 200$, $N = 10$, $\alpha = 0.5$, $\gamma = 0.5$, $\beta = 0.5$ were optimized via grid search to maximize final F_2 score. Documents falling below the empirically determined $S_{final} < 0.8$ threshold are excluded, ensuring only high-confidence passages proceed to answer generation.

E. LLM question answering

Given a related article, we know that each question only relates to a small piece of the article, our approach strategically addresses the granular nature of legal information requirements, where individual yes/no or essay questions typically reference specific segments within lengthy statutory articles, while multiple-choice questions may implicate distinct provisions across different terms of the related articles. To optimize context precision, we implement a hierarchical document decomposition protocol aligned with Vietnamese legal structure (Article $\rightarrow$ Section $\rightarrow$ Clause $\rightarrow$ Point), segmenting comprehensive articles into discrete semantic units using rule-based parsing with regular expressions. This structural normalization mitigates context window limitations while preserving juridical coherence.

The core QA framework employs Qwen2.5-7B [3], leveraging its enhanced reasoning capabilities through specialized prompt architectures (Table III). Retrieval operates at the subsection level, where:

- Semantic search (BGE-M3) and lexical search (BM25) independently identify the single most relevant legal section (top-1). If semantic search and lexical search return different sections, we put both terms to the LLM context. This technique is similar to [7] but using article's section level instead of sentence level to keep the intra-

Table II
LLM REWRITE QUESTION FOR FULL TEXT SEARCH PROMPT SETTING

Type	Value
Original question	Người đã hoàn thành cai nghiện ma túy tự nguyện đã chấp hành xong quyết định đưa vào cơ sở cai nghiện không chịu sự theo dõi hành vi sử dụng trái phép chất ma túy trong thời hạn 12 tháng kể từ ngày hoàn thành việc cai nghiện, đúng hay sai?
System prompt	Bạn là một chuyên gia về ngữ pháp tiếng Việt hoạt động trong lĩnh vực luật pháp.
User prompt	Câu hỏi của người dùng: {question}
Related article	Luật Phòng, chống ma túy article_id = 40 ... Người đã chấp hành xong quyết định áp dụng biện pháp xử lý hành chính đưa vào cơ sở cai nghiện bắt buộc chịu sự quản lý sau cai nghiện ma túy trong thời hạn 02 năm kể từ ngày chấp hành xong quyết định. ...
Full text search with original question	[]
Full text search with rewritten question	[Document(metadata="law_id": "Luật Phòng, chống ma túy", "article_id": "40")]

information of each legal section and most relevant points from both lexical and semantic queries.

- For multiple choices question, with each choice we form a query text using template $\{question\}\{choices\}$ to feed to semantic and lexical algorithm for retrieving the most related section to that choice.

IV. EXPERIMENT AND RESULTS

A. Data analysis

This year’s competition dataset consists of two corpora: (1) an official task-specific collection of 82 legal questions and 3,368 legal articles spanning 28 juridical domains, and (2) a supplementary Zalo document retrieval dataset containing 3,196 queries and 61,425 legal articles.

Text length analysis (Figure 2) reveals significant lexical variation, with article lengths following a heavy-tailed distribution (range: 4–3,169 words) that necessitates specialized preprocessing. The mean article length (239.5 words) aligns with standard BERT-family model input constraints.

Question-answering queries show similar variation with a mean length of 20.66 words (min: 4, max: 90) and left-skewed dispersion, indicating frequent verbose inquiries alongside compact professional formulations.

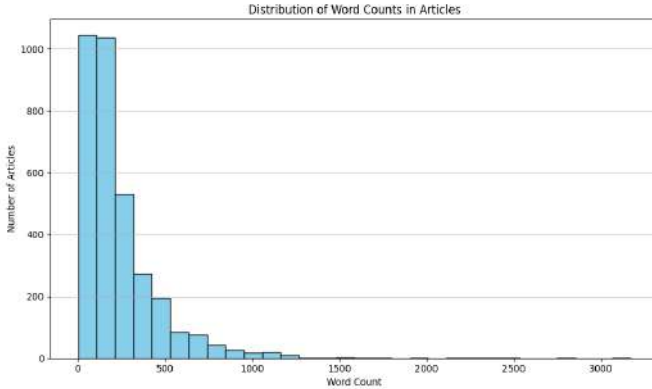


Figure 2. Legal article word count distribution

B. Finetuning classification models

For the topic classification task—which identifies the specific legal domain of a given question to optimize downstream retrieval. We curated a specialized dataset comprising:

- Original samples: 729 legally annotated questions from the ALQAC 2025 training set.
- Augmented samples from original questions: 729 synthetically generated questions from 729 original question.
- Augmented samples from original legal law: for each article of 28 topics, we used LLM prompt to generate up to 5 questions. We removed the pruned questions (empty, English/Chinese questions) and got 31,965 synthetic questions.

The combined dataset (33,423 samples) spans 28 distinct legal domains, indexed as Law IDs 0–27. Each question was mapped to its corresponding Law ID. We partitioned this data using a 90:10 train-test split (30,120 training samples, 3,303 test samples) while maintaining strict domain balance across splits.

For model development, we fine-tuned a DistilBERT transformer multilingual backbone [1], the best finetuned checkpoint achieves **92.67% accuracy** on the topic classification task. For the relevance classification, we defined a retrieved document is relevant (Class 1) or not relevant (Class 0) to a given legal question if the document can be used to answer the given question. We structured input data using a standardized template:

$\{document\} \{question\}$

This Vietnamese-language format preserves the original semantic context while providing a consistent structure for model training. The dataset construction followed three principles:

- For each original question we augmented another question using LLM, so we have 729 synthetic questions generated from 729 original questions.
- For each positive sample (relevant) we make 2 negative samples (not relevant) using legal articles from the same topic randomly.

Finally, we got a dataset of 19,654 sample with 1,997 samples in test set and 17,567 in training set using a 90:10 train-test split. The best-performing model achieved **98.8% accuracy** on

Table III
LLM PROMPT FOR QUESTION ANSWERING

Question type	Prompt type	Value
Yes/No	System prompt	Bạn là một trợ lý ảo hỗ trợ hỏi đáp pháp luật Việt Nam, bạn sẽ trả lời các câu hỏi dạng trắc nghiệm Đúng/Sai. Hãy cho điểm nhị phân 'Đúng' hoặc 'Sai', 'Đúng' có nghĩa là câu hỏi đúng dựa trên các thông tin được cung cấp.
	User prompt	Thông tin tham khảo: {context} Câu hỏi: {question} Chỉ trả về một trong hai giá trị 'Đúng' hoặc 'Sai', không giải thích gì thêm!
Multiple-choices	System prompt	Bạn là một trợ lý ảo hỗ trợ hỏi đáp pháp luật Việt Nam, bạn sẽ trả lời các câu hỏi dạng trắc nghiệm A/B/C/D. Sử dụng các đoạn thông tin đã truy xuất được để trả lời câu hỏi.
	User prompt	Thông tin tham khảo: {context} Câu hỏi: {question} {choices} Chỉ trả về một trong các giá trị A/B/C/D, không giải thích gì thêm!
Free-text	System prompt	Bạn là một trợ lý ảo hỗ trợ hỏi đáp pháp luật Việt Nam. Sử dụng các đoạn thông tin đã truy xuất được để trả lời câu hỏi. Hãy tạo ra câu trả lời ngắn gọn và súc tích gồm đầy đủ thông tin được cung cấp!
	User prompt	Thông tin tham khảo: {context} Câu hỏi: {question} Câu trả lời của bạn:

the test data. Consistent hyperparameters were applied across both fine-tuned models (DistilBERT [1]).

Embedding generation utilized the BGE-M3 multilingual model through Hugging Face Transformers, with all vector operations accelerated via GPU inference. LLM serving was handled by Ollama (v0.3.2) running Qwen2.5:7b [3] for legal question answering tasks.

This infrastructure operated on a dedicated hardware configuration featuring:

- GPU: NVIDIA GeForce RTX 3060 12GB
- CPU: Intel® Core™ i7-13700K 16 cores/24 threads 5.4GHz
- Memory: 32GB DDR5 RAM

C. Results

For rigorous evaluation across competition tasks, we implemented structured submission protocols: In Task 1 (Legal Document Retrieval), three distinct runs were generated corresponding to top-M document returns ($M = 1, 3, 5$) per query to progressively evaluate precision-recall tradeoffs. For Task 2 (Legal Question Answering), we submitted complementary runs comprising: (1) exclusive BM25-based matching, (2) hybrid BM25-semantic fusion, and (3) full-article context integration – systematically isolating the impact of retrieval granularity on answer accuracy. This tiered experimental design enabled comprehensive ablation analysis while adhering to ALQAC evaluation frameworks.

Our overall performance reveals a notable dichotomy between the effectiveness of our document retrieval system and our question-answering capability.

In the Legal Retrieval task (Table IV), our team ranked 9th out of 13 with an F2 score of **0.714**. This placement in the lower third of the ranking, with a significant performance

Table IV
LEGAL RETRIEVAL RESULT

Team	F2
Flame Reavers	0,9434765774
vund	0,89701897
UIT legalGrep	0,89701897
NOWJ	0,8766937667
UTEHY-NLU	0,8482384822
Hallucineers	0,827913279
D2	0,8238482383
UIT-Darkcow	0,7872628725
ABCD	0,7730352302
BAMBOOVN	0,7140921409
Warriors	0,6897018969
LAD	0,6588533415
VnLegalLLM25	0,5189701896

Table V
LEGAL QUESTION ANSWERING RESULT

Team	Correct	Accuracy
NOWJ	81	0,98780487800
Hallucineers	81	0,98780487800
UIT-Darkcow	81	0,98780487800
UTEHY-NLU	80	0,97560975600
FlameReavers	79	0,96341463400
vund	79	0,96341463400
BAMBOOVN	78	0,95121951200
ABCD	77	0,93902439000
D2	77	0,93902439000
VnLegalLLM25	74	0,90243902400
Warriors	67	0,81707317100
UIT legalGrep	62	0,75609756100
LAD	25	0,30487804900

gap to the top team, indicates that our retrieval pipeline was less effective at identifying the most relevant legal articles compared to the leading approaches.

Conversely, in the Legal Question Answering task (Table V), we demonstrated a much stronger performance. We achieved an accuracy of **0.951**, correctly answering 78 out of 82 questions, which earned us 7th place. Our accuracy is highly competitive, trailing the top-performing teams by a very narrow margin.

This disparity highlights a key strength in our approach: our system excels at synthesizing answers from a given set of documents, even when that set is not optimally retrieved. Our lower retrieval score implies we may have worked with a noisier or less complete set of candidate articles. However, our high answering accuracy proves that our answer generation pipeline—which leverages powerful Language Models (LLMs) and reasoning techniques—was robust enough to identify the correct solution from suboptimal retrieval results. This shows that our advanced answering component can partially compensate for deficiencies in the retrieval stage.

V. LIMITATIONS & FUTURE WORKS

Despite promising results, this work faces several limitations. The system struggles with long legal documents, and its dependence on prompting techniques with smaller LLMs (<10B parameters) increases instability and hallucination risks. Our rudimentary reranking mechanism lacks the sophistication of modern neural rankers like BGE-M3, and hardware constraints prevented full exploitation of the ZALO dataset. Future efforts will prioritize hierarchical text decomposition, advanced reasoning methods like chain-of-thought, and hybrid ranking integrating neural approaches.

CONCLUSION

This study empirically validates the efficacy of Large Language Models (LLMs) in legal-domain Retrieval-Augmented Generation (RAG) pipelines, demonstrating their dual capacity for synthetic data augmentation and operational support within knowledge-intensive workflows. Through synergistic integration of LLMs with transformer-based architectures, our resource-constrained implementation—deployed on limited hardware—achieved competitive performance in the ALQAC 2025 benchmark (top 7 in task 2). The results establish that strategically deployed LLMs can overcome computational limitations while maintaining accuracy in specialized legal contexts, offering a viable pathway for efficient legal AI systems where infrastructure scalability remains constrained.

REFERENCES

- [1] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, 2019, arXiv preprint arXiv:1910.01108
- [2] Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, Chen, Jianlv and Xiao, Shitao and Zhang, Peitian and Luo, Kun and Lian, Defu and Liu, Zheng, arXiv preprint, 2024.
- [3] Qwen and An Yang and Baosong Yang and Beichen Zhang and Binyuan Hui and Bo Zheng and Bowen Yu and Chengyuan Li and Dayiheng Liu and Fei Huang and Haoran Wei and Huan Lin and Jian Yang and Jianhong Tu and Jianwei Zhang and Jianxin Yang and Jiaxi Yang and Jingren Zhou and Junyang Lin and Kai Dang and Keming Lu and Keqin Bao and Kexin Yang and Le Yu and Mei Li and Mingfeng Xue and Pei Zhang and Qin Zhu and Rui Men and Runji Lin and Tianhao Li and Tianyi Tang and Tingyu Xia and Xingzhang Ren and Xuancheng Ren and Yang Fan and Yang Su and Yichang Zhang and Yu Wan and Yuqiong Liu and Zeyu Cui and Zhenru Zhang and Zihan Qiu, “Qwen2.5 Technical Report”, 2025, <https://arxiv.org/abs/2412.15115>
- [4] Hai-Long Nguyen and Dieu-Quynh Nguyen and Hoang-Trung Nguyen and Thu-Trang Pham and Huu-Dong Nguyen and Thach-Anh Nguyen and Thi-Hai-Yen Vuong and Ha-Thanh Nguyen, “NeCo@ALQAC 2023: Legal Domain Knowledge Acquisition for Low-Resource Languages through Data Enrichment”, 2023, <https://arxiv.org/abs/2309.05500>.
- [5] Top 2 at ALQAC 2024: Large Language Models (LLMs) for Legal Question Answering, Pham, Huy Quang and Nguyen, Quan Van and Tran, Dan Quang and Nguyen, Thang Kien-Bao and Nguyen, Kiet Van, International Journal of Asian Language Processing, volume 35, number 01, pages 2450010, 2025, World Scientific.
- [6] NOWJ@ALQAC 2024: Leveraging Document Ranking Methods and Large Language Models for Legal Document Retrieval and Question Answering, Mai, Ngoc-Duy and Le, Xuan-Bach and Nguyen, Minh-Tuan and Nguyen, Huu-Dong and Nguyen, Van-Huan and Nguyen, Ha-Thanh and Vuong, Thi-Hai-Yen, 2024 16th International Conference on Knowledge and System Engineering (KSE), pages 447–452, 2024, IEEE
- [7] NOWJ@ALQAC 2023: Enhancing legal task performance with classic statistical models and pre-trained language models, Nguyen, Tan-Minh and Nguyen, Xuan-Hoa and Mai, Ngoc-Duy and Hoang, Minh-Quan and Nguyen, Van-Huan and Nguyen, Hoang-Viet and Nguyen, Ha-Thanh and Vuong, Thi-Hai-Yen, 2023, <https://arxiv.org/abs/2309.09070>.
- [8] Enhancing legal document retrieval: a multi-phase approach with large language models, Nguyen, Hai-Long and Nguyen, Duc-Minh and Nguyen, Tan-Minh and Nguyen, Ha-Thanh and Vuong, Thi-Hai-Yen and Satoh, Ken, 2024, arXiv preprint arXiv:2403.18093.
- [9] XOSA-NOWJ@ALQAC 2024: Efficient Legal Information Processing with the Combination of Semantic-Based Model and LLMs, Nguyen, Hai-Long and Nguyen, Duc-Minh and Nguyen, Quang-Anh and Chu, Huy-Quang and Nguyen, Huu-Dung and Nguyen, Khanh-Huyen and Pham, Cong-Minh and Vuong, Thi-Hai-Yen and Nguyen, Ha-Thanh, 2024, 16th International Conference on Knowledge and System Engineering (KSE), pages 433–439, 2024, IEEE.
- [10] From TF-IDF to Instruction-Tuned LLMs: Hybrid Legal Reasoning Systems for COLIEE 2025, Euijin Baek, Jiayi Dai, H M Quamran Hasan, Yeji Kim, Housam Babiker, Mi-Young, Kim and Randy Goebel, 2025, COLIEE 2025.
- [11] SIL@COLIEE2025: A Cascading framework for finding relevant Case Laws, Bhavya Jain, Pooja Harde, Taha Sadikot, Eric Namit Kujur and Sarika Jain, 2025, COLIEE 2025.
- [12] NOWJ@COLIEE2025: A Multi-stage Framework Integrating Embedding Models and Large Language Models for Legal Retrieval and Entailment, Hoang-Trung Nguyen, Thi-Hai-Yen Vuong, Tan-Minh Nguyen, Xuan-Bach Le, Tuan-Kiet Le, Khanh-Huyen Nguyen, Ha-Thanh Nguyen and Le-Minh Nguyen, COLIEE 2025.
- [13] JNLP@COLIEE 2025: Hybrid LLM-based Framework for Legal Information Retrieval and Entailment|JNLP at COLIEE 2025: Hybrid Large Language Model-based Framework for Legal Information Retrieval and Entailment, The Hai Nguyen, Khac Vu Hiep Nguyen, Ngoc Anh Trang Pham, Ngoc Minh Nguyen, Hoang An Trieu, Nguyen Khang Le, Dinh Truong Do and Le Minh Nguyen, COLIEE 2025.
- [14] se7enese@ALQAC 2024: Experimenting New Released Models in Legal Document Retrieval and Question Answering, Le, Hoang-Bao and Zhou, Liting and Gurrin, Cathal, 2024 16th International Conference on Knowledge and System Engineering (KSE), 2024, IEEE.
- [15] AIEPU at ALQAC 2023: deep learning methods for legal information retrieval and question answering, Hoang, Long and Bui, Tung and Nguyen, Chau and Nguyen, Le-Minh, 2023 15th International Conference on Knowledge and Systems Engineering (KSE), 2023, IEEE.

UTEHY-NLU@ALQAC 2025: Dynamic Weighted Ensemble and Adaptive Reasoning for Vietnamese Legal Text Processing

Truong-Phuc Nguyen, Quy-Nhan Nguyen, Manh-Cuong Phan,
Chi-Hai Cao, Trinh-Hoai-An Duong, and Minh-Tien Nguyen*
Hung Yen University of Technology and Education, Hung Yen, Vietnam
{nguyentruongphuc_12421tn, nguyenquynhan, tiennm}@utehy.edu.vn;
cuongpm@spkt.edu.vn; {caochihai1710, anduongtrinhhoai}@gmail.com

Abstract—This paper presents our approach to the ALQAC 2025 competition for Vietnamese legal text processing. For information retrieval, we propose an Advanced Dynamic Weighted Ensemble method combining three dense retrieval models with query-specific confidence calculation and dynamic weighting, followed by cross-encoder reranking. For question answering, we develop an Adaptive Few-shot Chain-of-Thought Reasoning approach using Qwen3-4B/8B models with specialized prompting and dynamic classification between extractive and abstractive QA strategies. Our results demonstrate strong performance: 5th place in Legal Information Retrieval (F2-score: 0.8482, 7.51% improvement over single models) and 4th place in Legal Question Answering (97.56% accuracy). Reasoning enhancement provides significant gains, with 5.60% improvement in Multiple Choice and 12.50% in Extractive QA tasks. Our adaptive Free-Text approach outperforms pure strategies by 2.12%, validating the effectiveness of ensemble methods and reasoning-enhanced approaches for Vietnamese legal AI systems.

Index Terms—Legal AI, information retrieval, question answering, large language models, ensemble models, reasoning models

I. INTRODUCTION

The rapid proliferation of legal documents in Vietnam has created an urgent need for efficient information access and processing systems. Legal documents often require specialized knowledge due to numerous technical terms and intricate regulations, making them challenging for non-experts to understand and utilize effectively. The Automated Legal Question Answering Competition (ALQAC) [1], held annually since 2021, provides a crucial platform for advancing Vietnamese legal text processing. ALQAC 2025 focuses on two fundamental tasks: Legal Information Retrieval, which involves finding relevant legal documents based on specific questions, and Legal Question Answering, which encompasses responding to True/False, Multiple Choice, and Free-Text legal queries.

This paper presents our comprehensive approach to both tasks. For Legal Information Retrieval, we propose an Advanced Dynamic Weighted Ensemble method leveraging three

dense retrieval models through combination strategies, followed by cross-encoder reranking. For Legal Question Answering, we develop an Adaptive Few-shot Chain-of-Thought Reasoning approach combining advanced prompting techniques with LLMs reasoning capabilities. Our methodology achieves strong performance: 5th place with F2-score 0.8482 in Legal Information Retrieval and 4th place with 97.56% accuracy in Legal Question Answering.

The remainder of this paper is organized as follows. Section 2 reviews related work on Vietnamese legal text processing. Section 3 presents our Advanced Dynamic Weighted Ensemble for retrieval and Adaptive Few-shot Chain-of-Thought Reasoning for QA. Section 4 discusses the experimental setup and results analysis. Section 5 concludes with our key findings and future research directions.

II. RELATED WORKS

A. Legal Information Retrieval

Early approaches using rule-based systems and statistical methods (n-gram models, HMMs) demonstrated limitations in capturing the semantic nuances and contextual complexities of legal documents. The emergence of neural network models, particularly Transformer-based architectures [2] such as BERT [3] and its variants, has revolutionized legal text processing by enabling better semantic understanding and contextual representation. These models have shown remarkable success in capturing the intricate relationships between legal concepts and improving document understanding capabilities.

In the ALQAC competition series, several notable approaches have emerged. For ALQAC 2024, the winning team NOWJ [4] employed a sophisticated multi-layer ensemble strategy combining lexical matching techniques (BM25, QLD) with advanced semantic models (mBERT, sBERT, Vi-MonoT5). Their approach integrated LightGBM for feature combination and utilized GPT-4 for post-processing refinement, achieving the highest F2-score of 0.8771 in Task 1. XOSA-NOWJ [5] implemented a four-stage pipeline incorporating lexical ranking with BM25, followed by semantic rank-

*Corresponding author: Minh-Tien Nguyen (tiennm@utehy.edu.vn).

ing using multiple BERT variants and ColBERT, concluding with ensemble optimization through grid-search techniques.

In ALQAC 2023, the NeCo team [6] demonstrated the importance of domain adaptation by combining BM25 with multilingual BERT (mBERT) through ensemble strategies. They enhanced performance through data enrichment techniques, fine-tuning mBERT on 21,476 crawled legal samples to improve semantic understanding in the Vietnamese legal domain. This approach achieved an F2-macro score of 86.61% and secured first place in Task 1, highlighting the critical role of domain-specific data augmentation.

B. Legal Question Answering

Legal question answering presents unique challenges due to the specialized knowledge requirements, complex reasoning processes, and diverse question formats encountered in legal domains. Recent research has focused on developing sophisticated approaches that can handle the intricacies of legal reasoning while maintaining high accuracy across different question types. In ALQAC 2024, multiple teams demonstrated innovative approaches. The NOWJ team [4] utilized three state-of-the-art large language models (LLama-3 [7], Gemini [8], GPT-4 [9]) with specialized prompt templates tailored to each question type. They employed deterministic generation (temperature=0) to ensure consistent outputs and implemented comprehensive post-processing pipelines, followed by ensemble voting mechanisms to select optimal answers. UIT-DarkCow [10] focused on prompt engineering optimization, systematically testing various prompting strategies with the vietcuna-7b-v3-AWQ model and selecting configurations that demonstrated superior performance on public test sets.

The ALQAC 2023 competition witnessed significant advances in prompt engineering and multilingual approaches. AIEPU [11] pioneered comprehensive prompt tuning methodologies using three large language models: Flan-T5-XXL and Flan-Alpaca-XXL [12] for English-based prompting (incorporating Google Translate for cross-lingual transfer), and MT0-XXL for native Vietnamese processing. They evaluated over 50 prompts incorporating Zero-shot CoT and Legal Reasoning techniques, achieving 86.37% accuracy through ensemble voting of Vietnamese and English prompt variants, securing first place and demonstrating multi-prompt ensemble effectiveness.

III. METHODOLOGY

A. Legal Information Retrieval

Given a legal question q and a corpus of legal documents $D = \{d_1, d_2, \dots, d_n\}$, the task aims to retrieve the most relevant articles that can adequately answer the question. The retrieval system can be formulated as finding the optimal article d^* from the corpus D :

$$d^* = \arg \max_{d \in D} R(q, d) \quad (1)$$

where $R(q, d)$ represents the relevance score between query q and article d . For dense retrieval approaches, this relevance is typically computed as:

$$R(q, d) = \text{sim}(\mathbf{e}_q, \mathbf{e}_d) = \frac{\mathbf{e}_q \cdot \mathbf{e}_d}{\|\mathbf{e}_q\| \|\mathbf{e}_d\|} \quad (2)$$

where $\mathbf{e}_q$ and $\mathbf{e}_d$ are the dense embeddings of query q and document d respectively, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity between vectors.

Figure 1 illustrates our proposed Advanced Dynamic Weighted Ensemble pipeline, consisting of five main key stages: triplet dense retrieval, query-specific confidence calculation, advanced dynamic weighting, ensemble score fusion, and cross-encoder reranking.

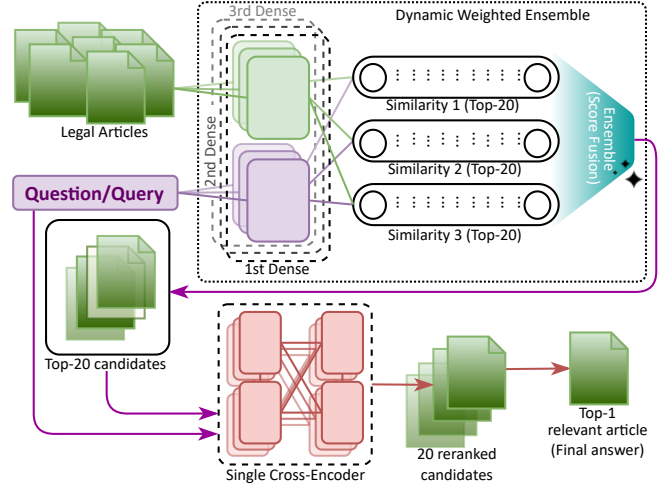


Figure 1. The Overview of Advanced Dynamic Weighted Ensemble pipeline.

1) *Triplet Dense Retrieval Ensemble*: Our main contribution employs an ensemble of three dense retrieval models selected based on comprehensive empirical evaluation of individual model performance on Vietnamese legal texts:

- **AITeamVN/Vietnamese_Embedding** [13]: Vietnamese-specific embedding model fine-tuned with 300000 hard negative samples.
- **BAAI/bge-m3** [14]: Multilingual dense retrieval model with proven cross-lingual capabilities.
- **intfloat/multilingual-e5-large** [15]: Large-scale multilingual embeddings with extensive language coverage.

Each model independently processes the input query q and generates embeddings for both the query and all legal articles. Using FAISS similarity search with inner product distance, each model retrieves its top-20 most similar articles, resulting in a total of 60 legal document candidates with potential overlaps across models, which enables capturing complementary retrieval strengths.

2) *Query-Specific Confidence Calculation*: A critical innovation in our approach is the development of query-specific confidence scores that adapt to each model's characteristic score distributions. For each model i , we calculate confidence using the top-3 similarity scores to ensure statistical stability while maintaining computational efficiency:

$$c_i = \frac{1}{3} \sum_{j=1}^3 \text{normalize}(s_{i,j}) \times \beta_i \quad (3)$$

where $s_{i,j}$ represents the j -th highest cosine similarity score from model i . The normalization function adapts to model-specific score distributions observed during empirical analysis:

$$\text{normalize}(s) = \text{clip} \left(\frac{s - s_{\min}^{(i)}}{s_{\max}^{(i)} - s_{\min}^{(i)}}, 0, 1 \right) \quad (4)$$

The model-specific parameters $s_{\min}^{(i)}$ and $s_{\max}^{(i)}$ are determined through systematic analysis of score distributions on the training set: - Vietnamese_Embedding: $s_{\min} = 0.30$, $s_{\max} = 0.85$ (reflecting domain-specific score characteristics) - bge-m3: $s_{\min} = 0.40$, $s_{\max} = 0.90$ (higher baseline due to multilingual pre-training) - multilingual-e5-large: $s_{\min} = 0.60$, $s_{\max} = 0.95$ (highest score range among the 3 models).

The quality boost parameter β_i reflects domain specialization and is empirically determined based on retrieval-only performance analysis from Table III: $\beta_{\text{vietnamese}} = 1.2$ (20% boost for Vietnamese-general domain expertise), $\beta_{\text{bge}} = 1.0$ (baseline reference), $\beta_{\text{e5}} = 0.8$ (20% penalty for general-purpose nature). This confidence mechanism ensures fair comparison across models with different score distributions while rewarding language-specific expertise.

3) *Advanced Dynamic Weighting Strategy*: Unlike static ensemble approaches that use fixed weights, our method dynamically adjusts model weights based on both historical F2-performance and real-time query confidence, enabling adaptive responses to varying query characteristics:

$$w_i = w_{\text{base},i} \times (0.7 \times w_{f2,i} + 0.3 \times w_{\text{conf},i}) \quad (5)$$

where the historical performance component $w_{f2,i} = \frac{F2_i}{F2_{\text{best}}}$ leverages individual model validation performance from the retrieval-only experiments in Table III. Specifically, $F2_i$ represents each model's F2-score in the single retrieval setting: Vietnamese_Embedding achieves $F2_{\text{vietnamese}} = 0.8255$, bge-m3 achieves $F2_{\text{bge}} = 0.7414$, and multilingual-e5-large achieves $F2_{\text{e5}} = 0.7226$. The term $F2_{\text{best}} = 0.8255$ corresponds to the highest-performing single model (Vietnamese_Embedding), serving as the normalization reference.

The confidence adjustment component $w_{\text{conf},i} = 0.7 + (c_i \times 0.6)$ provides query-adaptive scaling in the range $[0.7, 1.3]$, allowing models with high confidence to receive up to 30% weight boost while ensuring minimum 70% retention for low-confidence scenarios. The optimal 70-30 split ratio between historical performance and real-time confidence was determined through comprehensive grid search optimization across ratios from 50-50 to 90-10 with 10% increments, evaluated on the training set using F2-score as the optimization criterion. This configuration ensures that proven model reliability dominates the weighting scheme while providing sufficient adaptability to query-specific model confidence patterns.

4) *Ensemble Score Fusion with Rank-Based Discounting*: The 60 candidates from all models are merged using a sophisticated weighted combination strategy that incorporates both similarity scores and ranking positions:

$$S_{\text{final}}(q, d) = \sum_{i=1}^3 w_i \times S_i(q, d) \times \frac{1}{1 + \text{rank}_i(d) \times 0.1} \quad (6)$$

where $\text{rank}_i(d)$ represents the ranking position of article d in model i . The rank-based discount factor of 0.1 was determined through systematic empirical evaluation on the training set, testing values from 0.0 to 0.3 with 0.05 increments using 5-fold cross-validation. This mechanism prevents over-reliance on any single model's ranking while ensuring that consistently high-ranking articles across multiple models receive appropriate emphasis.

After score aggregation and deduplication of overlapping candidates, the top-10 articles with highest ensemble scores are selected for the final reranking stage, providing a focused set of high-quality candidates for precise relevance assessment.

5) *Cross-Encoder Reranking*: The final stage employs a cross-encoder architecture to perform precise pairwise relevance scoring between the query and each of the top-10 ensemble candidates. We utilize either **BAAI/bge-reranker-v2-m3** [14] for multilingual capability or **AITeamVN/Vietnamese_Reranker** [13] for Vietnamese-specific optimization. The cross-encoder performs precise pairwise relevance scoring between the query and each top-10 ensemble candidate, with the highest-scored article constituting the final system output.

B. Legal Question Answering

Given a legal question q and relevant articles $A = \{a_1, a_2, \dots, a_k\}$ retrieved from Task 1, our system addresses three distinct question types: True/False, Multiple Choice (MCQ), and Free-Text. The task is formulated as:

$$\hat{y} = f(q, A, \theta) \quad (7)$$

where $f(\cdot)$ represents the question answering function parameterized by θ , and $\hat{y}$ is the predicted answer. Evaluation employs accuracy metrics for True/False and MCQ questions, while Free-Text responses undergo expert assessment. Figure 2 illustrates our comprehensive question answering pipeline for ALQAC 2025 Task 2.

For True/False and Multiple Choice questions, we use specialized system prompts (#1, #2) with legal context and question content. For Free-Text questions, we found 41.10% require Extractive QA while the remainder need Abstractive QA. We implement adaptive classification (prompt #3) to determine the optimal strategy, then apply either Extractive QA (prompt #4) or Abstractive QA (prompt #5).

1) *Prompting Format*: Modern LLMs support ChatML templates, enabling deeper intervention in response generation through sophisticated system prompting. To address the three

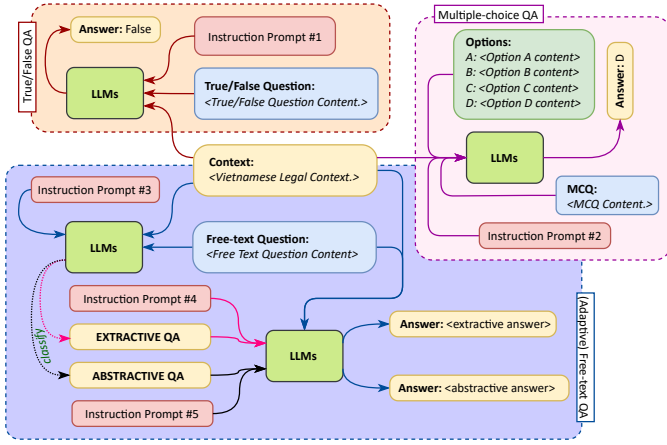


Figure 2. The Overview of Legal Question Answering Pipeline in ALQAC 2025.

distinct question types in ALQAC 2025 Task 2, we developed specialized system prompts incorporating established techniques including CoT reasoning and few-shot prompting. For True/False questions, our system prompt provides comprehensive guidance covering legal analysis procedures, evaluation criteria for accuracy assessment, common legal domain knowledge, standardized output formats, and concrete examples within detailed role-based instructions. The MCQ prompt extends this framework with additional answer selection criteria, systematic elimination strategies, and question type categorization guidance. We developed specialized system prompts incorporating CoT reasoning and few-shot prompting for each question type. True/False and Multiple Choice prompts provide comprehensive legal analysis guidance, while Free-Text prompts implement two-stage classification between extractive and abstractive strategies with confidence thresholds (0.6). Complete system prompts are available at: <https://github.com/ntphuc149/UTEHY-NLU-ALQAC-2025>.

Table I
USER PROMPTING FORMAT USED IN ALQAC 2025 LEGAL QA TASK.

Question Type	User Prompt Format	
	Vietnamese	English
True/False, Free Text (classification, extractive QA, abstractive QA)	TRÍCH DẪN PHÁP LÝ: <legal context>	LEGAL CITATION: <legal context>
	CÂU HỎI: <question>	QUESTION: <question>
Multiple Choice	TRÍCH DẪN PHÁP LÝ: <legal context>	LEGAL CITATION: <legal context>
	CÂU HỎI: <question>	QUESTION: <question>
	CÁC LỰA CHỌN: A: <option A> B: <option B> C: <option C> D: <option D>	OPTIONS: A: <option A> B: <option B> C: <option C> D: <option D>

The formatting prompts for users are shown in Table I. True/False and Free-Text QA both use the same prompt format

including legal context and question. Meanwhile, Multiple Choice QA requires selecting the correct answer from a set of options based on the legal text and question provided, so it is necessary to include the answer options. Both system prompts and user prompts are formatted as input for LLMs to process and provide responses.

2) *Reasoning Mode*: Models such as Qwen3 (4B/8B) are used to leverage reasoning capabilities for answering complex legal questions that require multi-step processing to produce the final answer. Qwen3 enabling and disabling the reasoning feature via the `enable_thinking` argument in the `apply_chat_template()` method from the tokenizer. When not using reasoning, we limit the number of tokens generated to 3, 1, 30, 500, and 4,096 for True/False, MCQ, Free-Text question classification, Extractive QA (free-text span extraction), and Abstractive QA (free-text answer generation), respectively. When reasoning is enabled, the number of new tokens is increased to 32,768 to ensure that the reasoning process is completed and the desired answer is obtained after the reasoning process. In this study, we did not use Qwen2.5 or Llama3.2 because these LLMs do not support built-in step-by-step reasoning capabilities like Qwen3.

IV. EXPERIMENTS AND RESULTS

A. Dataset

We use the ALQAC 2025 dataset provided for training to evaluate the performance of the models on two main tasks. Detailed statistics on training data and private test are presented in Table II.

Table II
STATISTICS OF QUESTION TYPE IN THE ALQAC 2025 DATASETS. T/F AND MCQ DENOTE TRUE/FALSE AND MULTIPLE CHOICE QUESTION, RESPECTIVELY.

	Training set			Private test set		
	T/F	MCQ	Free-text	T/F	MCQ	Free-text
Num. questions	387	286	56	11	60	11
Max question length	66.00	60.00	53.00	27.00	69.00	76.00
Min question length	11.00	4.00	6.00	13.00	7.00	8.00
Avg. question length	26.67	15.67	19.57	20.18	19.82	32.18

B. Experiments Setup

We evaluated three retrieval settings: single embedding model, embedding with reranker, and our proposed ensemble combination. Experiments used FAISS for similarity search on Nvidia P100 GPU (16GB VRAM). For QA task, 4B models ran on a single P100 GPU while 8B models required 2 T4 GPUs (2×15GB VRAM).

C. Evaluation Metrics

Legal Information Retrieval performance is evaluated using precision, recall, and F2-measure:

$$\text{Pre}_i = \frac{C_i}{R_i}; \text{Rec}_i = \frac{C_i}{G_i}; \text{F2}_i = \frac{5 \times \text{Pre}_i \times \text{Rec}_i}{4 \times \text{Pre}_i + \text{Rec}_i} \quad (8)$$

where Pre_i and Rec_i denote precision and recall for question i , and C_i , R_i , and G_i denote correctly retrieved, total retrieved, and ground truth relevant articles for question i . The macro-average F2-measure serves as the principal metric.

Legal Question Answering employs accuracy as the primary metric. True/False and Multiple Choice questions are evaluated by direct answer comparison, while Free-Text answers are evaluated by human legal experts.

D. Results

1) *Legal Information Retrieval*: Table III demonstrates a clear performance hierarchy: retrieval-only methods achieve the lowest performance, followed by retrieval with reranking, while our proposed ensemble achieves the highest performance. Vietnamese_Embedding achieves the highest single-model F2-score (0.8255) due to Vietnamese hard negative samples fine-tuning, outperforming bge-m3 (0.7414) and multilingual-e5-large (0.7226). Vietnamese_Reranker consistently outperforms bge-reranker (0.8862 vs 0.8572). Our ensemble method achieves F2-score 0.8875, representing 7.51% improvement over the best single model and 0.15% enhancement over two-stage systems.

Table III
TEAM’S PIPELINE PERFORMANCE FOR LEGAL IR TASK IN ALQAC 2025 TRAINING SET. **BOLD** VALUES INDICATE THE BEST PERFORMANCE IN EACH SETTING, WHILE UNDERLINED VALUES INDICATE THE SECOND BEST. ‘-’ REPRESENTS SETTING WITHOUT RERANKER MODEL.

Retrieval	Reranker	Precision	Recall	F2-Score
Vietnamese_Embedding		0.8285	0.8251	0.8255
bge-m3		<u>0.7462</u>	<u>0.7407</u>	0.7414
multilingual-e5-large	-	0.7257	0.7222	0.7226
Qwen3-Embedding-0.6B		0.6968	0.6934	0.6938
multilingual-e5-base		0.6296	0.6262	0.6266
Vietnamese_Embedding		0.8711	0.8656	0.8662
bge-m3	bge	0.8587	0.8532	0.8538
multilingual-e5-large	-reranker	0.8546	0.8491	0.8497
Qwen3-Embedding-0.6B	-v2-m3	0.8395	0.8340	0.8346
multilingual-e5-base		0.8450	0.8395	0.8401
Vietnamese_Embedding		0.8861	0.8820	0.8825
bge-m3		<u>0.8724</u>	<u>0.8676</u>	<u>0.8682</u>
multilingual-e5-large	Vietnamese	0.8615	0.8567	0.8572
Qwen3-Embedding-0.6B	_Reranker	0.8464	0.8416	0.8421
multilingual-e5-base		0.8601	0.8546	0.8552
Vietnamese_Embedding	bge			
+ bge-m3	-reranker	<u>0.8797</u>	<u>0.8742</u>	<u>0.8748</u>
+ multilingual-e5-large	-v2-m3			
Vietnamese_Embedding	Vietnamese	0.8875	0.8827	0.8832
+ bge-m3	_Reranker			
+ multilingual-e5-large				

Table IV shows competition results. We ranked 5th with F2-score 0.8482, trailing top teams by 0.0284-0.0952 points. The methodology’s consistent performance across training and private test sets suggests potential transferability to other legal systems through domain-appropriate model substitution and parameter recalibration.

2) *Legal Question Answering*: For Task 2, we evaluate five distinct QA configurations: True/False, MCQ, Free Text Extractive QA (span extraction), Free Text Abstractive QA

Table IV
RESULT OF LEGAL IR TASK IN ALQAC 2025 PRIVATE TEST SET.

Team	F2	Rank
Flame Reavers	0.9434765774	1
vund	0.8970189700	2
UIT legalGrep	0.8970189700	2
NOWJ	0.8766937667	4
UTEHY-NLU	0.8482384822	5
Hallucineers	0.8279132790	6
D2	0.8238482383	7
UIT-Darkcow	0.7872628725	8
ABCD	0.7730352302	9
Warriors	0.6897018969	10
LAD	0.6588533415	11
BAMBOOVN	0.5555555554	12
VnLegalLLM25	0.5189701896	13

(answer generation), and Adaptive Free-Text QA (dynamic classification-based switching). To accelerate evaluation on Free-Text questions, we employ Qwen3-8B with thinking mode for automated answer correctness assessment against ground truth labels. To validate reliability, five members independently evaluated 50 randomly sampled answers using criteria of relevance, factual accuracy, and ground truth alignment, achieving 98.40% consensus with the LLM judge decisions. Experimental results on the ALQAC 2025 training dataset are presented in Table V.

Table V
TEAM’S PIPELINE PERFORMANCE FOR LEGAL QA TASK IN ALQAC 2025 TRAINING SET. **BOLD** AND UNDERLINED VALUES INDICATE BEST AND SECOND-BEST PERFORMANCE. REASONING IMPRV. SHOWS GAINS FROM REASONING ENHANCEMENT. EQA, AQA, ADAPTIVE DENOTE EXTRACTIVE QA, ABSTRACTIVE QA, AND ADAPTIVE ROUTING STRATEGIES.

QA Task	Accuracy		Reasoning Imprv.	
	Reasoning: No / Yes	Qwen3-4B	Qwen3-8B	(4B / 8B)
True/False		89.66 / <u>92.51</u>	91.99 / 93.28	+2.85 / +1.29
Multiple Choice		85.31 / <u>90.91</u>	89.51 / 92.31	+5.60 / +2.80
Free Text	EQA	69.64 / 82.14	80.36 / <u>83.93</u>	+12.50 / +3.57
	AQA	71.43 / 80.36	78.57 / 82.14	<u>+8.93</u> / +3.57
	Adaptive	76.79 / <u>83.93</u>	80.36 / 85.71	+7.14 / +5.35

Table V demonstrates consistent performance patterns across question types and model configurations. True/False questions achieve the highest accuracy (89.66-93.28%), outperforming MCQ tasks (85.31-92.31%) due to binary versus four-option selection complexity. Free-Text QA exhibits lower performance (69.64-85.71%), reflecting open-ended question difficulty. Model scale significantly impacts performance, with 8B models consistently outperforming 4B variants across all tasks. Qwen3-8B achieves 2.33% improvement over Qwen3-4B on True/False questions, with similar patterns across MCQ (4.20% improvement) and Free-Text tasks.

Reasoning enhancement provides substantial gains across all configurations. Remarkably, 4B models with reasoning consistently outperform 8B models without reasoning: True/False (92.51% vs 91.99%), MCQ (90.91% vs 89.51%), EQA (82.14% vs 80.36%), AQA (80.36% vs 78.57%), and Adaptive Free-Text QA (83.93% vs 80.36%). This demonstrates rea-

soning can effectively compensate for model size limitations, offering promising directions for resource-constrained deployments. The most significant reasoning improvements occur in Extractive QA (12.50% for 4B, 3.57% for 8B), Abstractive QA (8.93% and 3.57%) and Adaptive Free Text QA (7.14% and 5.35%) while Multiple Choice shows moderate gains (5.60% for 4B, 2.80% for 8B).

Our Adaptive Free-Text approach achieves optimal performance (85.71%), outperforming pure EQA (83.93%) and AQA (82.14%), with 4B Adaptive models significantly outperforming single strategies (7.15% and 5.36% improvements), highlighting intelligent strategy selection value for resource-constrained deployments. Table VI presents the final rankings on the ALQAC 2025 private test set as announced by the competition organizers.

Table VI
RESULT OF LEGAL QA TASK IN ALQAC 2025 PRIVATE TEST SET.

Team	Accuracy	Rank
NOWJ	0.98780487800	1
Hallucineers	0.98780487800	1
UIT-Darkcow	0.98780487800	1
UTEHY-NLU	0.97560975600	4
FlameReavers	0.96341463400	5
vund	0.96341463400	5
BAMBOOVN	0.95121951200	7
ABCD	0.93902439000	8
D2	0.93902439000	8
VnLegalLLM25	0.90243902400	10
Warriors	0.81707317100	11
UIT legalGrep	0.75609756100	12
LAD	0.30487804900	13

Our team (UTEHY-NLU) achieved 4th place with an accuracy of 0.9756, demonstrating the effectiveness of our Adaptive Few-shot CoT Reasoning QA approach. Our method achieved competitive performance, trailing the top three teams (NOWJ, Hallucineers, and UIT-Darkcow) by only 0.0122 points, all of whom achieved identical scores of 0.9878. This narrow margin of 1.22 percentage points indicates that our reasoning-enhanced methodology performs competitively against state-of-the-art approaches in Vietnamese legal QA.

V. CONCLUSION

This paper presented our methodology for the ALQAC 2025 competition, achieving 5th place in Legal IR (F2-score: 0.8482) and 4th place in Legal QA (97.56% accuracy), trailing top teams by only 1.22 percentage points. Our Advanced Dynamic Weighted Ensemble method combines three dense retrieval models with query-specific confidence calculation and dynamic weighting, achieving 7.51% improvement over single models. Our Adaptive Few-shot CoT Reasoning approach leverages LLMs reasoning capabilities with specialized prompting and dynamic classification between extractive and abstractive strategies.

Key findings include: (1) Reasoning enhancement enables 4B models to outperform 8B non-reasoning models across all tasks consistently; (2) Reasoning provides substantial gains

(1.29-12.50% improvement), with largest benefits in AQA (8.93%) and EQA (12.50%); (3) Adaptive Free-Text approach outperforms pure strategies by 2.12%. Our contributions demonstrate the effectiveness of ensemble methods and reasoning-enhanced approaches for Vietnamese legal AI systems, providing practical insights for resource-efficient legal support systems. Future investigations will include component-wise ablation analysis and detailed failure case studies to identify challenges with ambiguous terminology and long-context scenarios.

REFERENCES

- [1] N. H. Thanh, B. M. Quan, C. Nguyen, T. Le, N. M. Phuong, D. T. Binh, V. T. H. Yen, T. Racharak, N. Le Minh, T. D. Vu, P. V. Anh, N. T. Son, H. T. Nguyen, B. Butr-indr, P. Vateekul, and P. Boonkwan, "A summary of the alqac 2021 competition," in *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, 2021, pp. 1–5.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [4] N.-D. Mai, X.-B. Le, M.-T. Nguyen, H.-D. Nguyen, V.-H. Nguyen, H.-T. Nguyen, and T.-H.-Y. Vuong, "Nowj@ alqac 2024: Leveraging document ranking methods and large language models for legal document retrieval and question answering," in *2024 16th International Conference on Knowledge and System Engineering (KSE)*. IEEE, 2024, pp. 447–452.
- [5] H.-L. Nguyen, D.-M. Nguyen, Q.-A. Nguyen, H.-Q. Chu, H.-D. Nguyen, K.-H. Nguyen, C.-M. Pham, T.-H.-Y. Vuong, and H.-T. Nguyen, "Xosanowj@ alqac 2024: Efficient legal information processing with the combination of semantic-based model and llms," in *2024 16th International Conference on Knowledge and System Engineering (KSE)*. IEEE, 2024, pp. 433–439.
- [6] H.-L. Nguyen, D.-Q. Nguyen, H.-T. Nguyen, T.-T. Pham, H.-D. Nguyen, T.-A. Nguyen, T.-H.-Y. Vuong, and H.-T. Nguyen, "Neco@alqac 2023: Legal domain knowledge acquisition for low-resource languages through data enrichment," 2023.
- [7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.
- [8] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
- [9] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [10] H. Q. Pham, Q. V. Nguyen, D. Q. Tran, T. K.-B. Nguyen, and K. V. Nguyen, "Top 2 at alqac 2024: Large language models (llms) for legal question answering," *International Journal of Asian Language Processing*, vol. 35, no. 01, p. 2450010, 2025.
- [11] L. Hoang, T. Bui, C. Nguyen, and L.-M. Nguyen, "Aiepu at alqac 2023: deep learning methods for legal information retrieval and question answering," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2023, pp. 1–6.
- [12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [13] N. N. Q. Nguyen Nho Trung, "Vietnamese_embedding: Embedding model in vietnamese language," 2025.
- [14] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," 2024.
- [15] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," *arXiv preprint arXiv:2402.05672*, 2024.

The Evolution of Defined Terms in Law: A Diachronic Study Using Open Legal Data

1st Makoto Nakamura
dept. Engineering
Niigata Institute of Technology
Niigata, Japan
mnakamur@niit.ac.jp

2nd Daichi Saito
Conphic Co.,Ltd.
Tokyo, Japan
daichi@conphic.co.jp

Abstract—This study examines the diachronic evolution of legal terms in Japanese statutes using open government data. Defined terms were extracted from XML-formatted laws provided through e-Gov, and changes across amendments were tracked to analyze additions, deletions, and revisions over time. This method enables quantitative observation of changes in legal language, even within a limited time span. The approach is designed to support future extensions to long-term historical data and comparative legal analysis through compatibility with international standards such as Akoma Ntoso.

Index Terms—language evolution, legal text processing, legal terms, Japanese statutes, terminology

I. INTRODUCTION

In this paper, we present our ongoing research on the analysis of legal terminology from the perspective of language evolution. A central aim of evolutionary linguistics is to explain how the structure of language emerged and changed over time [Hurford, 2002]. The field has become increasingly interdisciplinary and now involves collaboration among linguists, philosophers, biologists, cognitive scientists, roboticists, and computational modelers [Lyon et al., 2007].

Simulation studies have played an important role in exploring hypothetical scenarios of language emergence [Cangelosi and Parisi, 2002]. However, such models are often highly abstract and struggle to capture the complexity of real-world language use. Another persistent challenge in studying language evolution is the scarcity of empirical data, since spoken language leaves few, if any, lasting records. As a result, many models operate without grounding in actual linguistic history.

Legal language provides a unique way to address this problem. Although it is not a direct record of the spoken language used during human evolution, it represents a carefully documented system of linguistic expressions that are explicitly defined and revised over time. The evolution of legal terminology, especially defined terms in statutes, can be traced in detail and may serve as a proxy for observing structured linguistic change in a real-world setting. Thus, legal terms in statutory corpora offer a rare opportunity to bridge the gap between abstract models and historically grounded language data.

This study builds on our previous research [Nakamura et al., 2016]. At that time, access to Japanese legal data was limited, and researchers had to rely on paid services offered by private

companies. It was also difficult to obtain versions of laws that were valid at specific points in time. With the advent of open access to Japanese legislation, however, the availability and usability of legal data have improved dramatically. Since comprehensive statutory data became publicly available, it has become possible to retrieve historical versions of laws as they stood on particular dates. Moreover, the data are structured in XML and accessible via an API, enabling systematic and precise access to specific provisions.

The goal of this study is to use Japan’s open legislative data to systematically collect defined terms and analyze their diachronic evolution from a language-change perspective. In earlier work, we attempted to simulate the evolution of legal terminology using historical data as input to an evolutionary model [Nakamura et al., 2017]. However, the dataset used in that study was limited in both size and scope. With the improved accessibility and granularity of current legal data, it is now feasible to construct a more comprehensive model of linguistic change in Japanese legal language, grounded in real-world evidence.

The remainder of this paper is organized as follows. Section II discusses the role of legal terms from the perspective of legal informatics. Section III introduces the Japanese Legal Standard XML Schema (JLS) and the structure of legal data in e-Gov. Section IV situates our study within the broader context of language evolution. Section V explains the method for extracting defined terms and tracking their diachronic changes. Section VI presents our analysis and results. Finally, Section VII concludes the paper and outlines future directions.

II. LEGAL TERMS IN LEGAL INFORMATICS

In this section, we explain our work from the viewpoint of legal informatics.

A. Provision of Defined Terms

In statutory texts, key terms are generally defined explicitly before their first use. This study focuses on such defined legal terms, each consisting of a pair of a term and its corresponding explanation. Definitions are usually stipulated in Article 2, immediately after the article specifying the purpose of the act, and are therefore considered to apply to the entire statute.

[as of May 31, 2017]

(Definitions)

Article 2 (1) The term “retail service” as used in this Act means supplying gas via pipelines to meet general demand (in the case of generating gas at a simplified gas generating facility specified by Cabinet Order (hereinafter referred to as a “specified gas generating facility”) and supplying the gas via pipelines, limited to the case of supplying gas at not less than 70 gas service points within one housing complex).

(2) The term “gas retail business” as used in this Act means the business of providing a Retail Service (excluding the part that constitutes general gas pipeline service business, specified gas pipeline service business or gas manufacturing business).

(*omitting*)

Fig. 1. Example of defined terms (1)

(Definitions)

Article 2 In this Act, the meanings of the terms set forth in the following items are as prescribed respectively in those items:

(i) “specified telecommunications” means sending telecommunications aimed at reception by unspecified persons (meaning telecommunications as defined in Article 2, item (i) of the Telecommunications Business Act (Law No. 86 of 1984); hereinafter the same applies in this item) (excluding sending telecommunications aimed at direct reception by the public);

(ii) “specified telecommunications facilities” means telecommunications facilities (meaning telecommunications facilities as defined in Article 2, item (ii) of the Telecommunications Business Act) which are used for specified telecommunications; (*omitting*)

Fig. 2. Example of defined terms (2)

The scope of what constitutes a “legal term” depends on the research purpose [Lame, 2005], [Höfler et al., 2011], [Winkels and Hoekstra, 2012]. In this paper, we regard legal terms as those explicitly defined in the law prior to their use, represented as a pair of a quoted term and its explanatory sentence. Such definitions typically appear in one of the following forms:

- (1) A stand-alone provision
- (2) A parenthetical statement inserted in the text

In case (1), an article often contains multiple paragraphs, each defining a distinct legal term. These definitions employ

formulaic expressions linking the term and its explanation, as illustrated by the underlined examples in Fig. 1. The English translations shown there are based on the Japanese Law Translation Database System (JLT)¹. Some paragraphs further include itemized conditions or additional defined terms, as shown in Fig. 2, where subordinate indices such as (i), (ii), etc., are used instead of higher-level ones like (1) and (2).

In case (2), a term defined in parentheses follows a phrase within the main text, serving as its explanatory note. Abbreviations are often introduced in this way. For instance, in Fig. 1, the double-underlined expression “specified gas generating facility” defines the term as a simplified gas generating facility specified by Cabinet Order.

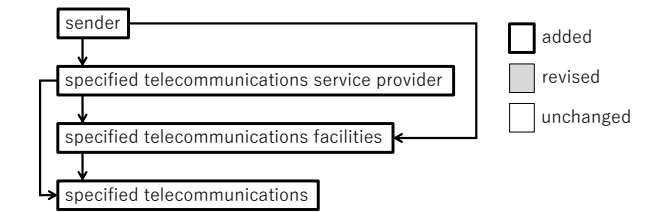
B. Diachronic Changes in Legal Terms

We cite the Act on Measures for Addressing Rights Infringements Arising from the Circulation of Information via Specified Telecommunications (Act No. 137 of 2001)². Hereafter, we refer to this act as the Provider Liability Act (Act No. 137 of 2001), which we use as an example to explain diachronic changes in legal terms. As of July 2025, this act had been amended five times, two of which included revisions to terms and definitions in Article 2. Fig. 3 shows the diachronic changes in these terms at three time points. The bold boxes indicate newly added terms, the thin boxes indicate terms that remain unchanged, and the gray background indicates revised definitions:

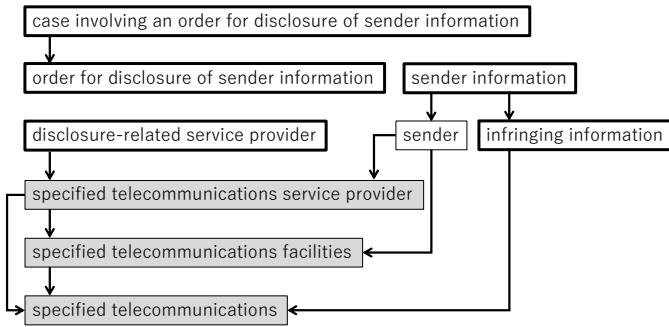
- (1) At the time of its original enactment (Fig. 3(a)), only four terms were defined: “specified telecommunications,” “specified telecommunications facilities,” “specified telecommunications service provider,” and “sender.” Among these, three terms, “specified telecommunications,” “specified telecommunications facilities,” and “specified telecommunications service provider,” were defined by referring to other terms that had already been established in the Act, while “sender” was defined independently. The arrows indicate these definitional dependencies.
- (2) Five terms were newly added, and some existing definitions were revised by the Act on the Partial Revision of the Act on Measures for Addressing Rights Infringements Arising from the Circulation of Information via Specified Telecommunications (Act No. 27 of 2021), which came into effect on October 1, 2022 (Fig. 3(b)). Note that, unlike changes in natural language, the meanings of legal terms are formally changed on the enforcement date.
- (3) The number of terms defined in the Provider Liability Act (Act No. 137 of 2001) increased to 14 with the Act on the Partial Revision of the Act on Measures for Addressing Rights Infringements Arising from the

¹<https://www.japaneselawtranslation.go.jp>

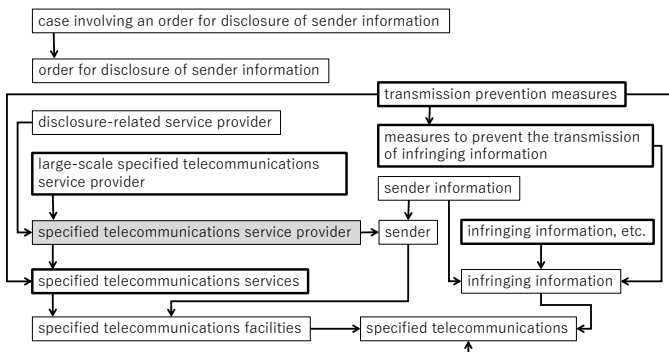
²The title of this law was amended from Act on the Limitation of Liability for Damages of Specified Telecommunications Service Providers and the Right to Demand Disclosure of Identification Information of the Sender (Act No. 137 of 2001) in 2024.



(a) Legal terms and relations in the Provider Liability Act (Act No. 137 of 2001)



(b) Legal terms and relations as of the enforcement of the Act on the Partial Revision of the Act on Measures for Addressing Rights Infringements Arising from the Circulation of Information via Specified Telecommunications (Act No. 27 of 2021)



(c) Legal terms and relations as of the enforcement of the Act on the Partial Revision of the Act on Measures for Addressing Rights Infringements Arising from the Circulation of Information via Specified Telecommunications (Act No. 25 of 2024)

Fig. 3. Diachronic changes in defined terms (Provider Liability Act)

Circulation of Information via Specified Telecommunications (Act No. 25 of 2024), which came into effect on April 1, 2025 (Fig. 3(c)). Some definitions were revised to express relationships among terms more concisely. For example, the newly defined term “specified telecommunications services” was introduced to simplify the definitions of “specified telecommunications service provider” and the newly added term “transmission prevention measures.”

III. JAPANESE LEGAL STANDARD XML SCHEMA (JLS)

In October 2016, the Ministry of Internal Affairs and Communications (MIC) launched the e-Legislative Affairs Work Support System (e-LAWS), which digitized the formerly paper-based process of drafting and publishing laws and



Fig. 4. Screenshot of e-Gov (in Japanese)

regulations. The system, now managed by the Digital Agency, is built on the Japanese Legal Standard XML Schema (JLS), an XML Schema specifically designed for Japanese legislation. XML Schemas define the structure of XML documents, specifying elements, attributes, and their hierarchy. They also provide a more flexible and XML-based alternative to DTDs.

JLS was developed through an analysis of existing laws and regulations to formalize their logical structure in XML. This made it possible to draft, review, and compile bills electronically, automating processes that were previously done on paper. Consequently, all legislative stages, from bill drafting to official publication, are now fully digital. The introduction of e-LAWS also shifted the publication format on e-Gov from HTML to XML, enhancing data reusability. Figure 4 shows the Provider Liability Act (Act No. 137 of 2001) on e-Gov,³ and Figure 5 presents its XML representation.

Structuring legal data with JLS separates content from presentation. For example, law titles and article headings are marked with the `LawTitle` and `ArticleCaption` tags, while formatting styles are defined independently. This separation improves machine readability, facilitates automated processing, and enhances transparency in legislative drafting.

Before e-LAWS, the Japanese Law Translation Database System (JLT)⁴ was the first system to employ XML for Japanese laws [Toyama et al., 2011], [Toyama et al., 2012]. JLT provides free English translations of major statutes, a bilingual dictionary of legal terms, and related translation resources. Its database stores bilingual legal texts in XML format, defined by a document type specification containing

³<https://elaws.e-gov.go.jp> (in Japanese)

⁴<https://www.japaneselawtranslation.go.jp>

```

<Law Era="Heisei" Lang="ja" LawType="Act" Num="137"
Year="13" PromulgateMonth="11" PromulgateDay="30">
<LawNum>Act No. 137 of November 30, 2001</LawNum>
<LawTitle>Act on the Limitation of Liability for Damages...</LawTitle>
<TOC>
<TOCLabel>Table of Contents</TOCLabel>
<TOCChapter Num="1">
<ChapterTitle>Chapter 1 General Provisions</ChapterTitle>
<ArticleRange>(Article 1, Article 2)</ArticleRange>
</TOCChapter>
[...]
```

Fig. 5. Example of JLS (translated in English)

TABLE I
JAPANESE LEGAL XML

Acronym	Release	Purpose	XML	Akoma Ntoso converter
JLT	2009	Japanese law translation database system	DTD	[Kawachi et al., 2016]
e-LAWS	2016	Japanese e-legislation system	JLS	[Nakamura, 2022]

103 element and 75 attribute declarations. Table I outlines the evolution of Japanese legal XML, showing how JLS succeeded JLT and expanded the definitions to 134 elements and 121 attributes after a comprehensive structural analysis.

Through e-Gov, users can now access laws promulgated since April 1, 2016, including those not yet in force. Although this study focuses on the definition provisions in Japanese statutes, many jurisdictions now publish their legislation in XML. With the adoption of international standards such as Akoma Ntoso [Palmirani and Vitali, 2011], processing methods can be unified, and Japanese legislative data can be converted accordingly [Kawachi et al., 2016], [Nakamura, 2022]. This compatibility opens possibilities for future comparative research on the evolution of legal terminology across different legal systems.

IV. LEGAL TERMS IN LANGUAGE EVOLUTION

In this section, we explain our work from the viewpoint of language evolution.

A. Record of Language Change

When we consider language as something that is constantly changing, it becomes clear that in any language, new words are continually created and eventually disappear, and sometimes such changes can spread throughout the entire speech community. These phenomena are most likely to occur in environments where words are invented and discarded rapidly, such as in the spoken language of young people or in internet slang [Crystal, 2008], [Crystal, 2011]. In recent years,

trends on platforms such as X (formerly Twitter) have likely accelerated these cycles even further.

In contrast, legal language can be regarded as one of the most conservative domains in this respect. One of the advantages of exploring legal documents from the perspective of language evolution is that the trajectory of change is thoroughly described and preserved in written form. Spoken language is often said to leave no fossils [Bates, 1961], which makes legal documents, where even the meanings of terms are explicitly defined, a valuable linguistic resource.

For example, pidgins often emerge spontaneously for practical reasons such as trade between different linguistic groups. They may disappear once they are no longer needed, or in some cases become stabilized and develop into creoles as native languages [Todd, 1974]. The pidgins known today are only those that happened to be recorded by chance [Sebba, 1997].

In other words, one of the main difficulties in studying language evolution has been the lack of records for spoken language, whereas linguistic phenomena that are documented are much easier to investigate.

B. Variation and Selection of Legal Terms

The diachronic evolution of terminology can be viewed as a unified field of study. It investigates the historical formation of a domain's terms and their subsequent development to the present [Kageura, 2002]. In the legal domain, such analysis is feasible because comprehensive sets of legal terms can be extracted from statutory corpora covering all enacted acts.

From the perspective of language evolution, the emergence and adoption of legal terms can be interpreted as a process of variation and selection. New words often appear spontaneously in society, sometimes as a result of social or technological innovation. When such words are widely accepted and acquire stable usage, they may eventually be incorporated as legal terms, where their meanings are explicitly defined and codified. In this sense, legal terminology reflects a selective process, in which linguistic variants compete and only those aligned with social and institutional needs survive to be formalized.

This approach also allows the synchronic comparison of legal terms to uncover latent interrelations among acts. Bills are often proposed in response to social changes such as economic or political shifts, wars, and natural disasters, which may lead to the simultaneous enactment of multiple acts. Although similar terms are defined across acts with comparable explanations, identifying such links is challenging without prior knowledge. Using natural language processing techniques, similarities between terms and their definitions can be quantified. Hence, diachronic legal terminology provides a framework for analyzing the dynamics of legal change.

V. EXTRACTION OF DEFINED TERMS

Thanks to XML-formatted open data, it has become possible to automatically extract data and to handle diachronic information. For this study, all laws available from e-Gov as of July

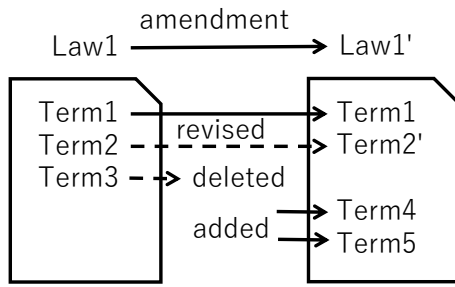


Fig. 6. Change in defined terms by amendment

2025 were downloaded and analyzed, including laws that had been promulgated but had not yet come into force. Among these, 2,107 laws were selected for analysis. Only those laws with articles whose headings included the word “definitions” were collected, and definition provisions were extracted.

Defined terms were extracted from the versions of laws that were in effect as of April 1, 2016. For laws enacted after that date, all changes were tracked from their enactment onward. For laws already in force as of that date, changes in defined terms were tracked only from that point forward.

In this study, we focused on terms defined in dedicated provisions established specifically for the purpose of providing definitions. Although some definitions appear within the main text of provisions using parentheses, such cases were excluded. Among the definition provisions, some define terms by paragraph, some by item (subparagraph), and some use a mixture of both; only those that could be extracted using regular expressions were included. In cases where a single paragraph or item defines multiple terms in parallel, these were counted as a single definition.

The legal texts were collected for each amendment, and from those, only the definition provisions were extracted. Edit distance was then calculated for each extracted definition. When the edit distance was one or greater, the corresponding defined term was considered to have changed and was tracked accordingly.

Even within definition provisions, the styles of expression vary widely. Since the purpose of this study is to track the diachronic changes of defined terms, overly strict extraction rules were not applied.

Fig. 6 illustrates how defined terms change through legal amendments. Law1 is amended to become Law1'. Term1 remains unchanged. Term2 retains the same term label, but its definition is altered. Term3 is deleted, and Term4 and Term5 are added. As a result, the total number of defined terms in the amended Law1' is four; two terms were added, one term was deleted, and one term was revised.

Based on this, the following values were calculated:

- The mean amendment frequency
- The number of new terms added, deleted, or revised in a single amendment

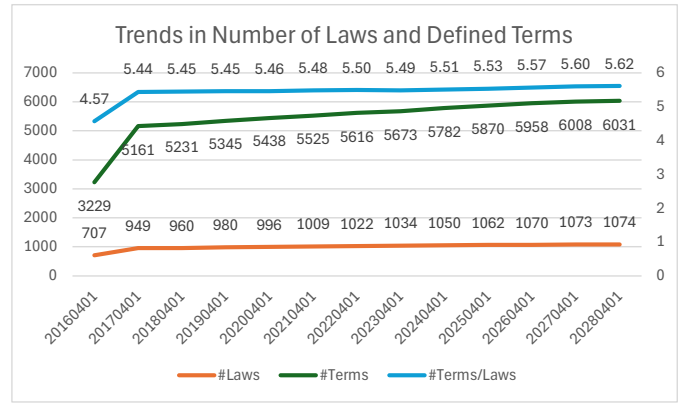


Fig. 7. Trends in numbers of laws and defined terms

TABLE II
STATISTICAL INFORMATION ON CHANGES IN LEGAL TERMINOLOGY

Amendment Type		Added Terms	Deleted Terms	Revised Terms
All amendments (including zero changes)	μ (σ^2)	1.123 (2.582)	0.430 (1.280)	1.201 (1.197)
Amendments with changes only	μ (σ^2)	2.737 (3.443)	1.990 (2.125)	1.536 (1.148)

VI. ANALYSIS

Among the laws containing defined terms, 1,076 were amended at least once between April 1, 2016 and July 2025. The total number of amendments was 1,226, with an average of 1.139 amendments per law. Of these, 690 amendments involved changes to definition provisions, with an average of 0.641 per law.

Fig. 7 shows the number of laws containing definition provisions, the total number of defined terms, and the average number of defined terms per law for each fiscal year. The values for the first year are relatively low because the dataset includes only laws that were amended on or after April 1, 2016. Thus, the data for that year reflect only the laws that came into effect on that specific date.

Statistical information on changes in legal terminology due to amendments is presented in Table II. The table shows the mean and variance of the number of terms added, deleted, and revised in a single amendment. For each category, the upper row represents the results including amendments in which the number of changes was zero, while the lower row shows the results excluding such cases.

Although fewer than ten years have passed, which may be too short a period to fully observe linguistic change, this framework has made it possible to observe real-world language evolution as it occurs. In the future, expanding the dataset to cover a longer historical period would allow for the analysis of linguistic change on a larger time scale.

VII. CONCLUSION

In this study, we proposed a framework for analyzing the diachronic evolution of legal terminology using Japan's

open legislative data. By extracting defined terms from XML-formatted statutes provided through e-Gov and tracking changes across amendments, we demonstrated that it is possible to observe and quantify how legal definitions evolve over time.

Although the period covered is currently less than ten years, our method has enabled empirical analysis of linguistic change in legal documents, which has traditionally been difficult due to the lack of longitudinal data. The ability to compute edit distances and track additions, deletions, and revisions of defined terms opens new avenues for studying the dynamics of legal language.

Furthermore, because Japanese legal data can be converted into international formats such as Akoma Ntoso, this framework is potentially extensible to comparative studies of legal language across jurisdictions. With extended historical datasets, this approach is expected to support the analysis of long-term trends in the evolution of legal terminology.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP24K03230.

REFERENCES

- [Bates, 1961] Bates, M. (1961). *Man in Nature*. Prentice Hall, NJ.
- [Cangelosi and Parisi, 2002] Cangelosi, A. and Parisi, D., editors (2002). *Simulating the Evolution of Language*. Springer, London.
- [Crystal, 2008] Crystal, D. (2008). *Txtng: The Gr8 Db8*. OUP Oxford.
- [Crystal, 2011] Crystal, D. (2011). *Internet Linguistics: A Student Guide*. Routledge.
- [Höfler et al., 2011] Höfler, S., Bünzli, A., and Sugisaki, K. (2011). Detecting Legal Definitions for Automated Style Checking in Draft Laws. Technical Report CL-2011.01, University of Zurich, Institute of Computational Linguistics.
- [Hurford, 2002] Hurford, J. R. (2002). The Roles of Expression and Representation in Language Evolution. In *The Transition to Language*, pages 311–334. Oxford University Press, Cambridge.
- [Kageura, 2002] Kageura, K. (2002). *The Dynamics of Terminology*. John Benjamins.
- [Kawachi et al., 2016] Kawachi, G., Nagai, A., Nakamura, M., Ogawa, Y., Ohno, T., and Toyama, K. (2016). Applying the Akoma Ntoso XML Schema to Japanese Legislation. *Journal of Law, Information and Science*, 24(2):49–102.
- [Lame, 2005] Lame, G. (2005). Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations. In *Law and the Semantic Web*, LNAI3369, pages 169–184. Springer.
- [Lyon et al., 2007] Lyon, C., Nehaniv, C., and Cangelosi, A., editors (2007). *Emergence of Communication and Language*. Springer.
- [Nakamura, 2022] Nakamura, M. (2022). Development of applications for open data for Japanese laws and regulations. *Journal of Open Access to Law*, 10(1):16 pages.
- [Nakamura et al., 2017] Nakamura, M., Hayashi, Y., and Matoba, R. (2017). Simulation of Language Evolution based on Actual Diachronic Change Extracted from Legal Terminology. In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017)*, pages 343–350.
- [Nakamura et al., 2016] Nakamura, M., Ogawa, Y., and Toyama, K. (2016). Development of Diachronic Terminology from Japanese Statutory Corpora. *Journal of Open Access to Law*, 4(1):16 pages.
- [Palmirani and Vitali, 2011] Palmirani, M. and Vitali, F. (2011). *Akoma-Ntoso for Legal Documents*, pages 75–100. Springer Netherlands, Dordrecht.
- [Sebba, 1997] Sebba, M. (1997). *Contact Languages: Pidgins and Creoles*. Modern Linguistics. Palgrave Macmillan.
- [Todd, 1974] Todd, L. (1974). *Pidgins and Creoles*. Routledge & Kegan Paul Ltd, London.
- [Toyama et al., 2011] Toyama, K., Saito, D., Sekine, Y., Ogawa, Y., Kakuta, T., Kimura, T., and Matsuura, Y. (2011). Design and Development of Japanese Law Translation Database System. In *Proceedings of Law via the Internet*, page 12 pages.
- [Toyama et al., 2012] Toyama, K., Saito, D., Sekine, Y., Ogawa, Y., Kakuta, T., Kimura, T., and Matsuura, Y. (2012). Design and Development of the Japanese Law Translation Database System (in Japanese). *Journal of Information Network Law Review*, 11:33–53.
- [Winkels and Hoekstra, 2012] Winkels, R. and Hoekstra, R. (2012). Automatic Extraction of Legal Concepts and Definitions. In *Legal Knowledge and Information Systems - JURIX 2012: The Twenty-Fifth Annual Conference*, pages 157–166.

Analysis of Section-Wise Syllable Usage and Its Relationship to Instrumentation in J-POP A Cappella Arrangements

Kohei Tsuchida

Dept. of Information Science
Tokyo Univ. of Science
Japan
6322061@ed.tus.ac.jp

Shun Sawada

Dept. of Information Technology
and Media Design
Nippon Institute of Technology
Japan
sawada@rs.tus.ac.jp

Kouichi Katsurada

Dept. of Information Science
Tokyo Univ. of Science
Japan
katsurada@rs.tus.ac.jp

Hidefumi Ohmura

Dept. of Information Science
Tokyo Univ. of Science
Japan
ohmura@is.noda.tus.ac.jp

Abstract—In a cappella arranging, “syllables” refers to non-lexical lyrics for chorus parts. Syllables are a key factor shaping a piece’s mood and listening comfort. At present, syllable choice in a cappella arranging is largely guided by rule of thumb, which poses a barrier for beginners. To lower this barrier, this study provides guidelines for syllable selection in a cappella arranging. These guidelines are derived by clarifying our hypotheses that (i) particular syllables are used depending on musical structure and (ii) particular syllables are used depending on the instruments. Specifically, we analyze 100 J-POP songs arranged for a cappella and published on YouTube, examining section-wise tendencies in syllable usage (intro, A-verse, B-verse, chorus) and the relationship between instruments used in the original tracks and syllables. The results reveal consistent regularities between sections and syllables, and partially confirm relationships between instruments and syllables. We also implement an automatic syllable-assignment program based on these findings.

Index Terms—a cappella, syllables, lyric assignment, J-POP, statistical analysis, qualitative evaluation

I. INTRODUCTION

In recent years, the popularity of a cappella has grown in the music industry. In this paper, we study a cappella—vocal music without instrumental accompaniment—using the following four parts:

- Lead: sings the main melody
- Chorus: provides harmonic support
- Bass: supports harmony and rhythm with the lowest tones
- Vocal percussion: creates rhythm

Recent media attention and accessibility have boosted public interest in a cappella. On the other hand, when creating a cappella arrangements, it is necessary to decide on non-lexical syllables for the chorus parts (hereafter, “syllables”). However, this choice often relies on empirical rules of thumb rather than systematic guidelines, which can be particularly burdensome for beginners. Moreover, there is limited research on a cappella in general and almost no studies focusing specifically on syllable usage.

Therefore, clarifying trends in syllable selection could prove useful for streamlining the a cappella arranging process. The objective of this study is to provide guidelines for syllable selection in a cappella arrangements. Specifically, we will empirically test two hypotheses: (1) that syllables are used according to musical structure (e.g., verse, chorus; see Section IV), and (2) that certain syllables correspond to the instruments used in the original composition (see Section V). Based on these results, we will construct rules for syllable selection. Through this work, we aim to develop a system that supports even a cappella arranging novices in making systematic, rule-based decisions, as well as to apply these findings to an automatic a cappella arrangement system (see Section VI).

II. RELATED WORK

In this section, we introduce previous research on a cappella arrangement and on syllables.

A. Scat Generation Research

Like syllables, scat refers to an improvised vocal technique without fixed lyrics, and it plays an important role in jazz and a cappella arrangements. However, there is a distinction between scat and syllables. While a syllable denotes a single utterance such as “oo” or “tu,” scat refers to a phrase combining syllables—for example, “shubidu-boo.” Nonetheless, these terms are sometimes used interchangeably. There has been research on generating scat using the high-quality analysis-conversion-synthesis system STRAIGHT [1]–[3]. In that work [4], acoustic adjustments were performed to produce natural scat vocalizations, but the choice of articulation—that is, which consonants and vowels to select and sing—had to be made manually by the user based on the melody’s tempo and rhythm. Prior work improved acoustic realism, but consonant-vowel choice remained manual, motivating rule-based automation.

B. A Cappella Vocal Ensemble Corpus

The jaCappella Corpus [5] provides ensemble and isolated part recordings of Japanese a cappella arrangements

TABLE I
CORRESPONDENCE BETWEEN SYLLABLE NOTATION AND
PRONUNCIATION

Score notation	IPA notation	Score notation	IPA notation
lu	[ru]	ta	[ta]
uh	[u]	ba	[ba]
fon	[fō]	wa	[wa]
fu	[fū]	lun	[rūn]
tu	[tū]	lan	[ran]
oh	[o]	ya	[ja]
ha	[ha]	won	[wō]
te	[te]	tan	[tan]
shu	[cū]	shan	[can]
bi	[bi]	han	[han]
du	[dū]	e	[e]
tun	[tūn]	pon	[pon]
ah	[a]	cha	[tca]
da	[da]	wan	[wan]
la	[ra]	i	[i]
pe	[pe]	wi	[wi]
pa	[pa]	chu	[tcu]
yun	[jūn]	fa	[fa]
yeah	[je]	li	[ri]
dan	[dan]	ti	[ti]
wo	[wo]	di	[di]
wow	[wau]	gi	[gi]

of 50 public-domain children’s songs, and has mainly been used for source separation. It also reports distributions of lexical vs. non-lexical syllables.

III. QUESTIONNAIRE SURVEY

A. Purpose

At present, there are very few publicly available datasets focused on syllable selection in a cappella arranging. Therefore, this study required collecting original data. In particular, by identifying which reference materials actual arrangers consult when creating a cappella arrangements, we can more accurately determine the appropriate targets for our analysis.

B. Survey Method

From May 1 to May 31, 2025, we invited 94 university a cappella clubs across Japan to participate in an online questionnaire. The single question was: “When arranging a cappella, which of the following resources do you refer to? (multiple answers allowed)” The six answer options were:

- A cappella videos on YouTube, etc.
- Online arrangement tutorial articles, etc.
- Textbooks on a cappella arranging, etc.
- Commercial scores available online or in music stores, etc.
- Advice from senior members, etc.
- Other

C. Results

Out of 334 total respondents, we analyzed the 176 individuals (52.7%) who had actual experience arranging a cappella. The resource most frequently cited was “a cappella videos on YouTube,” with 142 respondents (80.7%) indicating they used it. The breakdown for all options is shown below:

- A cappella videos on YouTube: 142 respondents (80.7%)
- Online arrangement tutorial articles: 41 respondents (23.3%)
- Textbooks on a cappella arranging: 18 respondents (10.2%)
- Commercial scores available online or in stores: 28 respondents (15.9%)
- Advice from senior members: 107 respondents (60.8%)
- Other: 40 respondents (22.7%)

In the “Other” category, respondents mentioned items such as “I refer directly to the original song” and “I also consult piano and Electone arrangement videos.” Taken together, these responses indicate that YouTube functions as the primary medium for discovering and sharing a cappella arrangements in our target community; accordingly, we focused data collection on publicly available YouTube videos. Moreover, the fact that 60.8% rely on “advice from senior members” highlights the absence of codified guidelines and the prevalence of tacit, empirical know-how. Given that J-POP arrangements dominate the Japanese a cappella scene we surveyed, we restricted the present corpus to J-POP; nevertheless, from a global perspective (e.g., US/UK collegiate scenes), Western pop and other repertoires are widely performed. Future work will therefore extend the corpus beyond J-POP/Japanese to test whether the section-wise tendencies and our rule-based priors generalize across languages, genres, and platforms.

IV. ANALYSIS OF SYLLABLES BY MUSICAL SECTION

A. Analysis Method

We divide each song into five sections—pre-melody, intro, A-verse, B-verse, and chorus—where “pre-melody” denotes any lyrical segment before the intro. To address orthographic variation (e.g., the same /u:/ sound written as “oo,” “uh,” or “woo”), we normalized all spellings to a canonical form according to the rules in TABLE I (notation $\leftrightarrow$ IPA mapping). To emphasize diversity rather than repetition, we counted unique syllables per section (each syllable counted at most once per section). Using the cleaned data, we analyze characteristic patterns of syllable usage by section.

B. Data Source

We analyzed a dataset of a cappella arrangements publicly available on YouTube, selecting 100 J-POP songs. Because our goal was to capture broad trends in syllable selection, we did not filter by performance skill or view count. To improve annotation accuracy, we prioritized recordings with clear enunciation and videos that showed mouth movements. This yielded a dataset well suited for phoneme identification.

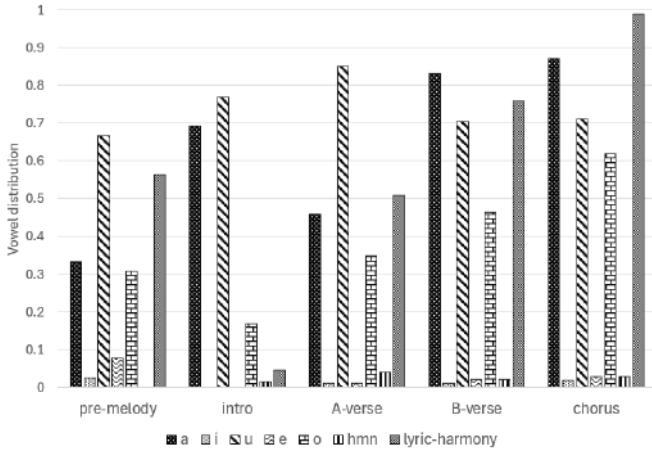


Fig. 1. Vowel distribution of syllables by musical section

C. Results

Fig. 1 plots, for each section (pre-melody, intro, A-verse, B-verse, chorus), the number of distinct syllables used per song. Vowel symbols are rendered as “a,” “i,” “u,” “e,” and “o.” The label “hmn” denotes humming, and “lyric-harmony” indicates passages where the harmony part sings the original lyrics.

- A-verses favor ‘u’, whereas B-verses and choruses shift toward “a”/ “o”.
- Choruses almost always include lyric harmony.
- Throughout all sections, syllables with “i” or “e” vowels are rarely used.

D. Discussion

The observed pattern—preference for “u”-vowel syllables in A-verses and increased “a” and “o” usage in choruses—appears to reflect intentional syllable selection in response to the song’s energy flow. Existing a cappella arranging practice holds that “u” offers a lighter, floating timbre suited to verses, “o” functions well as a connective vowel, and “a,” with its wide-open mouth shape, is ideal for the climactic chorus [6]. Our empirical results align with these established heuristics. Further, Amemiya et al. [7] report that, excluding “i” and “e,” the perceived brightness of single-vowel sounds ranks “a” > “o” > “u.” This ordering corresponds with our finding of heavier “a” and “o” usage in high-energy sections. The near-absence of “i” and “e” may be explained by effects of the singer’s formant (SF)—an acoustic phenomenon in which trained singers concentrate energy around 3 kHz in the third-to-fifth formant region, enhancing vocal clarity against accompaniment. Yamashita et al. [8] show that front vowels (“i,” “e”) and other vowels (“a,” “u,” “o”) exhibit different SF characteristics, suggesting that “i” and “e” may be less favorable in a cappella contexts. To test statistically whether syllable usage biases differ by section,

TABLE II
SYLLABLE USAGE FREQUENCY BY MUSICAL SECTION

	a	u	o	Total
Pre-melody	19	43	13	75
Intro	89	90	11	190
A-verse	79	156	37	272
B-verse	164	104	47	315
Chorus	158	113	64	335
Total	509	506	172	1187

we performed a chi-square test of independence:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count and E_i the expected count. We adjusted the data to ensure no more than 20 expected counts below 5 and that none fell below 1, removing “i”- and “e”-vowel categories due to their extreme sparsity. Although excluding these vowels risks overlooking their potential roles—and larger sample sizes or alternative methods (e.g., Fisher’s exact test or multinomial logistic regression) could mitigate this—in this study we prioritized satisfying chi-square assumptions. Consequently, we analyzed only the three vowel categories “a,” “u,” and “o.” This focus aligns with the primary vowels used in arranging and clarifies interpretation. The test on the cross-tabulation in TABLE II revealed a significant association between musical section and syllable- vowel category ($\chi^2(8) = 73.0, p = 1.24 \times 10^{-12}, p < 0.01$). This demonstrates that syllable choice is non-random and follows distinct statistical tendencies according to song structure.

V. ANALYSIS OF THE RELATIONSHIP BETWEEN INSTRUMENTAL SOUNDS AND SYLLABLES

A. Analysis Method

In this study, we focused exclusively on the instrumental sections (i.e., passages without vocals) of the original songs and analyzed the syllables assigned to the part carrying the melody. By isolating these sections, we aimed to eliminate the influence of song structure (intro, A-verse, B-verse, chorus, etc.) and reveal the effect of the instrument’s timbre itself on syllable choice. Instruments were classified into the following categories:

- Piano (Pf.)
- Strings (Str.)
- Electric Guitar (E. Gt.)
- Acoustic Guitar (A. Gt.)
- Glockenspiel (Glo.)
- Brass (B. W.)
- Synthesizer (Syn.)

For each instrument category, we identified which syllables were most frequently assigned and summarized the usage trends per instrument. To quantitatively evaluate the relationship between instrument timbre and syllable charac-

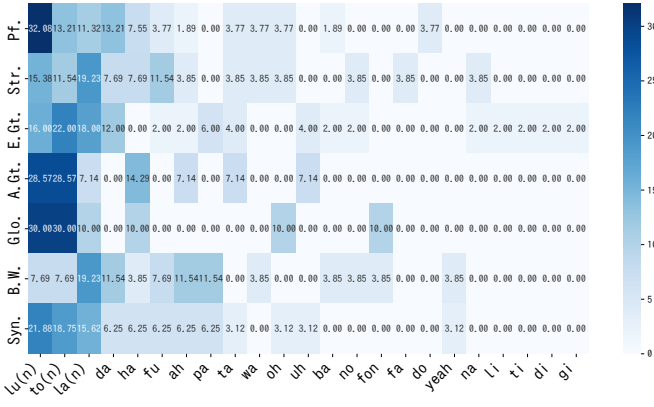


Fig. 2. Instrument-conditioned phoneme probability matrix

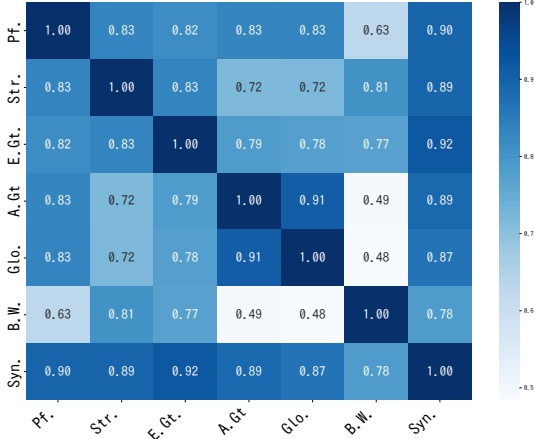


Fig. 3. Cosine similarity matrix of phoneme distributions between instruments

teristics, we computed cosine similarity. Cosine similarity is defined as:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

When measuring similarity between two instruments, A and B , we construct a vector x_A whose elements are the counts of each syllable used with instrument A , and similarly x_B , for instrument B , then compute $\cos(x_A, x_B)$. To assess similarity between two syllables, A and B , we build vectors y_A and y_B , where each element represents the usage count of that syllable with each instrument, and compute $\cos(y_A, y_B)$. This analysis clarified which instruments tend to share similar syllable assignments and how syllable usage patterns differ across timbres.

B. Data Source

The dataset used here is the same as in Section III (“Analysis of Syllables by Musical Section”): 100 J-POP a cappella arrangements publicly available on YouTube.

C. Results

Fig. 2 presents a phoneme probability matrix showing, for each instrument, the probability of assignment to

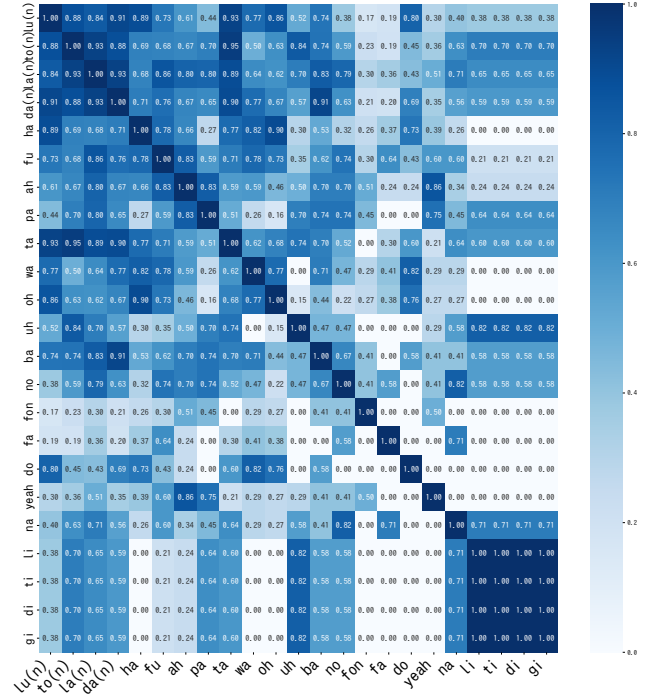


Fig. 4. Cosine similarity matrix between syllables

each phoneme. Fig. 3 shows cosine similarities between instruments, and Fig. 4 shows cosine similarities between syllables. From the phoneme probability matrix in Fig. 2, we observe that, irrespective of instrument, the syllables “lu” and “to” tend to be used most frequently. These syllables also often appear in combination with other syllables sharing the same vowel—e.g., “luto,” “dala.” Next, the inter-instrument cosine similarity matrix in Fig. 3 indicates that the cosine similarity values between brass (B.W.) and other instruments are generally low. By contrast, the inter-syllable cosine similarity matrix in Fig. 4 shows high similarity among “lu, to, la, da, ta.” In addition, “li, ti, di, gi” exhibit exactly the same high cosine similarity; this is because each of these syllables appeared only once, and in every case was assigned to the electric guitar.

D. Discussion

First, the frequent use of “lu” and “to” regardless of instrument likely relates to articulatory “ease of utterance.” For instance, “ha” and “fu” require relatively more breath, making continuous articulation difficult, whereas “lu” and “to” can be produced smoothly even in fast tempi and thus adapt readily to a wide range of instrumental timbres. The frequent appearance of combinations such as “luto” and “dala” can be interpreted as a strategy for handling the speed of the original instrumental lines. This phenomenon resembles the concept of double tonguing in brass and woodwind performance, where alternating different consonants facilitates the execution of rapid passages that are difficult to realize with a single articulation. singer’s formant (SF) may also influence how syllables

TABLE III
VOWEL FREQUENCY OF SYLLABLES ASSIGNED TO INSTRUMENTS

	a	u	Other	Total
Pf.	23	28	2	53
Str.	14	10	6	30
E. Gt.	23	22	5	50
B. W.	17	6	3	26
Syn.	14	16	2	32
Total	91	82	18	191

are combined. Stabilizing SF is considered conducive to producing an ideal singing tone. According to Takahashi et al. [9], no significant difference is observed between students and professionals for pitch changes within the same vowel, whereas significant differences are found for pitch changes across different vowels. In light of this, when combining syllables, there may be a tendency to pair same-vowel syllables rather than cross-vowel ones, as an adaptation to maintain vocal stability and smoothness. The generally low cosine similarities for brass may be related to the fact—evident in the phoneme probability matrix in Fig. 2—that “pa” is often assigned to brass timbres. Because “pa” is frequently used as an accent and thus functions differently from ordinary passagework, this usage likely reduces the similarity between brass and other instruments. However, this does not necessarily mean that “pa” behaves uniquely in terms of cosine similarity. While “pa” and “da” indeed have powerful attacks and are often used as accents, they are typically employed in combination with other syllables (e.g., “pala,” “dala”) rather than in isolation, which likely contributes to the observed patterns. Overall, although syllables are sometimes said to be chosen to imitate instrumental sounds, our results suggest that “ease of articulation” is often prioritized. In practice, syllable selection appears to consider not only timbral imitation but also singability and the flow of the musical phrase. To statistically test relationships between instrumental timbre and vowel/consonant categories, we conducted chi-square tests. We first examined whether the vowels of syllables used with each instrument were related to instrument type. Because excessively fine vowel categories would violate chi-square assumptions, we grouped “i, e, o”—whose frequencies were extremely low compared with “a, u”—into a single “other” category. In addition, because acoustic guitar and glockenspiel each appeared only five times (insufficient data), we excluded these two instruments from the test. Using the cross-tabulation in TABLE III, we performed a chi-square test. The result did not indicate a significant association between instrument timbre and syllable vowels ($\chi^2(8) = 12.5, p = 0.129, p > 0.01$). Note, however, that merging “o” with “i, e” may obscure “o”-specific tendencies. We then examined whether the consonants of syllables used with each instrument were related to instrument type. A direct test would have produced overly fine-grained categories due to the large number of consonant types, so we consolidated consonants

TABLE IV
CONSONANT FREQUENCY OF SYLLABLES ASSIGNED TO INSTRUMENTS

	Sonorants	Obstruents	Total
Pf.	28	25	53
Str.	14	12	26
E. Gt.	23	27	50
A. Gt.	7	7	14
Glo.	5	5	10
B. W.	13	13	26
Syn.	17	15	32
Total	107	104	211

into two categories: obstruents (e.g., stops, fricatives) and sonorants (e.g., nasals, liquids). Using the cross-tabulation in TABLE IV, we performed a chi-square test. Again, no significant association was found between instrument timbre and syllable consonant category ($\chi^2(6) = 0.726, p = 0.994, p > 0.01$). This outcome likely reflects the necessity of coarse category grouping. With a larger dataset, subdividing consonants into finer classes—stops, fricatives, nasals, liquids, etc.—could yield additional insights.

However, these tests are underpowered: several instrument classes (e.g., A. Gt., Glo.) occur infrequently and many cells are sparse, so we were forced to collapse categories and to exclude rare vowels to satisfy χ^2 assumptions. Therefore, we will both expand the corpus and adopt methods that cope with sparse counts—(multinomial) logistic or log-linear regression with regularization and mixed effects (song/arrangement as random factors), and exact/Fisher-type tests where appropriate—so that instrument effects can be evaluated without aggressive merging.

VI. AUTOMATIC SYLLABLE ASSIGNMENT PROGRAM

A. Design Policy

We consolidate the rule design (Sections IV–V) and the four assignment patterns into a single procedure (Fig. 5). First, section-weighted sampling reflects the empirical vowel mix a,u,o per section; syllable-level priors prioritize fast-articulable items lu,to,la,da,ta and down-weight breathy ha,fu. Then, per part: (i) Lead—preserve given lyrics; fill only previous blank notes by the sampler. (ii) Phrase copying—in the lead, detect runs of ≥ 3 notes that share onset grid and duration; copy their syllables to chorus parts to ensure cross-part coherence. (iii) Part 1 continuity—for notes not fixed by copying, if the immediately previous note has the same duration, inherit its syllable; otherwise sample as in (i). (iv) Parts 2–3—apply a priority map (inherit from Part 1 / Part 2 where defined), then duration-based inheritance, and finally sampling. We use a fixed random seed for reproducibility. This combination of section statistics, phrase-level copying, and duration continuity mirrors the observed tendencies and guarantees full coverage (no unassigned notes).

B. Functional Verification

We loaded a test input score into the program and opened the resulting MusicXML in a score viewer. The

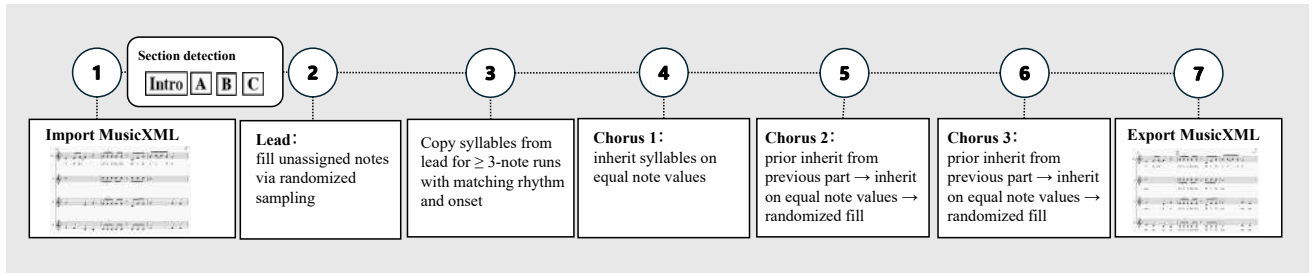


Fig. 5. Flowchart of the automatic syllable assignment program



Fig. 6. Output score produced by the program

lead part’s existing lyrics were preserved as-is, and syllables were reliably assigned to every previously blank note in all other parts. Notably, (i) no unassigned notes remained in any part, and (ii) the section-dependent vowel mix was visually apparent—both indicating that the processing steps defined in Section VI.A were implemented correctly. Fig. 6 shows an excerpt of the output score.

C. Planned Evaluation

We will run an A/B listener study with members of the university a cappella clubs who answered our questionnaire, comparing our rule-based outputs against a human-arranged reference (original arrangement) using identical scores in which only the chorus syllables differ; stimuli will be loudness-normalized and presentation order counterbalanced. For a small subset we will also include an ABX check and an auxiliary uniform-sampling baseline to confirm that effects are not explained by chance vowel choices. Participants will provide a forced-choice preference and brief 7-point ratings of naturalness and singability; results will be reported in follow-up work and used to tune weighting parameters and improve section detection.

VII. CONCLUSION

We analyzed how syllable choice in J-POP a cappella relates to musical structure and instrumentation. A-verses favored “u,” B-verses/choruses shifted toward “a/o,” and “i/e” were rare overall; selection appears driven not only by instrumental imitation but also by articulatory ease for rapid phrases. Building on these findings, we implemented a rule-based program that preserves lead

lyrics, enforces phrase copying and duration continuity, and fills remaining notes via section-weighted sampling; functional checks confirmed full coverage with section-dependent vowel mixes. Looking ahead, we will expand the dataset, model tempo and note-value effects, and apply finer statistical methods. In parallel, qualitative evaluations with university a cappella club members (naturalness/singability) and quantitative metrics (vowel ratios, phrase-copy success) will guide weighting-parameter tuning, improve section detection, and enhance overall usability.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 25H01169, 23K00243.

REFERENCES

- [1] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Munich, Germany, 1997, vol. 2, pp. 1303–1306.
- [2] H. Kawahara, I. Matsuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,” *Speech Commun.*, vol. 27, nos. 3–4, pp. 187–207, 1999.
- [3] H. Kawahara, “STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” *Acoust. Sci. & Technol.*, vol. 27, no. 6, pp. 349–353, 2006, doi: 10.1250/ast.27.349.
- [4] H. Kawahara and H. Katayose, “Scat singing generation using a versatile speech manipulation system STRAIGHT,” *J. Acoust. Soc. Am.*, vol. 109, no. 5 (Suppl.), p. 2425, 2001, presented at the 141st Meeting of the Acoustical Society of America, Chicago, IL, USA, Jun. 4–8, 2001.
- [5] T. Nakamura, S. Takamichi, N. Tanji, S. Fukayama, and H. Saruwatari, “JaCappella corpus: A Japanese a cappella vocal ensemble corpus,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095569.
- [6] Assy, *New A Cappella Perfect Book*. Tokyo, Japan: Doremi Music Publishing, 2023. (in Japanese).
- [7] T. Amemiya and S. Mizutani, “On the basic affective dimensions of Japanese onomatopoeia and the basic level of Japanese phonesthemes,” *Bulletin of the Faculty of Sociology, Kansai University*, vol. 37, no. 2, pp. 139–166, 2006. (in Japanese).
- [8] Y. Yamashita, M. Kayama, K. Ikeda, S. Yoshida, K. Hirabayashi, K. Itoh, and K. Asanuma, “Vowel characteristics of singing acoustic features related to skill development of singing voice,” *IPSJ SIG Tech. Rep.*, vol. 2018-MUS-118, no. 26, 2018. (in Japanese).
- [9] J. Takahashi and M. Tsuzaki, “Stability of singer’s formant in singing—Spectral analysis of singing including the transition of vowel—,” *IPSJ SIG Tech. Rep.*, vol. 2016-MUS-113, no. 13, 2016. (in Japanese).

Personalized Tonal Harmony Parsing via Scoring-Based Tree Selection

Daichi Hoshikoshi

*Department of Information Sciences
Tokyo University of Science
Chiba, Japan
6325544@ed.tus.ac.jp*

Shun Sawada

*Department of Information Technology and Media Design
Nippon Institute of Technology
Saitama, Japan
sawada.shun@nit.ac.jp*

Kouichi Katsurada

*Department of Information Sciences
Tokyo University of Science
Chiba, Japan
katsurada@rs.tus.ac.jp*

Satoshi Tojo

*Department of Data Science
Asia University
Tokyo, Japan
tojo_satoshi@asia-u.ac.jp*

Hidefumi Ohmura

*Department of Information Sciences
Tokyo University of Science
Chiba, Japan
ohmura@is.noda.tus.ac.jp*

Abstract—This study addresses individual differences in the harmonic analysis of chord progressions. We hypothesize that these differences arise from personal analytical preferences and propose a personalized analysis system based on a Generative Syntax Model (GSM). For a given progression, the GSM generates all possible parse trees via context-free grammar rules and scores them along multiple axes—syntactic conciseness, tonal consistency, frequency of typical progressions, and key consistency. Users subjectively set the weightings of these axes to reflect their own preferences, enabling personalized tonal harmony parsing. Experiments under various weighting conditions confirm the system’s potential for personalized analysis.

Index Terms—chord progression, context-free grammar, generative syntax model, harmonic analysis, harmonic structure, music parsing

I. INTRODUCTION

Chord progressions in music are fundamental elements that form the structure and emotion of a piece of music. These chord progressions are thought to have a structure similar to the grammar of language, and in fact, numerous methods for analyzing chord progressions from a linguistic perspective have been proposed [1], [2]. However, practical implementations [3], [4] typically present a single analysis, reflecting a hypothesis that there exists an objective, unique harmonic interpretation. In practice, analysts often disagree—for example, when encountering a non-diatonic chord, some may interpret it as a modulation, others as modal interchange or a secondary dominant. These divergences stem from individual knowledge, experience, and preferences: an analyst might favor interpreting an entire piece in one key, emphasize tonal harmonic functions, or prioritize secondary dominants. This study is based on the hypothesis that such differences in analyses are deeply rooted in the analyst’s individual musical knowledge, experience, and “preferences”.

Given these considerations, the purpose of this study is to focus on the differences in chord progression analysis between individuals and propose a new analysis method that

can flexibly reflect the analytical tendencies and preferences of individual users. Our two-stage approach first exhaustively generates all possible analyses for a given progression using a Generative Syntax Model (GSM), which employs production rules to output hierarchical parse trees. In the second stage, we evaluate each candidate against user-defined criteria—number of nodes, tonal consistency (e.g., modulations), frequency of secondary dominants and borrowed chords, and conformity with a specified key—and compute a weighted score. Users adjust the weights to reflect their analytical preferences, allowing the system to select the most fitting analysis.

Specifically, we formalize chord progressions as a context-free grammar based on functional harmony [5], parse candidates via chart parser (Earley’s algorithm) [6], and score each tree according to harmonic features. This flexible framework empowers personalized tonal harmony parsing optimized for individual analysts.

II. RELATED WORK

This section reviews existing studies that syntactically approach harmonic progressions. In particular, we focus on methods that describe and analyze the hierarchical structure of chord progressions using formal grammar, introducing the representative model GSM. Additionally, we discuss the suitability of various parsing algorithms [6] for handling the production rules defined in GSM.

A. Generative Syntax Model (GSM)

GSM proposed by Rohrmeier is a framework that grammatically describes chord progressions in Western tonal music as hierarchical and functional structures. In this model, the regularities of chord progressions are formalized using a CFG, and the entire chord progressions are analyzed as a parse tree. The fundamental concept of GSM is that chords have syntactic dependencies with each other and form structural groupings as hierarchical units.

- $\mathbb{P} = \{\text{piece}, P\}$
- $\mathbb{K} = \{\text{Cmaj}, \text{Cmin}, \text{C}\sharp\text{maj}, \text{C}\sharp\text{min}, \text{Dmaj}, \dots\}$
- $\mathbb{R} = \{\text{TR}, \text{SR}, \text{DR}\}$
- $\mathbb{F} = \{t, s, d, tp, sp, dp, tcp\}$
- $\mathbb{S} = \{I, II, III, IV, V, VI, VII\}$
- $\mathbb{O} = \{\text{Cmaj}, \text{Cmin}, \text{C}^\circ, \text{C}^\circ, \dots\}$

Fig. 1. Symbol categories defined for GSM

1) Rules of GSM:

GSM hierarchically generates the harmonic functions of chords. Here, we explain the basic harmonic functions. Harmonic functions can be classified into tonic (T), dominant (D), and subdominant (S). A progression such as T–S–D–T constitutes the fundamental pattern. We now describe the symbol sets (Fig. 1) handled by the rules of GSM: $\mathbb{P}$ denotes the entire chord progression. $\mathbb{K}$ denotes the key for each symbol. $\mathbb{R}$ denotes harmonic functions at the regional level. $\mathbb{F}$ denotes more fine-grained harmonic functions including substitutes. $\mathbb{S}$ denotes scale degrees based on the key. $\mathbb{O}$ denotes the surface chords.

TR, SR, and DR denote the Tonic Region, Subdominant Region, and Dominant Region, respectively, and they divide the entire chord progression into large functional blocks. Each of these denotes a fundamental harmonic function: tonic, subdominant, and dominant. The symbols *tp*, *tcp*, *sp*, and *dp* denote Tonic Parallel, Tonic Counter Parallel, Subdominant Parallel, and Dominant Parallel, respectively, and refer to chords that can function as substitutes for each corresponding harmonic function.

We now describe the rules of GSM (Fig. 2) that are important in this study. Rules 17 and 19 concern secondary dominants. A secondary dominant is a chord that approaches another chord X by a descending perfect fifth, expressed as V/X (slash). While such descending perfect fifth approaches provide a strong sense of resolution, secondary dominants can involve chords that fall outside the diatonic scale, and thus secondary dominants can sometimes obscure the sense of tonality. Rule 18 describes chords that approach a chord X by a descending fifth within the diatonic scale, also expressed as V/X . Since the approach is not necessarily by a perfect fifth, it does not provide as strong a resolution as that of secondary dominants. However, because all chords remain within the diatonic scale, tonal stability is preserved. A progression in which Rule 18 appears repeatedly is referred to here as a diatonic descending fifth.

B. Parsing Method

This study employs the GSM to generate all possible parse trees for a given chord progression and select the analysis that best aligns with the user's preferences through scoring. To achieve this, the parser must exhaustively enumerate all valid candidates defined by the grammar. However, GSM contains recursive and ambiguous structures, which can substantially

- 1) $\text{piece}_{\text{key}=x \in \mathbb{K}} \rightarrow P^+$
- 2) $P \rightarrow TR$
- 3) $\text{piece} \rightarrow TR_{\text{key}=x \in \mathbb{K}}$
- 4) $TR \rightarrow DR t$
- 5) $DR \rightarrow SR d$
- 6) $TR \rightarrow TR DR$
- 7) $XR \rightarrow XR XR \quad (\forall XR \in \mathbb{R})$
- 8) $TR \rightarrow t$
- 9) $DR \rightarrow d$
- 10) $SR \rightarrow s$
- 11) $t \rightarrow tp$
- 12) $t \rightarrow tcp$
- 13) $s \rightarrow sp$
- 14) $d \rightarrow dp$
- 15) $X_{\text{key}=y} \rightarrow TR_{\text{key}=\psi(X,y)} \quad (\forall X \in \mathbb{F}, \forall y \in \mathbb{K})$
- 16) $X_{\text{key}=y \text{ maj/min}} \rightarrow X_{\text{key}=y \text{ min/maj}} \quad (\forall X \in \mathbb{F}, \forall y \in \mathbb{K})$
- 17) $X \rightarrow D(X) X \quad (\forall X \in \mathbb{S})$
- 18) $X \rightarrow \Delta(X) X \quad (\forall X, \Delta(X) \in \mathbb{S})$
- 19) $D(X) \rightarrow \begin{cases} V/VI/X \mid VII/VI/X & \text{if } X \text{ is dim} \\ V/X \mid VII/X & \text{otherwise} \end{cases}$
- 20) $t \rightarrow I$
- 21) $t \rightarrow I IV I$
- 22) $s \rightarrow IV$
- 23) $d \rightarrow V \mid VII$
- 24) $tp \rightarrow \begin{cases} VI & \text{if key is major} \\ III & \text{if key is minor} \end{cases}$
- 25) $dp \rightarrow VII \text{ if key is minor}$
- 26) $sp \rightarrow \begin{cases} II & \text{if key is major} \\ VI, \flat II & \text{if key is minor} \end{cases}$
- 27) $tcp \rightarrow \begin{cases} III & \text{if key is major} \\ VI & \text{if key is minor} \end{cases}$
- 28) $X \rightarrow X^+ \quad (\forall X \in \mathbb{O})$

Fig. 2. Production rules defined for GSM

increase the number of parse candidates. While this is essential for capturing the interpretive richness of harmonic analysis, it also poses practical challenges: certain parsing methods may be unable to handle such structures, or may suffer from computational inefficiency due to exponential growth in parsing complexity.

Therefore, two key requirements must be met. First, the parser must correctly generate all parse trees for any context-free grammar, including those with left recursion and ambiguity. Second, its computational cost must remain practical to allow integration with the scoring function. Given these requirements, this section compares representative parsing methods—LL, LR (GLR), and chart—in light of GSM's structural characteristics, and explains the rationale for selecting the chart parser as the most suitable method.

1) LL(k) Parser:

The LL method is a top-down predictive parsing technique that is relatively easy to implement and offers fast parsing speed, making it efficient for syntactic analysis within a limited range of grammars. In particular, it excels in parsing tree structures such as nested structures that do not involve left

recursion or ambiguity. However, the LL method has clear limitations. Because it cannot handle left-recursive structures, it is difficult to apply to grammars that include left recursion and ambiguity, such as GSM. Moreover, converting an ambiguous grammar into an LL form often requires complex grammar transformations and excessive lookahead, significantly degrading grammar maintainability. Therefore, the LL method is not suitable for this study, which aims to exhaustively generate numerous parse candidates.

2) *GLR Parser:*

The LR method is a bottom-up shift–reduce parsing approach that generates parse trees in a table-driven manner. Generalized LR (GLR) extends LR to support ambiguity and is, in principle, capable of enumerating all possible parse candidates. However, when GSM’s modulation rules and substitute-chord combinations are incorporated into the grammar, the resulting LR tables grow to an intractable size, and standard parser generators can fail during table construction. In addition, GLR incurs substantial processing overhead, which makes it ill-suited to domains such as musical grammar, where syntactic structures must be revised repeatedly through trial and error.

3) *Chart Parser:*

The chart parser is a chart-based parsing algorithm compatible with any CFG, and it can handle left-recursive structures and ambiguous grammars. It is well suited to grammars with recursive and nested structures, such as GSM. In particular, this study requires exhaustive enumeration of all parse tree candidates, followed by selection of the optimal result using a scoring function, making the flexibility of the chart parser especially important. Furthermore, the chart parser employs a dynamic programming approach using intermediate chart data during parsing, allowing for reuse of partial parses and thus offering advantages in computational efficiency. In this study, we use the EarleyChartParser [7] implementation provided in the Python natural language processing library NLTK. NLTK offers high flexibility in grammar specification, robust support for parse tree visualization and error detection, and is well suited for grammar design and adjustment involving iterative experimentation.

III. PROPOSED METHOD

In this study, we propose a personalized harmonic analysis method for chord progressions that selects the parse tree most consistent with the user’s analytical preferences from among the various possible harmonic interpretations. The method performs an exhaustive search of parse trees using GSM, and selects among them based on a user-defined scoring function. The scoring function defines multiple evaluation axes grounded in music theory and parse tree, such as the number of nodes, the number of secondary dominants, diatonic descending fifth, the number of modulations, and the degree of conformity with a user-specified key. By assigning weights to each of these axes, users can reflect their analytical preferences. For exhaustive enumeration of parse trees, we adopt the chart parser, which can analyze all structures of a CFG, and

implement it using the nltk.EarleyChartParser [7] included in the Python NLTK library. This parser supports grammars that include left recursion and ambiguity, and is highly compatible with the hierarchical and recursive syntactic generation of GSM. The proposed method does not rely on model training, and allows for flexible modification of analysis policies via arbitrary parameter settings. As a result, it provides a new framework for harmonic analysis that can derive different interpretations for the same chord progression, depending on the user’s analytical preference.

A. *GSM Rules*

In this study, we use a CFG based on GSM, used for the syntactic analysis of chord progressions. While the syntactic rules of GSM are designed in accordance with the original work, we exclude from consideration certain structures involving modulation—specifically Rule 15 (Fig. 2) in order to balance computational efficiency and search scope. Rule 15 allows the left-hand side of a functional symbol in $\mathbb{F}$ to expand into a functional region in $\mathbb{R}$ belonging to a different key, thereby enabling modulation at any point in the chord progression. If this rule is permitted during parse tree construction, nonterminal symbols can transition into different keys, resulting in an explosive increase in the number of parse trees and exponentially growing computational cost. In terms of parsing, this means that nonterminal symbols expand into deeper syntactic levels, and the number of derivable parse trees increases exponentially. To avoid such a vast search space, this study places greater emphasis on the scoring function reflecting the user’s analytical preferences.

B. *Selection of Parse Trees via Scoring Function*

For each parse tree, we define multiple evaluation axes based on music theory, quantify and weight them, and compute a total score. The user can assign an arbitrary weight w_i to each evaluation axis to reflect their individual analytical preferences.

1) *Count of Nodes:* This refers to the total number of nonterminal nodes contained in the parse tree. The greater number of nodes indicates the more complex structure. Therefore, fewer nodes are preferred, and a penalty is applied as the number of nodes increases.

2) *Count of Secondary Dominants:* Chains of secondary dominants as defined by Rules 17 and 19 (Fig. 2) are descending perfect fifth that result in strong resolution. The number of such patterns in the parse tree is counted. When these occur outside the diatonic scale, the sense of tonality becomes weaker. Thus, analyses with more chains of secondary dominants are penalized more heavily in this study.

3) *Count of Diatonic Descending Fifth:* diatonic descending fifth as defined by Rule 18 (Fig. 2) refer to descending fifth within the diatonic scale. Although the sense of resolution is weaker compared to secondary dominants, these progressions maintain tonal consistency and provide stability. However, in this study, analyses with more such chains are also penalized. The reason is that interpretations based on harmonic functions

(such as tonic) provide a stronger and more stable sense of tonality.

4) *Count of Modulations*: This refers to the total number of nodes within the parse tree that belong to a key different from the overall key of “piece”. By counting modulations, we evaluate whether the analysis maintains tonal consistency. Fewer modulations are regarded as indicating clearer structure, and a penalty is applied for each modulation.

5) *Key Specification*: If the user explicitly specifies a target_key for analysis, we evaluate whether this key is assigned to the top-level node (“piece_X”) of the parse tree. This count function returns 0 if the key of “piece” matches the user-specified key, and 1 otherwise.

6) *Total Scoring Method*: For each of the above evaluation axes, we assign weighting coefficients w_1, w_2, w_3, w_4, w_5 , and define the total score as follows:

$$S(T) = w_1 N_1(T) + w_2 N_2(T) + w_3 N_3(T) + w_4 N_4(T) + w_5 N_5(T) \quad (1)$$

Here, T denotes a parse tree, $N_1(T)$ is the number of nodes, $N_2(T)$ is the number of secondary dominants, $N_3(T)$ is the number of diatonic descending fifth, and $N_4(T)$ is the number of modulations. The last term, $N_5(T)$, represents the key specification penalty. The target key is specified in the form “target_key=Cm”.

In the above scoring scheme, a lower total score $S(T)$ indicates that the analysis better matches the user’s analytical preferences, and is thus regarded as a personalized analysis. Ultimately, the parse tree T with the smallest score $S(T)$ is selected from the set of generated candidates and presented to the user. If the user wishes to adjust the analysis tendencies, they can modify the weight coefficient vector—for example, by reducing $(w_3, w_4, \dots)$ —to favor analyses that emphasize progressions such as secondary dominants or diatonic descending fifth, according to their preferences.

IV. EXPERIMENTS

A. Objective

The objective of this experiment is to evaluate the effectiveness of chord progression analysis tailored to the user’s analytical preferences. This study investigates whether varying the weights of evaluation axes in the scoring function according to the user’s analytical preferences can yield harmonic analyses that better match those preferences.

B. Target Chord Progression and Analysis Conditions

We present only a single running example in this paper. Our objective is to illustrate how the multiple count functions and the weights behave; hence this illustrative example suffices for exposition. However, to demonstrate the validity of the approach, further examples will be required.

Specifically, the target chord progression consists of eight chords:

$$C \rightarrow D \rightarrow G \rightarrow C \rightarrow Em \rightarrow Fm \rightarrow Bb \rightarrow G$$

For this progression, we conducted analyses under the following conditions by varying the weight coefficients w_i .

1) Condition 1: Emphasis on Tonal Harmonic Functions:

This condition assumes an analytical preference that prioritizes analysis based on tonal harmonic functions such as tonic, subdominant, and dominant, rather than focusing on secondary dominants or diatonic descending fifth. Under this preference, parse trees in which nodes such as “ $D(X)$ ” and “ $\Delta(X)$ ” (representing secondary dominants and diatonic descending fifth, respectively) appear less frequently, and nodes corresponding to functional symbols in $\mathbb{F}$ appear more frequently, are considered to better represent a tonal harmonic structure. If the weights w_2 and w_3 for secondary dominants and diatonic descending fifth are set high, then a parse tree with a higher frequency of $\mathbb{F}$ -symbol nodes can be regarded as an analysis based on tonal harmonic functions.

2) Condition 2: Emphasis on Secondary Dominants:

This condition assumes an analytical preference that actively favors syntactic structures utilizing secondary dominants, even at the cost of departing from diatonic tonality. Therefore, parse trees in which “ $D(X)$ ” nodes appear frequently are considered to include more progressions based on secondary dominants. If the weight w_2 for secondary dominants is set low, and “ $D(X)$ ” nodes appear frequently in the parse tree, the result can be interpreted as an analysis that emphasizes secondary dominants.

3) Condition 3: Key Specification “Cm”:

This condition assumes an analytical preference in which the chord progression is intended to be interpreted in the key of Cm. Under this preference, it is desirable for the entire parse tree to be developed as consistently as possible based on the Cm key. Parse trees that begin with “piece_Cm” and contain few modulations to other keys are preferred. By setting “target_key=Cm” and increasing the weights w_4 for the number of modulations and w_5 for key specification, if the resulting parse tree starts with “piece_Cm” and a high proportion of its constituent nodes match the Cm key, the analysis is regarded as highly consistent with the user’s intended tonal preference.

C. Experimental Results

A total of 2,688 parse trees were generated. For each condition, we present a summary of the analysis with the lowest score and a breakdown of the score by evaluation axis, including the contribution of each axis and the total score. In $X \times Y$, X represents the weight w_i , and Y is the corresponding feature count $N_i(T)$.

TABLE I
MINIMUM TOTAL SCORE BY CONDITION

	Condition 1	Condition 2	Condition 3
Nodes	1×33	1×32	1×34
Secondary Dominants	10×1	-10×2	1×1
Diatonic Descending Fifth	10×0	1×1	1×1
Modulations	1×14	1×13	10×15
Key Specification	1×0	1×0	50×0
Total Score	57	27	186

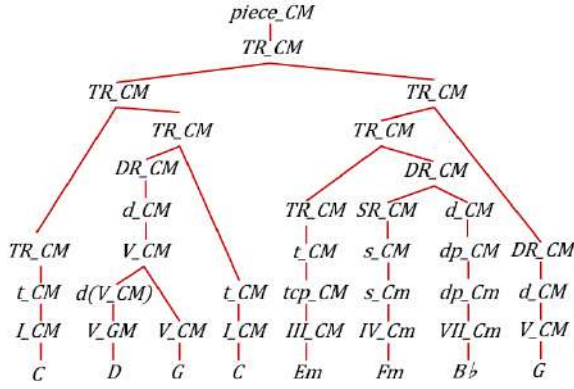


Fig. 3. Parse tree under Condition 1

1) Condition 1 (Emphasis on Tonal Harmonic Functions):

In this setting, the weights for “ $D(X)$ ” and “ $\Delta(X)$ ” (representing secondary dominant and diatonic descending fifth, respectively) were set to $w_2=w_3=10$ to impose penalties, while the weights for all other axes were set to $w=1$. The target key was not specified. In the lowest-scoring parse tree (Fig. 3), the entire structure was parsed as “piece_CM”, with a key switch to Cm occurring at the Fm–B $\flat$ section. In particular, the “D , G , C” progression was analyzed with only one occurrence of a “ $D(X)$ ” node, keeping the use of secondary dominants to a minimum. This is because “D , G” is outside the diatonic scale and necessarily requires one “ $D(X)$ ” node. Moreover, the progression “Fm , B $\flat$ ” was not analyzed using a “ $\Delta(X)$ ” node in the Cm key; instead, it was assigned to the “s” and “dp” nodes in the $\mathbb{F}$ category within the Cm key, thereby avoiding the use of “ $\Delta(X)$ ” node. With $w_2=w_3=10$, secondary dominants and diatonic descending fifth were minimized, favoring $\mathbb{F}$ -nodes and matching the tonal harmonic preference.

2) Condition 2 (Emphasis on Secondary Dominants):

In this setting, a negative weight of $w_2=-10$ was assigned to the occurrence count of “ $D(X)$ ” nodes, so that analyses with more occurrences received a lower score. All other weights were set to $w=1$, and the target key was not specified. In the lowest-scoring parse tree (Fig. 4), the “D , G , C” progression was interpreted using two “ $D(X)$ ” nodes, capturing two secondary dominants. Since the weight w_2 was set to a negative value, and “ $D(X)$ ” nodes appeared frequently in the parse tree, the result can be interpreted as emphasizing secondary dominants.

3) Condition 3 (Key Specification: Cm):

In this condition, a high weight of $w_5=50$ was assigned to analyses consistent with the specified target key (“target_key=Cm”), and $w_4=10$ was assigned for the number of modulations. All other weights were set to $w=1$.

In the lowest-scoring parse tree (Fig. 5), the entire structure is derived starting from piece_Cm. Since the user-specified target key is Cm (“target_key=Cm”), the key of “piece” matches; hence no penalty is applied under Key Specification (the corresponding count function returns 0). Chords such as C

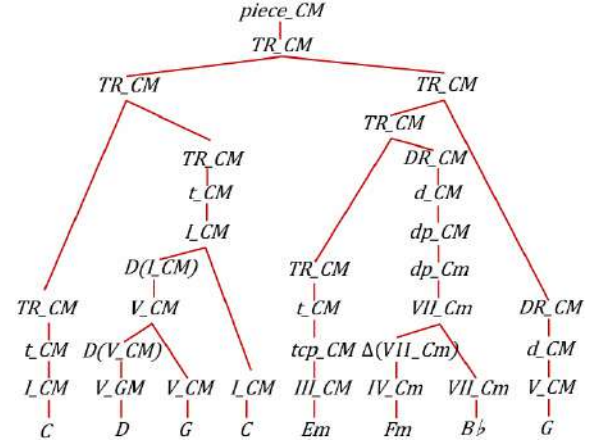


Fig. 4. Parse tree under Condition 2

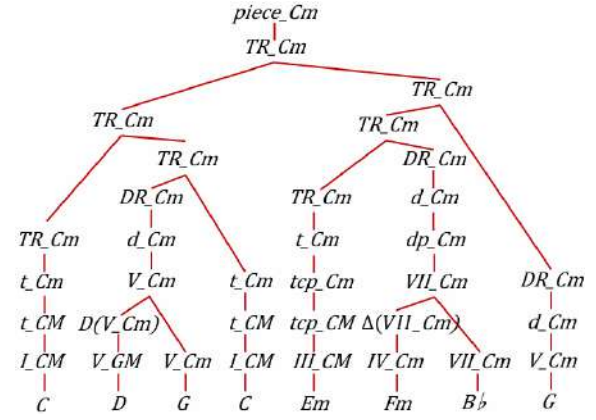


Fig. 5. Parse tree under Condition 3

and Em are diatonic chords in the key of Cm, and according to the grammar, they are derived through a modulation from the Cm key to the CM key, following the path: $t\_Cm \rightarrow tcp\_Cm \rightarrow tcp\_CM \rightarrow III\_CM \rightarrow Em$. On the other hand, chords like Fm, B $\flat$, and G are derived directly within the Cm key. Notably, G functions as the dominant (V) in both CM and Cm keys, but in the latter part of the progression, it is analyzed within the Cm key.

As a result of assigning $w_5=50$ for key specification and $w_4=10$ for modulation count, the parse tree began with “piece_Cm”, and the proportion of nodes matching the Cm key was higher compared to Conditions 1 and 2. Therefore, the analysis can be considered consistent with the analytical preference of interpreting the piece in the key of Cm.

V. DISCUSSION

In the comparative experiments under the three conditions, we examined how the weight settings w_i in the scoring function influence the analyses. The lowest-scoring parse tree under each condition differed significantly, clearly demonstrating that the weight configuration in the scoring function greatly affects the evaluation policy of the analysis algorithm. Furthermore, in all three conditions, it was confirmed that

modifying the weights w_i produced analyses aligned with the corresponding analytical preferences. From these results, we conclude that the scoring function is useful as a mechanism for generating parse trees that reflect analytical preferences, thus contributing to personalized analysis.

At the same time, the present experiment is preliminary: it uses a single chord progression. To establish generality and human alignment, we will evaluate across a curated set of diverse progressions and run a small expert study with trained musicians and analysts. For each progression, we will compare the scoring-selected analysis with the expert's preferred analysis and report agreement and plausibility under modest adjustments of the weighting w_i .

We acknowledge that the current method does not yet fully reproduce how experts interpret diverse chord progressions. To better capture and reflect analytical preferences, we will extend the system in three directions.

First is the validity of the weights w_i . For example, in Condition 3, setting the weights for modulation count and key specification to $w_4=w_5=10$ with “target_key=Cm” resulted in the selection of a parse tree starting with “piece_CM”. This is because the chord progression used in the experiment contains more chords from the CM key, and when attempting to interpret it in the Cm key, many CM key nodes had to be converted to Cm-key nodes, thereby increasing the modulation count penalty. As a result, in this experiment, we set the key specification weight $w_5=50$, higher than the modulation penalty, to better reflect the analytical preference. The present study uses hand-set weights w_i . Going forward, we will replace them with data-driven estimates and examine the stability of the scoring-based tree selection under small variations of w_i . In parallel, we will place probabilistic structure on the grammar—for example, a probabilistic context-free grammar (PCFG)—and integrate it into our framework. Recent machine-learning work on harmonic structure analysis indicates that such data-driven formulations are both feasible and informative [4], [8], [9], [10].

Second is the validity of the evaluation axes themselves. In this study, we used five evaluation axes, but we plan to introduce additional axes that can more accurately capture analytical preferences. For instance, an axis such as “enforce modulation to a specific key at a specific point in the progression” could be considered. Since interpretations involving modulation often give rise to multiple possible analyses, supporting more detailed user preferences like “I feel that this part modulates from key X to Y” would enhance the expressiveness of the framework.

Third, we outline the potential to extend the grammar rules. As noted in Section III-A, we left GSM Rule 15 (Fig. 2) disabled to avoid combinatorial explosion. We plan to re-enable it by controlling the search during parsing with pruning and beam search. Concretely, we compute partial scores for Earley-style chart items and use them to select and limit rule expansions as the parse proceeds, keeping only the top candidates within a fixed beam and discarding low-scoring branches using a threshold. This incremental scoring reduces

both the number of intermediate derivations and the number of completed parses, so even with Rule 15 enabled the overall runtime remains tractable. We also plan to add rules for seventh chords and for chord extensions. These additions will broaden coverage for real-world progressions and will expand the space of analytical preferences the system can accommodate.

From these considerations, we conclude that while the scoring function demonstrates a certain level of utility for producing analyses aligned with analytical preferences, it does not yet fully achieve the ideal of personalized analysis. We plan to continue extending the system to better capture and reflect diverse analytical preferences.

VI. CONCLUSION

In this study, we aimed to select harmonic analyses that reflect users' analytical preferences by utilizing GSM and a scoring function, which incorporated evaluation axes such as the number of nodes, occurrences of secondary dominants, diatonic descending fifth, number of modulations, and degree of match with a specified key. By adjusting the weights assigned to each axis, analyses were tailored to individual analytical preferences.

As a result, we confirmed that our framework can generate chord-progression analyses aligned with a user's analytical preferences. Going forward, we will replace the hand-set weights with data-driven estimates as suitable data become available and incorporate probabilistic structure (e.g., PCFG) into the grammar; in parallel, we will broaden the evaluation axes and extend the harmonic rules to further improve robustness and flexibility.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 25H01169, 23K00243.

REFERENCES

- [1] M. Rohrmeier, “Towards a generative syntax of tonal harmony,” *J. Math. Music*, vol. 5, no. 1, pp. 35–53, Mar. 2011.
- [2] F. Lerdahl and R. S. Jackendoff, *A Generative Theory of Tonal Music*. Cambridge, MA, USA: MIT Press, 1996.
- [3] Y. Ogura, H. Ohmura, S. Tojo, and K. Katsurada, “Estimating unexpectedness in jazz harmony with a probabilistic incremental parser,” in *Proc. 18th Sound and Music Computing Conf. (SMC)*, 2021, pp. 292–299.
- [4] D. Harasim, M. Rohrmeier, and T. J. O'Donnell, “A generalized parsing framework for generative models of harmonic syntax,” in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2018, pp. 152–159.
- [5] H. Riemann, *Harmony Simplified: Or the Theory of the Tonal Functions of Chords*, H. Bewerunge, trans. London, U.K.: Augener, 1896.
- [6] J. Earley, “An efficient context-free parsing algorithm,” *Commun. ACM*, vol. 13, no. 2, pp. 94–102, 1970.
- [7] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009.
- [8] D. Harasim, T. J. O'Donnell, and M. Rohrmeier, “Harmonic syntax in time: Rhythm improves grammatical models of harmony,” in *Proc. 20th Int. Soc. Music Information Retrieval Conf. (ISMIR)*, Delft, The Netherlands, pp. 335–342, Nov. 2019.
- [9] R. Cartuyvels, J. Koslovsky, and M.-F. Moens, “The potential of unsupervised induction of harmonic syntax for jazz,” *Trans. Int. Soc. Music Inf. Retr.*, vol. 8, no. 1, pp. 165–178, Jun. 2025.
- [10] M. Gotham, G. Micchi, N. Nápoles López, and M. Sailor, “When in Rome: A meta-corpus of functional harmony,” *Trans. Int. Soc. Music Inf. Retr.*, vol. 6, no. 1, pp. 150–166, Nov. 2023.

Enhancing Multivariate Time Series Anomaly Detection with Feature-Scaled Reconstruction in LSTM Autoencoders

1st Truong Son Pham
Le Quy Don Technical University
Hanoi, Vietnam
son.pham@lqdtu.edu.vn

2nd Quang Phong Nguyen
Le Quy Don Technical University
Hanoi, Vietnam
phongbdatt56@gmail.com

3rd Nguyen Phuong Do
Le Quy Don Technical University
Hanoi, Vietnam
nguyenphuongdo22@gmail.com

Abstract—In this paper, we address anomaly detection in multivariate time series for cybersecurity applications by leveraging Long Short-Term Memory (LSTM) autoencoders trained exclusively on benign (attack-free) data. We propose a novel loss function, Feature-Scaled Reconstruction Loss (FSRL), which dynamically weights the reconstruction error of each feature. FSRL assigns larger loss weights to features that are consistently well-reconstructed by the autoencoder (identified in a pretraining phase), thereby amplifying deviations on these reliable features. We design both a supervised weighting scheme (leveraging Random Forest feature importances from labeled data) and an unsupervised scheme (using per-feature reconstruction error from a pretrained model) to determine the feature weights. We evaluate the proposed FSRL-enhanced autoencoder on two diverse datasets: CICDDoS2019 (network traffic with DDoS attack labels) and the Server Machine Dataset (system metrics). Experimental results demonstrate that our approach significantly improves anomaly detection performance, achieving higher precision, recall, and F1-scores compared to conventional LSTM autoencoders using a standard mean squared error (MSE) loss or a Random Forest-weighted loss baseline. Notably, FSRL yields higher F1-scores than both the unweighted and RF-weighted loss setups, indicating its effectiveness in focusing the model on dependable features. This work highlights the importance of feature-specific loss scaling in LSTM autoencoders and contributes an effective new approach for robust anomaly detection in cybersecurity contexts.

Index Terms—Anomaly Detection, LSTM Autoencoder, Feature-Scaled Reconstruction Loss, Multivariate Time Series

I. INTRODUCTION

Anomaly detection in multivariate time series is a core task in cybersecurity [1], particularly in intrusion detection systems where identifying abnormal network behavior is critical to preventing attacks. Long Short-Term Memory Autoencoders (LSTM-AEs) [2] have been widely adopted for this purpose due to their ability to capture temporal dependencies and reconstruct normal traffic patterns. When trained exclusively on benign data, these models can detect anomalies as deviations that yield high reconstruction errors.

However, conventional LSTM-AE approaches use a uniform reconstruction loss, implicitly assuming all input features are equally important and equally stable. This assumption is problematic in high-dimensional time-series data where

features vary significantly in both reconstruction difficulty and informativeness. Some features are consistently well-reconstructed and serve as strong indicators of abnormality, while others are inherently noisy or less relevant. Treating these features equally may dilute the model’s sensitivity to meaningful anomalies.

To address this, we propose a novel *Feature-Scaled Reconstruction Loss* (FSRL), which introduces per-feature weighting into the autoencoder’s loss function. These weights are derived from each feature’s average reconstruction error during a pretraining phase, where lower-error (more stable) features are given higher importance. Alternatively, weights can also be computed using supervised feature importance scores from a Random Forest classifier trained on labeled data. By emphasizing reliable features, our method improves the model’s ability to detect subtle but meaningful deviations in time-series behavior.

We evaluate our proposed FSRL-LSTM-AE model on two benchmark datasets: CICDDoS2019 [3], which contains labeled DDoS attack traffic, and the Server Machine Dataset (SMD) [4], a multivariate time series dataset of server metrics with labeled anomalies. Our method outperforms both the standard MSE loss and a baseline using Random Forest feature weighting, achieving higher precision, recall, and F1-score across both domains.

In summary, our contributions are threefold:

- We propose FSRL, a new loss function for LSTM-AEs that incorporates feature-specific weighting to improve anomaly detection sensitivity.
- We design both unsupervised and supervised strategies for estimating feature importance to guide reconstruction.
- We demonstrate consistent performance improvements on two diverse real-world datasets relevant to cybersecurity and system monitoring.

The remainder of the paper is organized as follows. Section II discusses related work in time-series anomaly detection and feature-aware modeling. Section III introduces the LSTM-AE framework and the rationale for feature weighting. Section IV details our proposed approach. Section V presents datasets and experimental setup. Section VI reports

performance metrics and comparisons. Finally, Section VII concludes with a discussion of future directions.

II. RELATED WORK

Anomaly detection in multivariate time series has attracted extensive research attention due to its applications in cyber security, system monitoring, and industrial diagnostics. Traditional approaches such as ARIMA models [5] and control charts [6] rely on statistical assumptions and handcrafted features. While useful in simple cases, they often fail to capture non-linear temporal dependencies or high-dimensional interactions among features, limiting their effectiveness in modern, real-world scenarios.

With the rise of deep learning, sequence models, particularly Long Short-Term Memory (LSTM) networks [7], have become a popular choice for modeling temporal dynamics. LSTM-based Autoencoders (LSTM-AEs) have proven effective for unsupervised anomaly detection by learning to reconstruct normal patterns and flagging sequences with high reconstruction error as anomalous. Malhotra et al. [8] pioneered this idea for multi-sensor data, showing that LSTM encoder-decoder architectures can detect anomalies in time-series without labeled attack samples. Wei et al. [9] further applied this approach in cybersecurity, using an LSTM-AE to detect DDoS attacks in network traffic by training on benign flows and identifying abnormal ones via reconstruction loss.

Building upon deterministic models, probabilistic frameworks such as variational autoencoders (VAEs) have been introduced to improve robustness and uncertainty modeling. Yokkampon et al. [10] propose a Multi-Scale Convolutional Variational Autoencoder (MSCVAE) for unsupervised anomaly detection in multivariate time series, motivated by the need to capture both temporal dependencies and inter-series correlations without labeled data. They construct multi-scale attribute matrices to represent system states at different levels, then use a convolutional VAE for reconstruction and an attention-based ConvLSTM to learn temporal patterns. A novel ERR-based thresholding strategy is introduced to handle class imbalance more effectively than ROC-based thresholds, and experiments on four datasets show MSCVAE outperforms competing methods in both accuracy and robustness. In parallel, Li et al. [11] introduced MAD-GAN, a GAN-based framework where the generator learns to replicate normal sequences while the discriminator helps distinguish real from generated data, yielding powerful anomaly scores.

To better handle long-range dependencies and feature correlations, attention mechanisms and Transformer-based models have also been proposed. Xu et al. [12] presented the Anomaly Transformer, which uses a novel association discrepancy-based attention mechanism to learn pairwise correlations between time steps. This method achieved state-of-the-art results on several benchmark datasets. Likewise, graph-based models such as those by Deng et al. [13] use graph neural networks to model inter-feature dependencies across time, offering improved context-aware anomaly detection in complex sensor systems.

In parallel, several works have emphasized the role of feature importance in improving detection performance. Li et al. [14] introduced SA2E-AD with an attention mechanism within the layers of a stacked autoencoder, enabling the model to focus preferentially on the most informative features during reconstruction. By dynamically assigning importance to input dimensions based on the data itself, this feature-aware autoencoder improves detection performance, especially in complex multivariate time-series scenarios. Other approaches utilize external models, such as Random Forests, to rank features by their predictive power and use these rankings to adjust loss weights or select subsets of features.

Our work builds upon these insights but takes a distinct direction. Instead of relying on external attention modules or supervised feature rankings, we propose a simple yet effective *Feature-Scaled Reconstruction Loss* (FSRL), which leverages feature-wise reconstruction error statistics learned during unsupervised pretraining. By assigning higher loss weights to features that are typically reconstructed well, the model naturally becomes more sensitive to deviations in those stable dimensions. To our knowledge, this direct use of reconstruction behavior to guide loss weighting within LSTM-AEs has not been explored previously and complements the broader trend of feature-aware anomaly detection.

III. BACKGROUND

A. Long Short-Term Memory Auto-Encoder

A Long Short-Term Memory auto-encoder (LSTM-AE) [2] is an encoder-decoder neural architecture for sequence data that leverages LSTM recurrent units in both the encoder and decoder. The encoder LSTM processes an input time-series $X = \{x_t\}_{t=1}^T$ (with $x_t \in \mathbb{R}^D$) step by step and compresses it into a fixed-length latent representation. Typically, the encoder's final hidden state $\mathbf{z} = \mathbf{h}_T$ acts as this latent vector capturing the salient information of the sequence. The decoder LSTM then uses this vector $\mathbf{z}$ to reconstruct the sequence, producing an output $\hat{X} = \{\hat{x}_t\}_{t=1}^T$ that aims to match the original input sequence. In practice, the latent vector can be provided to the decoder at each time step (e.g. by “repeating” $\mathbf{z}$ as the decoder input for all steps) to ensure the decoder has access to the encoded context at every time step. This encoder-decoder setup forces the model to learn a compact representation $\mathbf{z}$ that retains the essential temporal features needed to regenerate the sequence, thereby “capturing latent features from time-series data”.

To train the LSTM-AE, one minimizes a reconstruction loss that measures the discrepancy between the input and its reconstruction. The mean squared error (MSE) is the standard choice for this loss. For a single sequence, the reconstruction MSE can be written as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T} \sum_{t=1}^T \sum_{d=1}^D (x_{t,d} - \hat{x}_{t,d})^2, \quad (1)$$

i.e., the average of squared differences between each input value and its reconstructed value (summed over all time

steps and features). The LSTM-AE's parameters (encoder and decoder weights θ_e, θ_d) are learned via gradient descent to minimize this loss on the training data. A well-trained LSTM-AE reconstructs the original input from the compressed representation, facilitating the detection of anomalies by noting when the reconstruction error is high. In practice, the encoder and decoder architectures are often symmetric (e.g. same hidden state size), and training uses techniques like Adam optimization and early stopping for stability. The LSTM-AE framework has been widely used for sequence anomaly detection and time-series feature learning, owing to its ability to model temporal dependencies and reconstruct normal patterns with low error.

B. Incorporating Feature Importance in Weighted Loss Generation

In the context of anomaly detection, not all input features contribute equally to distinguishing between normal and abnormal behavior. Some features carry more discriminative power or are more stable under benign conditions. To better capture this insight during training, we adopt a *weighted loss* approach that integrates feature importance into the reconstruction objective of the LSTM Autoencoder.

In conventional autoencoder training, the model minimizes the mean squared reconstruction error across all features equally. However, this uniform treatment can dilute the model's sensitivity to critical features that are either more informative or more robustly reconstructed. To address this, we introduce feature-aware weighting strategies into the loss function.

Specifically, we explore two mechanisms to compute feature importance for use in the loss:

- **Supervised Feature Importance:** A Random Forest classifier [15] is trained on labeled data (benign vs. attack) to assess the discriminative power of each feature. The resulting feature importance scores are normalized and used as weights in the loss function [16]. This approach integrates prior knowledge from classification tasks into the reconstruction-based detection pipeline.
- **Unsupervised Reconstruction-Based Importance:** An LSTM Autoencoder is first pre-trained on benign data using the standard MSE loss. After training, the average reconstruction error per feature is computed. Features that are reconstructed consistently well (i.e., with low error) are considered more reliable indicators of abnormality. Their weights are set inversely proportional to the reconstruction error, emphasizing accurate reconstruction of these features during anomaly detection training.

In both cases, the weighted reconstruction loss function takes the form:

$$\mathcal{L}_{\text{weighted}} = \frac{1}{T} \sum_{t=1}^T \sum_{d=1}^D w_d \cdot (x_{t,d} - \hat{x}_{t,d})^2, \quad (2)$$

where w_d is the importance weight for feature d , derived from either supervised or unsupervised analysis.

This generation-aware weighted loss formulation encourages the model to focus on features that are either statistically important or diagnostically stable, thereby improving both anomaly detection accuracy and robustness across diverse scenarios.

IV. PROPOSED MODEL

A. Long Short-Term Memory Autoencoder with Feature-Scaled Reconstruction Loss

In many applications the *scale* of reconstruction errors can vary significantly across features. Some features may inherently have higher reconstruction error (due to higher noise or variability) even under normal conditions, while others are reconstructed very precisely. Feature-Scaled Reconstruction Loss (FSRL) is a modification of the loss function that introduces per-feature scaling (or weighting) of the reconstruction error based on statistical characteristics of each feature's error distribution on the training data.

The key idea is to normalize the reconstruction error for each feature to account for differences in error variance. By scaling errors feature-wise, the autoencoder's loss (and subsequently its learning and anomaly decisions) is not dominated by any single feature solely because that feature has a larger error scale under normal conditions. This yields a fairer evaluation across features, highlighting truly anomalous deviations in typically low-error features and down-weighting large errors in high-variance features that might otherwise be considered "normal" noise.

The proposed method utilizes a Long Short-Term Memory Autoencoder (LSTM-AE) to model normal network traffic behavior and detect anomalies through reconstruction error. To enhance sensitivity to deviations in well-modeled features, we introduce a Feature-Scaled Reconstruction Loss (FSRL), which assigns higher importance to features that are typically reconstructed accurately.

Let $X = \{x_t\}_{t=1}^T$ be an input time-series sequence, where each $x_t \in \mathbb{R}^D$ is a multivariate observation with D features. The encoder LSTM processes the input sequentially, yielding a final hidden state h_T and cell state c_T :

$$(h_t, c_t) = f_{\text{enc}}(x_t, h_{t-1}, c_{t-1}), \quad t = 1, \dots, T, \quad (3)$$

with initial states $(h_0, c_0) = (0, 0)$. The decoder LSTM reconstructs the sequence starting from the encoded states:

$$(\tilde{h}_t, \tilde{c}_t) = f_{\text{dec}}(\hat{x}_{t-1}, \tilde{h}_{t-1}, \tilde{c}_{t-1}), \quad \hat{x}_t = g(\tilde{h}_t), \quad (4)$$

where $g(\cdot)$ is a linear output layer, and $(\tilde{h}_0, \tilde{c}_0) = (h_T, c_T)$. The model is trained so that the reconstructed sequence $\hat{X} = \{\hat{x}_t\}_{t=1}^T$ closely matches the input X .

The standard Mean Squared Error (MSE) loss is defined as in equation 1. To incorporate feature importance, we introduce a weight w_d for each feature d as shown in equation 2. These weights are computed using the reconstruction error of a pre-trained LSTM-AE trained only on benign data. After training,

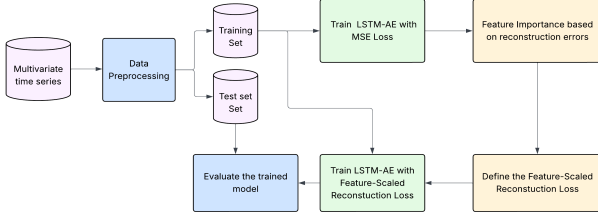


Fig. 1. General Flow of Proposed Approach

we evaluate the model on the same benign dataset and compute the average MSE for each feature:

$$\varepsilon_d = \mathbb{E}_t \left[(x_{t,d} - \hat{x}_{t,d})^2 \right]. \quad (5)$$

To assign more importance to features that are reconstructed well (i.e., low ε_d), we define the weights as the normalized inverse of the feature-wise MSE:

$$w_d = \frac{\varepsilon_d^{-1}}{\sum_{j=1}^D \varepsilon_j^{-1}}. \quad (6)$$

A small constant ϵ is added in practice to avoid division by zero. These weights are then used in the final training phase of a new LSTM-AE model trained with the weighted loss $\mathcal{L}_{\text{weighted}}$.

B. General Flow of the Approach

As shown in Figure 1, we adopt a one-class learning strategy to detect anomalies by training a model exclusively on normal (benign) data. The proposed method, LSTM Autoencoder with Feature-Scaled Reconstruction Loss (FSRL-LSTM-AE), is designed to learn the normal patterns in multivariate time-series data and identify deviations that indicate abnormal or malicious activity. The overall process consists of two main phases. In the first phase, an LSTM Autoencoder is pre-trained using only benign sequences. This model is optimized with a standard mean squared error (MSE) loss to accurately reconstruct normal data. After training, the model is used to evaluate the reconstruction quality of each feature, allowing us to estimate how well each dimension of the data is learned under normal conditions.

Based on this evaluation, we calculate feature importance scores by measuring the average reconstruction error for each feature. Features that are consistently reconstructed with low error are considered more reliable indicators of abnormality. Accordingly, each feature is assigned a weight inversely proportional to its reconstruction error, and these weights are normalized to ensure they sum to one.

In the second phase, a new LSTM Autoencoder is trained using these computed weights to guide the learning process. The loss function is modified to prioritize accurate reconstruction of the more reliably modeled features. This helps the model focus on subtle deviations in stable features, which are more indicative of anomalies.

Model performance is evaluated using standard classification metrics. Anomalies are flagged when the reconstruction loss of a given sequence exceeds a predefined threshold, which may be fixed or dynamically determined based on validation statistics or empirical heuristics. This evaluation strategy allows us to assess both the detection capability and robustness of the proposed approach.

V. EXPERIMENTS

A. Datasets

CICDDoS2019 [3] is a benchmark dataset designed to support research and development of DDoS (Distributed Denial of Service) detection techniques. It addresses the limitations of earlier datasets and provides a realistic and comprehensive collection of modern DDoS attacks, particularly amplification through reflection techniques. The dataset includes attack types such as DNS, NTP, LDAP, NetBIOS, SSDP, and UDP floods, as well as exploitation-based SYN flood attacks. The feature set is consistent with other datasets from the Canadian Institute for Cybersecurity, such as CIC-IDS2017, CIC-IDS2018, and CIC-DoS2017, ensuring consistency for comparative studies. The dataset was processed using CICFlowMeter, which converted raw PCAP files into CSV format, yielding over 80 flow-level statistical features.

The Server Machine Dataset (SMD) [4] is a multivariate time series dataset collected for the purpose of anomaly detection in data center server operations. It contains five weeks of server monitoring data from 28 different servers. Each time step in the dataset consists of 38 features that reflect system-level metrics. Each server subset is structured with normal data in the training set and labeled anomalies in the test set, making it ideal for unsupervised or semi-supervised anomaly detection.

B. Data Preprocessing

To prepare the CICDDoS2019 dataset for model training, we first selected 20 relevant features from the original set of over 80 flow-level features. These features were chosen based on their importance in characterizing network traffic patterns associated with DDoS activity. The `Class` column was retained as the label to differentiate between `Benign` and `Attack` flows. All rows containing missing (NaN) values were removed to ensure data integrity. Feature values were then normalized to the range [0,1] using `MinMaxScaler` to facilitate stable training and to avoid domination by features with larger numeric scales. To incorporate temporal context, the data was segmented into sequences using a sliding window of 10 time steps, resulting in samples of shape (10, 20), where each sample captures 10 consecutive flows. The training set was constructed exclusively from sequences that consisted entirely of benign traffic, yielding 97.822 training samples. In contrast, the test set included a mix of benign and attack sequences and comprised 431.362 samples, enabling comprehensive evaluation of the model's anomaly detection capability.

For the SMD dataset, the time-series data was similarly reshaped into fixed-length sequences using a window size of 10. This produced multivariate sequences of shape (10, 38), with each sample representing the temporal evolution of 38 server performance metrics. Each server machine subset in the dataset was treated as a separate task for both training and evaluation, reflecting the distributed and independent nature of server monitoring. The training set for each server contained only normal system behavior, while the test set included both normal and anomalous time segments, with explicit labels indicating the presence of anomalies. This structure allowed the model to learn normal operational patterns during training and to identify deviations during evaluation.

C. Evaluation Metrics

To assess the performance of the proposed anomaly detection method, we adopt three widely used metrics: *Precision*, *Recall*, and *F1-score*. These metrics are computed based on the counts of true positives (TP), false positives (FP), and false negatives (FN) when classifying each test sequence as either Normal or Anomaly.

Precision measures the proportion of detected anomalies that are actually true anomalies:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (7)$$

A high precision value indicates that the model produces few false alarms.

Recall (also known as detection rate or sensitivity) measures the proportion of actual anomalies that the model successfully detects:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (8)$$

A high recall value indicates that the model is effective at identifying most anomalies present in the data.

The F1-score is defined as the harmonic mean of precision and recall, offering a unified metric that balances the influence of both false positives and false negatives:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (9)$$

To determine whether a test sequence is anomalous, we compute its reconstruction loss and compare it to a threshold derived from the distribution of training loss values. Specifically, we apply a percentile-based approach: several candidate thresholds are tested by varying the percentile (e.g., 95th, 97th, 99th) of the reconstruction loss distribution on the benign training set. The final threshold is selected based on the percentile that yields the best F1-score on a validation set. This method ensures that the threshold is data-driven and adapted to the scale and characteristics of each dataset. Precision, recall, and F1-score are then computed using the ground truth labels of the test data.

VI. RESULTS AND DISCUSSION

This section presents the evaluation results of three loss functions applied to anomaly detection using LSTM Autoencoders: standard Mean Squared Error (MSE) loss, weighted MSE loss using Random Forest (RF) feature importance, and the proposed Feature-Scaled Reconstruction Loss (FSRL). The detailed results are shown in Tabel I.

TABLE I
PERFORMANCE COMPARISON OF LOSS FUNCTIONS ON CICDDoS2019 AND SMD DATASETS

Dataset	MSE Loss	Weighted Loss (RF)	FSRL (Proposed)
CICDDoS2019			
Precision	0.9745	0.9602	0.9648
Recall	0.9430	0.9602	0.9766
F1-Score	0.9585	0.9602	0.9706
SMD - Server Machine Dataset			
Precision	0.8131	0.8426	0.9454
Recall	0.7529	0.8272	0.9095
F1-Score	0.7818	0.8207	0.9271

A. CICDDoS2019 Dataset

The CICDDoS2019 dataset, focused on network traffic anomalies related to DDoS attacks, yielded high performance across all methods. The standard MSE loss achieved the highest precision (0.9745), indicating strong reliability in positive predictions. However, the recall was notably lower (0.943), suggesting that some attack instances were missed. The FSRL method, in contrast, significantly improved recall to 0.9766 while maintaining competitive precision (0.9648), resulting in the highest F1-score (0.9706) among the three approaches. This suggests that emphasizing features that are well-reconstructed during pretraining enables the model to better generalize and identify subtle anomalies.

The weighted loss based on Random Forest feature importance achieved a balanced but slightly lower performance than FSRL, with both precision and recall at 0.9602. This indicates that while RF-based weighting introduces some improvement over uniform treatment of features, it is less effective than weights derived directly from reconstruction behavior.

B. SMD Dataset

On the more complex SMD dataset, which includes multivariate server monitoring data, the proposed FSRL approach demonstrated substantial improvements. It achieved the highest precision (0.9454) and recall (0.9095), leading to an F1-score of 0.9271. These results show that the model was highly effective in both minimizing false positives and capturing true anomalies across server performance metrics.

In comparison, the standard MSE loss lagged behind with an F1-score of 0.7818, primarily due to its low recall (0.7529), indicating difficulty in capturing diverse anomaly patterns. The RF-weighted loss performed moderately better with an F1-score of 0.8207 but still fell short of the performance achieved by FSRL.

C. Summary

Overall, the Feature-Scaled Reconstruction Loss outperformed both baselines on both datasets. Its ability to derive feature importance directly from model reconstruction behavior provides a tailored and data-driven mechanism to guide training. Unlike external models like Random Forests—which may not fully capture temporal reconstruction characteristics—FSRL tightly integrates weighting into the autoencoding framework. The approach is particularly beneficial in complex multivariate time-series contexts, such as server monitoring, where subtle deviations in typically stable features can be critical indicators of anomalies.

VII. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel *Feature-Scaled Reconstruction Loss* (FSRL) scheme tailored for LSTM-based autoencoders in anomaly detection. By leveraging pre-trained feature-wise MSE values to derive normalized loss weights, the proposed FSRL scales each feature’s contribution to the total reconstruction error. This effectively balances features of varying scales and importance, leading to improved anomaly detection performance. Through experiments on the CICD-DoS2019 network traffic dataset and the SMD (Server Machine Dataset) industrial time series, we demonstrated that FSRL significantly improves detection accuracy. For both benchmarks, FSRL achieved higher precision, recall, and F1-scores than the baseline loss functions. Notably, it outperformed both the conventional unweighted MSE loss and an alternative weighted loss based on Random Forest feature importance, underscoring the advantage of calibrating the reconstruction loss on a per-feature basis.

Moving forward, several avenues for future work can further enhance and extend the FSRL framework. First, we plan to evaluate FSRL on additional and more diverse anomaly detection datasets to assess its generality beyond the domains considered in this study. Second, integrating the FSRL approach into real-time inference pipelines is a key objective, in order to enable live monitoring applications that require low-latency anomaly detection; in this context, we also aim to study the scalability of the method to larger datasets and longer time series for practical deployment. Third, we intend to extend our loss-weighting strategy to other sequence modeling architectures, such as Transformer-based autoencoders, to investigate its effectiveness beyond LSTM models. Finally, we will explore automated weight calibration strategies beyond the current inverse-MSE heuristic. For example, meta-learning or optimization-based techniques could be employed to learn optimal feature weighting instead of using a fixed inverse-error scheme. We anticipate that addressing these future directions will broaden the applicability of FSRL and further enhance its impact on sequential anomaly detection tasks.

REFERENCES

- [1] Max Landauer, Florian Skopik, Branka Stojanović, Andreas Flatscher, and Torsten Ullrich. A review of time-series analysis for cyber security analytics: from intrusion detection to attack prediction. *International Journal of Information Security*, 24(1):3, 2025.
- [2] Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS one*, 12(7):e0180944, 2017.
- [3] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A Ghorbani. Developing realistic distributed denial of service (ddos) attack dataset and taxonomy. In *2019 international carahan conference on security technology (ICCST)*, pages 1–8. IEEE, 2019.
- [4] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2828–2837, 2019.
- [5] Robert H Shumway and David S Stoffer. Arima models. In *Time series analysis and its applications: with R examples*, pages 75–163. Springer, 2017.
- [6] Cynthia A Lowry and Douglas C Montgomery. A review of multivariate control charts. *IIE transactions*, 27(6):800–810, 1995.
- [7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [8] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.
- [9] Yuan Yuan Wei, Julian Jang-Jaccard, Fariza Sabrina, Wen Xu, Seyit Camtepe, and Aeryn Dunmore. Reconstruction-based lstm-autoencoder for anomaly-based ddos attack detection over multivariate time-series data. *CoRR*, abs/2305.09475, 2023.
- [10] Umaporn Yokkamon, Abbe Mowshowitz, Sakmongkon Chumkamon, and Eiji Hayashi. Robust unsupervised anomaly detection with variational autoencoder in multivariate time series data. *IEEE Access*, 10:57835–57849, 2022.
- [11] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks*, pages 703–716. Springer, 2019.
- [12] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*.
- [13] Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4027–4035, 2021.
- [14] Mengyao Li, Zhiyong Li, Zhibang Yang, Xu Zhou, Yifan Li, Ziyang Wu, Lingzhao Kong, and Ke Nai. Sa2e-ad: A stacked attention autoencoder for anomaly detection in multivariate time series. *ACM Trans. Knowl. Discov. Data*, 18(7), June 2024.
- [15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [16] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1):213, 2009.

Improving Targeted Password Guessing with Pattern-Guided Vector Quantization

1st Duc Minh Nguyen
Le Quy Don Technical University
HaNoi, VietNam
minhndmta700789@gmail.com

2nd Viet Hung Nguyen
Le Quy Don Technical University
HaNoi, VietNam
hungnv@lqdtu.edu.vn

Abstract—Targeted password guessing, which uses personally identifiable information (PII) to compromise specific accounts, remains a significant and evolving cybersecurity threat. While deep learning models excel at generating common passwords for large-scale trawling attacks, they struggle with personalized and structurally diverse passwords due to their reliance on common patterns and a lack of embedded structural awareness. To address these shortcomings, we introduce VQPassGPT, a Transformer-based framework designed specifically for targeted password guessing. This method combines pattern conditioning with dual-stage vector quantization, enabling the model to learn compact and semantically rich representations that capture the structure of PII-based passwords. Our evaluation demonstrates that VQPassGPT outperforms existing models in targeted guessing scenarios, showing particular strength in generating long and complex passwords. In a simulated limited-attempt guessing task, the model achieves a success rate of over 0.9 % against a set of realistic, user-generated passwords. These results highlight its effectiveness in scenarios where both guessing accuracy and generative diversity are critical.

Index Terms—VQPassGPT, Password Guessing, Pattern-Guided Generation, Vector Quantization.

I. INTRODUCTION

As technology progresses at an increasing pace, passwords continue to serve as a crucial means of user authentication and authorization on systems. While other possibilities exist, such as biometrics or multi-factor authentication, passwords remain because they are simple and can be used across platforms. Nonetheless, self-selected passwords often must follow known and easily guessed rules [1] and are even used on more than one account, thereby becoming a low-hanging fruit for guess-based prediction of passwords.

Password prediction techniques can be broadly categorized into two types: *trawling attacks* [2], which involve testing large sets of common or leaked passwords against many accounts; and *targeted attacks* [3], which focus on a specific individual by leveraging personal information such as their full name, date of birth, or email address prefix to guess their password. Although models designed for trawling, such as PassGAN [4], PassGPT [5], and PagPassGPT [6] have proven highly effective at modeling password distributions from large-scale leaks, they tend to struggle to generalize to less common or personalized password structures that are influenced by user-specific characteristics.

To mitigate these limitations, we propose VQPassGPT, an extension of PagPassGPT that integrates a vector quantization mechanism within the GPT-2 Transformer architecture [7]. Specifically, we embed two vector quantization (VQ) modules into the decoder-only network: one is placed after the initial token embeddings, and the second is inserted midway through the Transformer’s decoder stack. This mechanism quantizes continuous hidden representations into discrete codes, enabling the model to learn compact and semantically rich representations. This provides a more efficient means to learn complex and rare password patterns, particularly those influenced by personally identifiable. Following the autoregressive generation paradigm, the model estimates the conditional probability of a password sequence $x = (x_1, x_2, \dots, x_T)$ given a prefix pattern p by factorizing the joint probability as follows [5], [6]:

$$P(x | p) = \prod_{t=1}^T P(x_t | x_{<t}, p), \quad (1)$$

where $x_{<t}$ represents the previously generated tokens and p is a particular structural pattern such as L5N2 or L3S1N2 [6]. The pattern prefix gets prepended to the input string and guides the password randomization process.

Current models still lack the robustness required for targeted password guessing. User-chosen passwords often do not conform to population-wide patterns; instead, they deviate based on individual habits and personal information. While PagPassGPT [6] uses a pattern-conditioned generation approach to ensure structural alignment, it struggles to infer complex, personalized passwords derived from personally identifiable information (PII), representing a critical vulnerability in real-world attack scenarios.

To mitigate this gap, we propose VQPassGPT, a model that enhances the GPT-2 Transformer architecture by employing dual-stage vector quantization. Specifically, we augment the model by inserting two vector quantization (VQ) modules within its decoder stack, which discretize the internal latent representations. This quantization enables the model to learn compact yet expressive latent codes, allowing it to generalize more effectively to rare, complex, or PII-influenced passwords. Additionally, the VQ layers [8] promote generative diversity and reduce duplication, which is critical for modeling high-entropy password behaviors.

The remainder of this paper is organized as follows. Section II reviews related work on existing password guessing techniques and neural password generation. Section III details the proposed framework and the integration of vector quantization into the pattern-guided Transformer model. Section IV describes the experimental setup, evaluation metrics, and results. Finally, Section V concludes the paper and outlines potential directions for future work.

II. BACKGROUND AND RELATED WORK

A. Traditional Password Guessing Techniques

Traditional password guessing methods have long used rule-based dictionaries and probability models. Well-known techniques include Probabilistic Context-Free Grammars (PCFG) [9] and Markov models [10], which both try to find common password structures and patterns between characters. Tools like John the Ripper [11] and Hashcat [12] add to this with smart rules for changing words and methods for trying different character combinations. These methods are good at guessing simple and reused passwords, but they struggle to understand deeper meanings or adapt to the new ways people create passwords today.

B. Neural Password Generation Models

Deep learning models have been applied effectively in password generation. Models can learn directly from large datasets of leaked passwords. PassGAN [4] was one of the first models to use Generative Adversarial Networks (GANs) for this task, and it was better at guessing new passwords than older rule-based methods. However, GAN models [13] often have problems like "mode collapse" (getting stuck on a few password types) and lack of diversity in the passwords they create.

To fix these issues, PassGPT [5] used a GPT-2 Transformer model trained on real leaked passwords. This greatly improved the quality and variety of the generated passwords by predicting them one character at a time. Building on that, PagPassGPT [6] added a feature to guide the generation with structural patterns (like L5N2), to make sure the outputs have the correct format. PagPassGPT has some limitations, however. While it is good at copying very common passwords for large-scale trawling attacks, it struggles to create personalized or rare passwords, especially those that use personal information (PII) or are very complex.

C. Vector Quantization in Language Models

Vector Quantization (VQ) is a technique for converting complex internal data (latent representations) into simpler, discrete codes. It has been used successfully in areas like image generation [8] and speech recognition. In language models, VQ can help improve the structure and diversity of the generated text by learning more compact and meaningful representations. Models like VQ-VAE [14] and PassVQT [5] have shown that adding a VQ layer to a Transformer [15] can improve the quality and variety of the output. However, VQ has not been widely used for password generation. Therefore,

in our work, we use a two-stage VQ process within a GPT-2 Transformer decoder. With this change, the model can learn richer and more compact internal representations that better match the structure and meaning of real passwords.

D. Targeted Online Password Guessing

While large-scale data breaches get a lot of attention, targeted online password guessing is a serious and often underestimated threat. Unlike offline attacks where attackers can try endless guesses [11], [12], online services usually limit login attempts (for example, to fewer than 100 tries per account, as recommended by NIST [16]). Because of these limits, attackers must make their guesses count by generating passwords that are highly likely to be correct for a specific user. The TarGuess framework, introduced by Wang et al. [3] showed just how effective this can be. They found that by using basic personal info (like name, birthday, and email), they could guess a user's password with over 20.18% success in just 100 attempts. They also found that about 60% of users include personal information in their passwords, often in predictable formats like <name><birthyear> or <email-prefix><number>.

These results show why new models are needed. Models like PassGAN, PassGPT, and PagPassGPT are good at finding common passwords for large-scale attacks, but they lack the ability to understand the meaning and relationships needed for successful targeted guessing. For example, they can easily generate a common password like `sunshine2020`, but would fail to create `David1982!` — which is a much more likely guess for a user named David born in 1982.

We believe this problem exists because current models don't properly learn personalized, context-based structures [17]. To fix this, our proposed VQPassGPT improves its internal learning power by using vector quantization. This helps the model better capture rare but meaningful patterns, especially those using PII, while also creating more varied passwords with fewer duplicates. This makes it a strong tool for creating effective targeted guesses under real-world limits.

III. METHODOLOGY

A. Pattern-Guided Preprocessing and Transformer-Based Training

To address two key challenges in password generation, improving accuracy through pattern-conformant outputs [6] and minimizing repetition, we propose a model architecture comprising two main components, as illustrated in Fig. 1.

VQPassGPT is based on the GPT-2 [7] decoder-only Transformer, known for its generative capacity and open-source availability. Like PassGPT [5], it employs an autoregressive mechanism to compute conditional probabilities for password generation. While alternative architectures such as LSTMs [18], GANs [13], and VAEs [14] can model password sequences, GPT-2 is especially effective due to its attention-based structure and efficient parallel training.

The training process involves two main stages. First, a custom tokenizer extracts each password and its associated

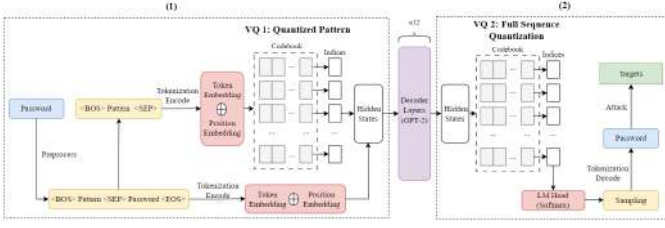


Fig. 1. Overview of the VQPassGPT architecture. The input password is first preprocessed to extract its structural pattern, forming a pattern-password sequence marked with special tokens (<BOS>, <SEP>, <EOS>). This sequence is tokenized and passed through two vector quantization stages: (1) VQ1 encodes the pattern segment into discrete latent codes; (2) after decoding, VQ2 quantizes the full hidden representation before final softmax output. The decoder stack (12 layers) autoregressively generates password tokens conditioned on the quantized pattern representations.

structural pattern, converting them into token indices. Unlike standard GPT-style models, VQPassGPT integrates two vector quantization layers. The first (VQ1) is applied to the pattern segment (before the <SEP> token) after token embedding. The second (VQ2) is applied to the entire sequence of decoder hidden states before output.

B. Model Architecture

VQPassGPT integrates two vector quantization (VQ) modules into the GPT-2 Transformer decoder [15]. These modules discretize latent representations at strategic points, enhancing compactness and diversity in password generation.

Quantization occurs in two stages: (1) VQ1 is applied to the embedded pattern segment (i.e., up to <SEP>), and (2) VQ2 is applied to the full decoded sequence. Each VQ module uses a learned codebook [8] to map continuous vectors to discrete indices by selecting the nearest entry via Euclidean distance. This process enables semantically meaningful clustering in the latent space, thereby improving the model’s ability to generalize to rare or user-specific passwords.

1) *Embedding and Decoder Stack*: First, each input token is transformed into a dense vector using a learned embedding function $E : \mathbb{Z} \rightarrow \mathbb{R}^d$. Positional information is captured through a trainable embedding matrix $P \in \mathbb{R}^{n \times d}$. The model’s input representation is created by summing these two embeddings:

$$h_0 = E(x) + P \quad (2)$$

where x is the input token sequence. This combined representation is then fed into a 12-layer Transformer decoder based on the GPT2Model [15].

2) *Vector Quantization Layers*: Our model integrates two independent VQ modules, VQ₁ and VQ₂, at two strategic points within the GPT-2 decoder stack. The first module quantizes the portion of the hidden states corresponding to the pattern prefix (i.e., tokens before and including <SEP>), while the second is applied to the full hidden states before they are passed to the output head.

Each quantization module contains a learnable codebook $C \in \mathbb{R}^{K \times d}$ with $K = 512$, a common choice in models

like VQ-VAE [8] that offers a good balance between expressiveness and efficiency. Given a latent vector z , the nearest codebook entry is selected based on Euclidean distance [19]:

$$k = \arg \min_j \|z - e_j\|_2^2, \quad z_q = e_k \quad (3)$$

where $e_k \in C$ is the chosen code vector. To ensure differentiability during training, the model uses the stop-gradient approach:

$$\tilde{z} = z + (z_q - z).detach(). \quad (4)$$

This allows the quantized vector $\tilde{z}$ to be used in the forward pass while enabling gradients to flow through the operation during the backward pass.

3) *Output head*: After decoding, the final hidden states are passed through a linear output layer $W \in \mathbb{R}^{d \times |V|}$ and a softmax function to obtain next-token probabilities:

$$p(y_t | x_{<t}) = \text{Softmax}(W \cdot h_t). \quad (5)$$

Password generation is performed autoregressively, conditioned on a structural pattern prefix and any previously generated tokens.

C. Training Objective

The total loss function combines the standard cross-entropy loss for next-token prediction with the quantization losses from two vector quantization modules, VQ₁ and VQ₂, integrated at distinct locations in the model architecture.

Following [8], the quantization loss for each VQ module is defined as:

$$\mathcal{L}_{VQ} = \underbrace{\|sg[z] - e\|_2^2}_{\text{Codebook Loss}} + \beta \cdot \underbrace{\|z - sg[e]\|_2^2}_{\text{Commitment Loss}}, \quad (6)$$

where z is the latent vector (either from the pattern embedding or decoder output), e is the nearest codebook entry, and $sg[\cdot]$ denotes the stop-gradient operator. The first term (codebook loss) updates the codebook embeddings to match the model’s latent representations, while the second term (commitment loss) encourages the latent outputs to remain close to their assigned codebook entries. We set $\beta = 0.25$ to maintain a balanced influence between codebook adaptation and representation commitment.

Specifically, $\mathcal{L}_{VQ1}$ is computed over the pattern segment (i.e., up to the <SEP> token) immediately after token embedding, while $\mathcal{L}_{VQ2}$ is computed over the full hidden sequence after the decoder.

The final training objective is therefore defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda(\mathcal{L}_{VQ1} + \mathcal{L}_{VQ2}), \quad (7)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for language modeling, and $\lambda = 1.0$ [8] is a weighting factor.

D. Implementation Notes

The model is built upon the GPT2LMHeadModel [15] from the Hugging Face Transformers library. The quantization modules are implemented in PyTorch using `nn.Embedding` for the codebook and an efficient Euclidean distance search. Efficient generation is supported through caching of past key-value states. The architecture is fully compatible with pattern-guided generation. While VQ_1 captures high-level structural patterns from the prefix input (e.g., pattern format), VQ_2 encourages diversity and compactness across the entire generated sequence.

IV. EXPERIMENT AND EVALUATION

A. Datasets and Experimental Setup

This section describes the datasets [6], [20] and methodology used for our evaluation, covering the preprocessing procedure and the construction of a specialized test set for targeted attacks. All models in this study are trained on the same large-scale dataset to ensure a fair comparison.

To better understand the characteristics of leaked passwords, we analyzed PII-related patterns in five commonly used datasets. Table I summarizes the proportion of passwords containing attributes such as names, birth years, phone numbers, and email prefixes. These statistics informed the construction of our training data and targeted test set. We identified PII patterns based on matching tokens for attributes such as abbreviated birth dates (`abbr_bd`), names (`abbr_name`), full dates of birth (`dob`), and others. Combinations like `name+dob` and `name+year` represent semantically meaningful structures often found in personalized passwords.

TABLE I
PREVALENCE OF PII-RELATED ATTRIBUTES IN LEAKED PASSWORD DATASETS (%)

Dataset	abbr_bd	abbr_name	dob	email	name	name+dob	name+year	phone	yob	others	Total
RockYou	2.80	32.32	0.03	0.51	29.77	1.29	2.09	0.43	0.12	95.68	14,344,373
LinkedIn	1.37	26.26	0.05	1.03	24.74	1.84	3.41	0.23	0.21	94.76	47,915,722
phpBB	4.22	45.42	0.09	0.18	40.49	0.50	1.30	0.13	0.05	98.52	184,388
MySpace	0.27	14.16	0.02	0.18	11.15	0.92	1.51	0.01	0.02	93.10	36,228
Yahoo	0.03	18.13	0.00	0.00	21.74	0.00	0.01	0.00	0.00	91.79	3,745,855

To ensure consistency, we applied the same preprocessing to all datasets: duplicate passwords were removed, entries shorter than 4 or longer than 15 characters were excluded, and non-ASCII or control-character-containing strings were filtered out. The resulting dataset comprises 63,378,468 unique passwords and was used for training all models without splitting.

Beyond standard evaluation, we constructed a separate test suite for targeted guessing, following the TarGuess framework [3]. As shown in Fig. 2, we extracted structured PII attributes names, usernames, phone numbers, and birth-dates and used them to generate realistic password patterns (e.g., `name+year`, `name+123`) per TarGuess-I and III. For TarGuess-II, we identified sister-password pairs across datasets and applied common transformations such as truncation or character substitution.

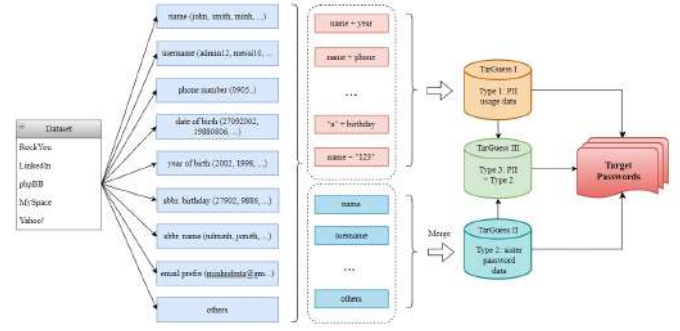


Fig. 2. Overview of the Target Password Generation Pipeline based on PII and sister-password data. The process extracts user-specific features from real-world datasets and constructs candidate password guesses using TarGuess I, II, and III frameworks [3].

This custom test suite realistically simulates an attacker performing a brute-force attack on specific targets with a limited number of attempts [16] (e.g., 100 per user). This benchmark allows for a detailed analysis of how well the models generalize beyond common passwords.

To better illustrate the structure of these personalized passwords, Table II presents examples of password targets generated using common PII-based patterns. These samples reflect real-world guessing strategies that use attributes like names, birth dates, and email prefixes.

TABLE II
EXAMPLES OF TARGET PASSWORDS BY PII-BASED PATTERNS

#	name+year	name+date	email prefix+phone	name+123
1	bagong1972	ssvadq0524	rhontin_0057	nicotin123
2	superbubble1975	combug0103	deadlylove4619	marysia123
3	chaquetita1991	peter0115	nicolegk_2000	baron123
4	madrep1989	cacosca0922	tangruknu5031	jessica123
5	josekike2003	peckenpaugh0204	berna_6730	dolnad123

The training configurations for VQPassGPT are detailed in Table III. The model was implemented using the Hugging Face Transformers library [15] and the GPT-2 architecture. The training process for VQPassGPT took approximately 9 days on a single RTX 3080 GPU. This is longer than PagPassGPT [6] primarily due to the additional computation required for the dual-stage vector quantization during the model's forward and backward passes. These configurations were chosen to balance model complexity, representational power, and computational efficiency.

B. Evaluation Metrics

We evaluate VQPassGPT against state-of-the-art models PCFG [9], OMEN [21], PassGAN [4], PassGPT [5], and PagPassGPT [6] using consistent training data and original configurations. Accuracy is measured by password length (7–15 characters) to assess the ability to generate complex, user-specific passwords. We also compute duplication rates under the Divide-and-Conquer Generation (D&C GEN) strategy [6] with $T = 5000$, reflecting redundancy in large-scale outputs. Lastly, Shannon entropy [22] is used to quantify

TABLE III
COMPARISON OF TRAINING CONFIGURATIONS: VQPASSGPT VS. PAGPASSGPT

Parameter	VQPassGPT	PagPassGPT
Batch size	512	512
Number of epochs	50	30
Optimizer	AdamW	AdamW
Initial learning rate	5×10^{-5}	5×10^{-5}
Maximum input tokens	32	32
Embedding size	384	384
Transformer layers	12	12
Attention heads/layer	8	8
Pattern conditioning	Prefix + Quantization	Prefix
Quantization	Dual-stage VQ	No
Library	GPT-2	GPT-2
Training hardware	RTX 3080	RTX 3080
Training time	~9 days	~5 days
Operating system	Linux	Linux

password complexity based on character distribution, where higher entropy indicates greater unpredictability. The entropy $H(X)$ of a password X is calculated as:

$$H(X) = - \sum_{i=1}^n p(x_i) \cdot \log_2 p(x_i) \quad (8)$$

where x_i represents the i -th unique character in the string, $p(x_i)$ refers to the frequency of occurrence of the character, and n is the number of distinct characters. The overall entropy is estimated by multiplying $H(X)$ by the password length.

C. Results and Discussion

Both models have been evaluated against the same test set. As shown in Figure 3, PagPassGPT is accurate for short passwords (length 7–8) for names or dates of birth; however, the performance declined as cases became longer and complex. On the contrary, VQPassGPT performed better with cases that used ≥ 9 or had higher entropy, as it is able to model some semantic structure and PII-based patterns. The use of vector quantization improves the model’s ability to generalize compared to non-quantized approaches. It is clear VQPassGPT has better overall performance within the testing dataset of targeted guessing cases, especially for high-entropy passwords that typically require more advanced heuristic-based guessing techniques to crack.

In the targeted password guessing scenario (Table IV), traditional probabilistic models such as PCFG and OMEN perform poorly, achieving only 0.0004% and 0.3826% accuracy respectively at 10^8 guesses. Deep learning models improve upon this: PassGAN reaches 0.0266%, PassGPT 0.3427%, and PagPassGPT slightly higher at 0.3745% with pattern-guided generation. However, all three remain limited in handling rare or personalized passwords.

VQPassGPT achieves the best results, with 0.9269% at 10^8 guesses and consistent gains across all dictionary sizes from 0.0008% at 10^5 to 0.0648% at 10^7 . This performance stems from the integration of dual-stage vector quantization,

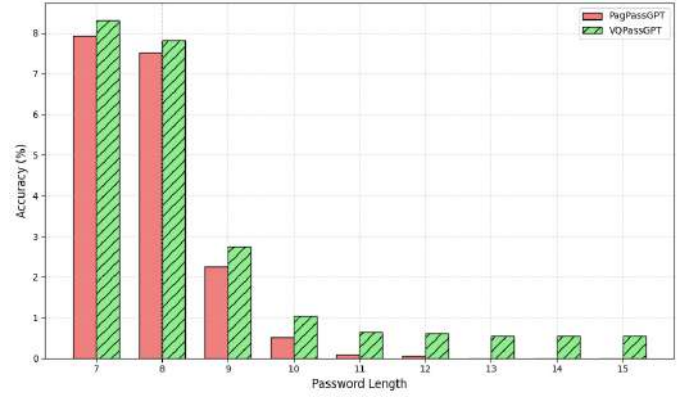


Fig. 3. Accuracy comparison by password length between PagPassGPT and VQPassGPT on the target password set. VQPassGPT achieves better performance on longer and personalized passwords, indicating enhanced generalization to structures influenced by PII.

TABLE IV
TARGET PASSWORD GUESSING ACCURACY (%) AT DIFFERENT DICTIONARY SIZES

Model	10^5	10^6	10^7	10^8
<i>Probabilistic Models</i>				
PCFG	0.0001	0.0004	0.0004	0.0004
OMEN	0.0005	0.0035	0.0586	0.3826
<i>Deep Learning Models</i>				
PassGAN	0.0001	0.0005	0.0043	0.0266
PassGPT	0.0005	0.0046	0.0445	0.3427
PagPassGPT	0.0005	0.0047	0.0447	0.3745
VQPassGPT	0.0008	0.0059	0.0648	0.9269

applied to both pattern embeddings and decoder outputs. By learning discrete latent codes, the model captures semantically meaningful structures more effectively, enabling robust generalization to high-entropy and user-specific passwords.

As shown in Fig. 4, when applying the Divide-and-Conquer (D&C GEN) strategy [6] with a threshold of $T = 5000$, the duplicate rates vary notably between models. PagPassGPT exhibits a relatively stable duplication rate, increasing slightly from 0.308% at 10^5 to 0.353% at 10^8 outputs. In contrast,

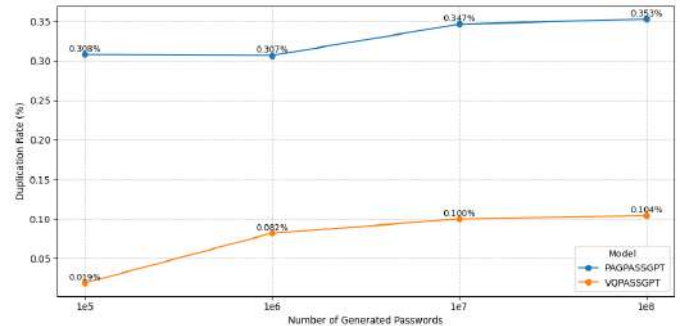


Fig. 4. Duplicate Rate of Generated Passwords by PagPassGPT and VQPassGPT using D&C generation with $T = 5000$

TABLE V
COMPARISON OF GENERATED PASSWORD COMPLEXITY BETWEEN
PAGPASSGPT AND VQPASSGPT

Metric	PagPassGPT	VQPassGPT
Average Length	7.63	8.64
Average Entropy (Shannon)	19.68 bits	23.90 bits
% Passwords with Length ≥ 8	75.04%	90.21%
% Passwords with Entropy ≥ 50 bits	7.04%	18.74%

VQPassGPT maintains a much lower rate, rising only from 0.019% to 0.104%, even as the output scale increases.

This reduction in redundancy is a result of the dual vector quantization layers integrated into VQPassGPT. By discretizing internal representations, the model encourages diversity in the latent space and avoids overfitting to high-frequency patterns. PagPassGPT, lacking such mechanisms, is more susceptible to repeated outputs in large-scale generation.

As shown in Table V, VQPassGPT produces longer and more complex passwords, with an average length of 8.64 characters and entropy of 23.90 bits. It also increases the proportion of high-entropy passwords (entropy ≥ 50 bits) from 7.04% to 18.74%, and over 90% of its outputs exceed 8 characters, compared to 75% for PagPassGPT. While the model does not inject pattern tokens at every step, the latent VQ layers effectively capture and promote structural diversity — an important factor in targeted attacks [3], where generating user-specific formats improves the chances of success.

V. CONCLUSION AND FUTURE WORKS

In this study, we focused on the growing criticality of targeted password guessing attacks, particularly considering the increasing exploitation of personal information (PII) by adversaries to compromise user accounts. While previous neural models, such as PagPassGPT, have made significant strides in global password guessing, they still have difficulties with complex, personalized passwords. This is particularly the case for passwords that incorporate user-specific PII and follow personalized or context-specific composition rules.

We proposed VQPassGPT, a Transformer-based model with dual-stage vector quantization, improving the generation of structurally rich and semantically meaningful passwords. We have shown that VQPassGPT outperforms any of the existing GPT-2-based models in targeted guessing, especially for longer or high-entropy passwords. The utility of VQPassGPT can also help estimate password strength and aid security audits.

In the future, we plan to investigate: (1) pattern-conditioned generation for individual users; (2) multi-codebook quantization to enhance latent representation; and (3) systematic guessing mechanisms for low-resource settings. We also aim to adapt the model for prefix-based generation involving unseen password structures with quantization-aware methods.

ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Science and Technology of Vietnam for their support provided through the

project “Multidimensional detection and automated response using artificial intelligence” (code NDT/CZ/24/01).

REFERENCES

- [1] Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano, “The quest to replace passwords: A framework for comparative evaluation of web authentication schemes,” in *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, IEEE, 2012, pp. 553–567.
- [2] A. Narayanan and V. Shmatikov, “Fast dictionary attacks on passwords using time-space tradeoff,” in *Proc. 12th ACM Conf. on Computer and Communications Security (CCS)*, pp. 364–372, 2005.
- [3] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang, “Targeted online password guessing: An underestimated threat,” in *Proc. 2016 ACM SIGSAC Conf. on Computer and Communications Security*, pp. 1242–1254, 2016.
- [4] W. Li, L. Xu, Y. Zhang, and H. Zhu, “PassGAN: A generative adversarial network for password generation,” in *Proc. 2017 ACM SIGSAC Conf. on Computer and Communications Security*, 2017.
- [5] J. Rando and F. Pérez-Cruz, “PassGPT: Password modeling and (guided) generation with large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.08518>
- [6] X. Su, X. Zhu, Y. Li, Y. Li, C. Chen, and P. Esteves-Verissimo, “Pag-PassGPT: Pattern Guided Password Guessing via Generative Pretrained Transformer,” in **Proc. 2024 54th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks (DSN)**, pp. 429–442, 2024.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.
- [8] L. D. Lingle, “Transformer-VQ: Linear-Time Transformers via Vector Quantization,” **arXiv preprint arXiv:2309.16354**, Sep. 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2309.16354>
- [9] M. Weir, S. Aggarwal, M. Collins, and H. Stern, “Password cracking using probabilistic context-free grammars,” in *Proc. 30th IEEE Symposium on Security and Privacy (S&P)*, pp. 391–405, 2009.
- [10] “Markov generator,” Openwall, 2017. [Online]. Available: <http://openwall.info/wiki/john/markov>
- [11] “John the Ripper,” Openwall, 2017. [Online]. Available: <http://www.openwall.com/john/>
- [12] “Hashcat Rules,” GitHub Repository, 2017. [Online]. Available: <https://github.com/hashcat/hashcat/tree/master/rules>
- [13] B. Vi, N. Tran, and T. The, “A GAN-based approach for password guessing,” in *Proc. RIVF Int. Conf. on Computing and Communication Technologies (RIVF)*, Hanoi, Vietnam, 2021.
- [14] K. Yang, X. Hu, Q. Zhang, J. Wei, and W. Liu, “VAEPass: A lightweight password guessing model based on variational autoencoder,” *Proc. 7th Int. Conf. on Cryptography, Security and Privacy (CSP)*, 2022.
- [15] Hugging Face, “Transformers Documentation,” [Online]. Available: <https://huggingface.co/docs/transformers/index>. [Accessed: July 23, 2025].
- [16] P. A. Grassi, M. E. Garcia, and J. L. Fenton, “Digital Identity Guidelines: Authentication and Lifecycle Management,” NIST Special Publication 800-63B, Jun. 2017. [Online]. Available: <https://doi.org/10.6028/NIST.SP.800-63b>
- [17] M. Scanlon, R. Kong, E. Nicholas, E. Lam, and M. L. Hale, “Context-Aware Password Guessing Using Generative Pretrained Transformers,” in **Proc. 36th IEEE Computer Security Foundations Symposium (CSF)**, 2023. [Online]. Available: <https://markscanlon.co/papers/ContextPasswordCrackingUsingGPTs.pdf>
- [18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [19] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [20] “The Password Project: Leaked password lists and dictionaries,” 2017. [Online]. Available: http://thepasswordproject.com/leaked_password_lists_and_dictionaries
- [21] M. Duermuth, F. Armknecht, C. Castelluccia, and D. Perito, “OMEN: Faster password guessing using an ordered Markov enumerator,” in *Proc. 17th European Symposium on Research in Computer Security (ESORICS)*, pp. 569–588, 2012.
- [22] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

Pri-Inv: A Novel Privacy-Preserving Digital Forensic Solution for Face Presence Verification in Criminal Investigations

1st Anh-Tu Tran

*Department of Information Security
Academy of Cryptography Techniques
Hanoi, Vietnam
tutran@actvn.edu.vn*

2nd Xuan-Hung Truong

*Cyber Security and High-Tech Crime Prevention Department
Ministry of Public Security; and
Academy of Cryptography Techniques
Hanoi, Vietnam
hungtx.ncs@actvn.edu.vn*

Abstract—Determining whether a known criminal suspect appears in personal photo collections seized from a target device can provide critical leads in digital forensic investigations. This task, known as targeted face presence search, focuses on verifying the occurrence of a specific face within a private image repository, without revealing other identities or sensitive image content. However, conducting such searches on private data raises serious privacy concerns, particularly when non-relevant individuals are present in the image dataset. In this work, we introduce Pri-Inv, a privacy-preserving forensic framework that enables secure face presence verification using homomorphic encryption (HE). Pri-Inv allows law enforcement agencies to query encrypted face representations of known suspects against encrypted facial features extracted from images on a seized device—without ever decrypting the underlying data during the computation process. By leveraging the properties of homomorphic encryption, the system ensures that neither the content of the images nor the identities of non-target individuals are exposed throughout the matching process. Our experimental evaluation demonstrates that Pri-Inv achieves high detection accuracy while maintaining strong cryptographic privacy guarantees and practical runtime performance. This work presents a viable and ethically sound solution for secure face-based investigations in privacy-sensitive digital forensic scenarios.

Index Terms—Homomorphic encryption, Privacy-preserving forensic, Image matching, Digital investigation, Digital forensic

I. INTRODUCTION

In the digital era, the proliferation of cybercrime and digital evidence has necessitated the development of sophisticated forensic tools capable of extracting meaningful information from personal devices while upholding individual privacy [1]. Modern smartphones, computers, and cloud-connected devices often store large volumes of personal data, including images that may serve as critical evidence in criminal investigations [2]. Among the various forensic tasks [3], one fundamental problem is the Targeted Face Presence Search: determining whether a specific, known criminal appears in the image collection of a suspect's seized device [4], [5].

This task differs from general-purpose face recognition or social association inference. It is narrowly focused on verify-

ing the presence or absence of a known face (e.g., a criminal from a law enforcement database) in a private image dataset, without revealing the identities of non-relevant individuals or exposing the image contents themselves. While this targeted approach provides valuable leads—such as establishing physical proximity or prior contact between individuals—it must be executed in a way that respects the privacy of those not involved in the investigation.

Traditionally, face matching in digital forensics has relied on comparing raw biometric data or image features, which risks exposing sensitive and unrelated personal information. This raises serious ethical and legal concerns, especially when processing images that include family members, friends, or other uninvolved parties. The core challenge lies in enabling secure comparison between private image data and reference criminal databases without decrypting or exposing the underlying information during the matching process.

To address this, we propose Pri-Inv, a novel privacy-preserving forensic framework that leverages Homomorphic Encryption (HE) to perform secure face search in digital forensic. Homomorphic encryption is a cryptographic technique that allows computations to be performed directly on encrypted data [6]. In the context of forensic analysis, this means facial features extracted from both the suspect's images and the law enforcement database can remain encrypted throughout the matching process. Only the final result—indicating whether the known criminal appears in any image—is revealed to investigators, while all other information remains confidential.

Our proposed system enables law enforcement to identify whether a target face appears in a suspect's photo collection, without ever decrypting the facial representations or accessing raw images. In contrast to earlier approaches that either lack formal privacy guarantees or suffer from inefficiencies when applied to large-scale datasets, Pri-Inv is designed for both privacy protection and scalability. It minimizes computational overhead while ensuring compliance with legal standards for data protection.

To validate the effectiveness of the approach, we conduct experiments using the Labeled Faces in the Wild (LFW)

The article was completed with support from the Academy of Cryptography Techniques and Ministry of Public Security.

dataset [7], comprising 1,000 image pairs evenly split between positive (same-person) and negative (different-person) samples. The evaluation demonstrates that Pri-Inv can perform accurate encrypted face matching while maintaining strict privacy guarantees and feasible computational cost.

In summary, this work addresses a critical need in digital forensics by introducing a secure and practical solution for targeted face presence search using homomorphic encryption. It contributes to the growing body of research on privacy-preserving forensic technologies, offering a scalable and ethically aligned method for image-based suspect verification in criminal investigations.

II. RELATED WORK

Research in privacy-preserving biometrics and secure face recognition has evolved significantly in recent years, driven by the increasing need for secure and responsible analysis of sensitive data, particularly in digital forensic contexts. Our work intersects with three major research domains: forensic face presence detection, Face Recognition Techniques, and homomorphic encryption for encrypted-domain computation.

A. Face Presence Detection in Forensics

Face recognition technologies have long been employed by law enforcement agencies for tasks such as suspect identification, surveillance, and forensic triage. Traditional face search systems match a probe image against a large database to identify potential matches [8], while recent efforts have focused on co-occurrence and relational analysis in group photos to uncover links between individuals [9]. However, these approaches often operate on unencrypted data, risking exposure of unrelated personal identities during forensic analysis.

Our work addresses a more targeted task—face presence verification—which determines whether a known individual (e.g., a registered criminal) appears in a set of personal photos, such as those retrieved from a suspect’s device. In contrast to full-scale face identification or associative search, our approach is narrowly scoped to reveal only whether the target individual is present, without disclosing the content or identities of other individuals in the dataset. This design is motivated by privacy concerns in forensic workflows and distinguishes our work from broader face recognition applications.

B. Biometric Template Protection

A wide array of methods has been proposed to secure biometric templates against misuse, typically categorized into cryptographic and pattern-recognition-based approaches. Cryptographic biometric template protection includes early works on biometric encryption, visual cryptography for biometric images, and fuzzy vault constructions, which allow for approximate matching in the encrypted domain [10]. While these methods provide strong privacy assurances, they often face trade-offs in accuracy and utility, particularly when handling real-valued face embeddings from modern deep learning models.

Various pattern recognition-based approaches have been proposed to enhance biometric data security, including cancelable biometrics, non-invertible transformations, and biometric key-binding techniques. These methods aim to support revocability and diversity by transforming the original biometric data or binding it to a cryptographic key. While effective in some scenarios, they often involve trade-offs between recognition accuracy and security strength. Recent surveys [11], [12] highlight that many of these schemes still face challenges such as vulnerability to linkage attacks or insufficient protection under active adversaries.

In contrast, our method avoids these trade-offs by leveraging fully homomorphic encryption (FHE), which maintains both strong privacy and high matching accuracy, even when working with high-dimensional face templates derived from modern neural representations.

C. Face Recognition Techniques

Modern face recognition systems have achieved remarkable performance owing to advances in deep learning, particularly convolutional neural networks (CNNs). Unlike earlier methods based on handcrafted features (e.g., Gabor filters, LBP), current systems rely on learned feature embeddings that capture identity-specific characteristics in a compact, discriminative space [13].

State-of-the-art frameworks such as FaceNet [14], SphereFace [15], ArcFace [16], VGGFace [17] and CosFace [18] generate high-dimensional vector representations (e.g., 128–512 dimensions) from raw face images, enabling highly accurate face verification and recognition. These embeddings form the basis of most biometric authentication pipelines and have been widely adopted in both academic and industry settings.

In our proposed framework, we leverage the representational power of such models—specifically, FaceNet and VGGFace—to extract face templates from both the probe and reference images. These embeddings are then compared securely in the encrypted domain using fully homomorphic encryption. Our approach demonstrates that high-accuracy deep feature representations can be effectively and efficiently used in privacy-preserving face presence verification, without requiring decryption at any stage.

Importantly, while these deep learning models offer high utility, integrating them with cryptographic techniques (like FHE) presents unique challenges in terms of data representation, dimensionality, and computational overhead. Our contribution lies in bridging this gap by optimizing the encoding and matching of deep face features within the homomorphic encryption pipeline, ensuring both accuracy and privacy.

D. Homomorphic Encryption for Face Matching

Homomorphic encryption allows arithmetic operations to be performed directly on encrypted data, making it particularly appealing for privacy-preserving biometric verification. Early works in this space used partial homomorphic encryption (PHE) schemes, such as the Paillier cryptosystem [19], for

secure fingerprint or iris matching. However, these approaches either required repeated client-server interactions [20] or provided no protection to the biometric database.

More recently, researchers have turned to somewhat and fully homomorphic encryption (FHE) for secure biometric computation. Ghostshell [21], for example, demonstrated binary iris matching using a somewhat homomorphic encryption scheme, but remained limited to binary template structures and incurred moderate computational cost (e.g., 0.5 seconds per comparison).

A breakthrough came with Gentry's FHE scheme [22], which enabled general-purpose computation on encrypted data. Troncoso-Pastoriza et al. [23] applied this to face matching using Gabor features, but their system was computationally expensive (e.g., 380 MB per template and 100 seconds per match). To address such inefficiencies, our framework builds on the Fan-Vercauteren (FV) FHE scheme [24], incorporating batching techniques that allow encrypted dot product computations with significantly lower overhead (e.g., 66 KB template size and 0.01 seconds per match). These improvements make our method practical for real-world forensic deployments.

Moreover, our system integrates face templates generated by state-of-the-art deep learning models, such as FaceNet and VGGFace, and performs all similarity computations directly on encrypted embeddings using the CKKS homomorphic encryption scheme [25]. This design ensures compatibility with high-accuracy face representation methods while maintaining end-to-end data confidentiality in the encrypted domain.

III. PRI-INV: A PRIVACY-PRESERVING DIGITAL FORENSIC APPROACH FOR SECURE IMAGE MATCHING

This section outlines the methodology of the Pri-Inv framework, a privacy-preserving digital forensic solution that utilizes HE for secure image matching in criminal investigations. The methodology focuses on ensuring data privacy while allowing accurate image comparison between suspect images and criminal databases.

A. Overview of Pri-Inv Framework

This section presents the methodology of the proposed Pri-Inv framework, which integrates deep learning-based facial feature extraction and Fully Homomorphic Encryption (FHE) to achieve privacy-preserving face matching for criminal investigations. The methodology is structured into four main phases: data acquisition and embedding generation, encryption, privacy-preserving matching, and result verification.

The proposed framework is designed to enable investigators to perform secure face matching while ensuring that sensitive biometric data remains confidential throughout the entire process. The system consists of three primary components: (i) a forensic terminal equipped with deep learning models and an FHE module, (ii) a facial embedding database containing preprocessed and normalized embeddings of known individuals, and (iii) a secure verification module responsible for decrypting results only when legally authorized.

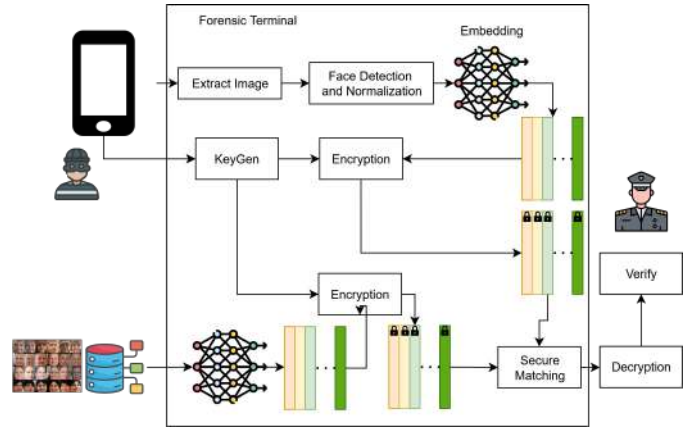


Fig. 1. Overview of the Pri-Inv investigation procedure

The incorporation of deep learning-based embedding extraction significantly improves the accuracy and robustness of facial feature representation. Meanwhile, the utilization of FHE ensures that all computations, including similarity evaluation, are performed entirely in the encrypted domain, thereby eliminating any potential risk of data leakage. The overall investigation procedure is illustrated in Figure 1.

1) *Data Acquisition and Embedding Generation*: During the investigation, the suspect is required to connect their personal device to the forensic terminal. The terminal incorporates advanced facial image processing models that perform four essential pre-processing steps:

- **Image Detection and Preprocessing**: Facial images are detected and extracted from available image files. Non-facial regions and background noise are removed.
- **Face Normalization**: Detected facial regions are normalized to a fixed resolution and standardized format to ensure compatibility with the embedding model.
- **Deep Learning-based Embedding Extraction**: A state-of-the-art convolutional neural network, such as FaceNet [14] or VGGFace [17], is used to generate high-dimensional facial embeddings that capture discriminative features while discarding redundant visual information.
- **Vector Encryption**: Following embedding extraction, the Embedding Encryption stage ensures the confidentiality of biometric features. To initiate the encryption process, the suspect is prompted to input a password, which is used to generate a unique FHE key. Both the suspect's embeddings and the embeddings stored in the database are encrypted using this key to ensure that subsequent computations are performed exclusively in the encrypted domain. The extracted embeddings are encrypted using FHE schemes such as BFV or CKKS [25], which allow for arithmetic operations to be performed directly on encrypted data. Compared to encrypting raw pixel-level images, encrypting only the compact embeddings significantly reduces the computational and storage overhead while retaining high matching accuracy.

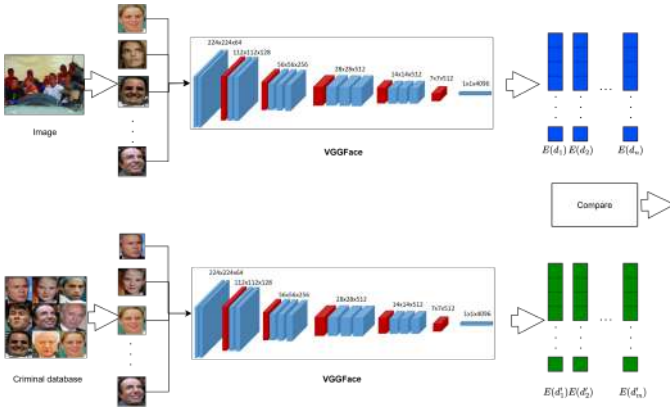


Fig. 2. Encrypted embedding matching process using VGGFace-based facial feature extraction and FHE

2) *Privacy-Preserving Matching*: In the Encrypted Embedding Matching stage, similarity computations (e.g., cosine similarity or Euclidean distance) are performed homomorphically, ensuring that all operations are executed entirely in the encrypted domain. Consequently, neither the system nor any external party can access the plaintext embeddings, thereby eliminating the risk of privacy breaches. Only the final matching score is decrypted to determine whether the suspect's facial features correspond to entries stored in the database. This matching process is illustrated in Figure 2, in which the embedding vectors are generated using the VGGFace model [17].

3) *Result Verification*: If a match is identified with a similarity score exceeding a predefined threshold, the corresponding encrypted image is isolated and stored in a secure, access-controlled partition. To ensure lawful verification, the investigator requests the suspect to re-enter the password to decrypt only the matched image. This selective decryption aligns with the principle of data minimization and ensures that only relevant evidence is exposed for judicial examination.

The proposed methodology effectively balances investigative efficiency and privacy protection. By combining deep learning-based embeddings with FHE-based encrypted matching, the Pri-Inv framework demonstrates its potential as a practical and legally compliant solution for privacy-preserving biometric analysis in forensic investigations.

B. Homomorphic Encryption with CKKS

To enable secure and privacy-preserving face matching, we adopt the CKKS (Cheon–Kim–Kim–Song) homomorphic encryption scheme [25], which is specifically designed for approximate arithmetic over real (or complex) numbers. Unlike conventional fully homomorphic encryption (FHE) schemes such as BFV or BGV that operate over integers, CKKS supports approximate computations on floating-point vectors, making it particularly well suited for encrypted machine learning applications, such as face searching based on embedding similarity.

CKKS operates over a polynomial ring $R_q = \mathbb{Z}_q[x]/(x^n + 1)$, where n is a power of two representing the ring degree, and q is the ciphertext modulus. Real-valued plaintext vectors are first scaled by a factor $\Delta \in \{2^{30}, \dots, 2^{40}\}$ to convert them into fixed-point representation before encoding into polynomials in R_q . This scaling helps preserve numerical precision during homomorphic operations.

A salient feature of CKKS is its support for *SIMD-style packing*, where multiple plaintext values are encoded into a single ciphertext, with each occupying a separate slot. This enables efficient parallel computation on vectors using slot-wise operations, such as addition and multiplication.

Given two embeddings $x, y \in \mathbb{R}^d$, the cosine similarity is defined as:

$$\text{sim}(x, y) = \frac{x^\top y}{\|x\| \|y\|} = \sum_{i=1}^d \tilde{x}_i \tilde{y}_i, \quad (1)$$

where $\tilde{x} = x/\|x\|$ and $\tilde{y} = y/\|y\|$ are the normalized embeddings. The corresponding cosine distance is then $d(x, y) = 1 - \text{sim}(x, y)$.

In the encrypted setting, both vectors are independently encrypted using CKKS, with their components packed into ciphertext slots. The similarity is computed by performing a homomorphic inner product between the two encrypted vectors. Although CKKS introduces small approximation errors due to its fixed-point arithmetic, the accuracy is sufficient for high-quality face matching, as demonstrated in our experiments.

C. Encrypted Cosine Similarity

We implement the cosine similarity computation entirely in the encrypted domain using the CKKS scheme. The overall protocol involves three main steps: homomorphic element-wise multiplication, encrypted slot summation, and decryption.

Let $\text{Enc}(\tilde{x})$ and $\text{Enc}(\tilde{y})$ denote the CKKS-encrypted, normalized embeddings, each packed into a single ciphertext. The encrypted cosine similarity is computed as follows:

1) *Homomorphic Hadamard Product*: We first compute the slot-wise (element-wise) product of the two ciphertexts:

$$c_{\text{prod}} = \text{Enc}(\tilde{x}) \odot \text{Enc}(\tilde{y}), \quad (2)$$

where each slot i in c_{prod} contains $\tilde{x}_i \cdot \tilde{y}_i$.

2) *Encrypted Slot-wise Summation*: To compute the inner product, we aggregate the slot values homomorphically using a sequence of left rotations and additions:

$$c_{\text{sum}} = \sum_{j=0}^{\log_2 d - 1} \text{RotLeft}(c_{\text{prod}}, 2^j), \quad (3)$$

where $\text{RotLeft}(\cdot, k)$ denotes a cyclic left rotation of ciphertext slots by k positions. This recursive doubling strategy ensures that all slot values are summed into slot 0.

3) *Decryption*: Finally, the client decrypts the result to retrieve the approximate cosine similarity:

$$\text{sim}(x, y) \approx \text{Dec}(c_{\text{sum}})[0], \quad (4)$$

where only the first slot is used; the remaining slots are either discarded or contain redundant/rotated values.

This method performs the entire similarity computation in a homomorphic manner, ensuring that only the final matching scores are revealed. By leveraging CKKS packing and SIMD-style processing, the scheme significantly reduces the number of ciphertexts and amortizes the cost of vector operations. While slot rotations introduce additional overhead, they are essential for slot-wise summation and do not substantially impact performance for moderate embedding dimensions (e.g., $d = 128$ or $d = 512$). Importantly, the protocol avoids the need for bootstrapping, as the required multiplicative depth remains low. Overall, this approach enables efficient, accurate, and privacy-preserving face matching with minimal interaction and strong confidentiality guarantees.

The cosine similarity-based face matching procedure is detailed in Algorithm 1:

Algorithm 1: Face Searching over Encrypted Embeddings

Input : Query embeddings $Q = \{q_1, \dots, q_m\} \subset \mathbb{R}^d$,
 database embeddings $D = \{d_1, \dots, d_n\} \subset \mathbb{R}^d$
 Similarity threshold $\tau \in [0, 1]$

Output: For each $q_i \in Q$, return
 $\mathcal{M}_i = \{j \mid \text{sim}(q_i, d_j) \geq \tau\}$

```

1 Client:
2 for  $i \leftarrow 1$  to  $m$  do
3   Normalize  $q_i$ :  $\tilde{q}_i \leftarrow q_i / \|q_i\|$ 
4   Encrypt:  $c_i \leftarrow \text{Enc}_{\text{CKKS}}(\tilde{q}_i)$ 
5   Send  $\{c_1, \dots, c_m\}$  to server;
6 Server:
7 for  $i \leftarrow 1$  to  $m$  do
8   for  $j \leftarrow 1$  to  $n$  do
9     Normalize  $d_j$ :  $\tilde{d}_j \leftarrow d_j / \|d_j\|$ 
10    Encode:  $p_j \leftarrow \text{Plain}(\tilde{d}_j)$ 
11    Hadamard product:  $c_{i,j} \leftarrow c_i \odot p_j$ 
12    Inner product:  $c_{i,j}^{\text{sum}} \leftarrow \text{SumSlots}(c_{i,j})$ 
13  Send all  $c_{i,j}^{\text{sum}}$  back to client;
14 Client:
15 for  $i \leftarrow 1$  to  $m$  do
16   Initialize match set:  $\mathcal{M}_i \leftarrow \emptyset$ 
17   for  $j \leftarrow 1$  to  $n$  do
18      $s_{i,j} \leftarrow \text{Dec}(c_{i,j}^{\text{sum}})[0]$ 
19     if  $s_{i,j} \geq \tau$  then
20        $\mathcal{M}_i \leftarrow \mathcal{M}_i \cup \{j\}$ 
21 return  $\{\mathcal{M}_1, \dots, \mathcal{M}_m\}$ 

```

D. Security Analysis

The system operates under an honest-but-curious threat model, in which all participants follow the protocol but may

attempt to infer additional information. The security is analyzed against three potential attack vectors:

Client Device. The client device is assumed to be secure as it retains the private decryption key θ_d and is the only point where raw facial images and plaintext embeddings are processed.

Communication Channel. All data transmitted between the client and the database are encrypted using CKKS. Without the decryption key, intercepted ciphertexts are computationally infeasible to decode.

Database The database stores only encrypted embeddings and public keys. Even if compromised, no meaningful biometric information can be inferred due to the semantic security of CKKS.

The security of CKKS relies on the hardness of the Ring Learning with Errors (RLWE) problem, which guarantees that reconstructing plaintext embeddings or extracting soft biometric attributes is computationally infeasible. Moreover, the approximate arithmetic inherent in CKKS further reduces the feasibility of reconstructing exact facial templates, thus strengthening privacy guarantees.

IV. EXPERIMENTS

The experiments conducted aim to evaluate the feasibility of using homomorphic encryption for privacy-preserving image matching in criminal investigations. This section outlines the details of the experiments, the datasets, the models used, and the configuration of the cryptosystem under test.

A. Environment

The experiments were carried out on a high-performance machine equipped with an Intel Core i9 CPU, 32GB of RAM, and an NVIDIA RTX 3090 GPU. The system ran a Linux environment via Windows Subsystem for Linux (WSL 2), using the Linux kernel version 5.15.167.4-microsoft-standard-WSL2 on an x86-64 architecture. The implementation was developed in Python using the TenSEAL library [26], which provides support for homomorphic encryption based on the CKKS scheme. This setup enabled efficient and secure computation of encrypted similarity scores, facilitating privacy-preserving face matching across large-scale datasets.

B. Data and Models

To evaluate the privacy-preserving face search capability of our system, we used the Labeled Faces in the Wild (LFW) dataset [7], which is a well-established benchmark for facial recognition tasks. Instead of verifying image pairs, we structured the experiment as a search problem: a set of query images was matched against a gallery (database) of facial embeddings to identify potential matches.

The gallery consisted of facial embeddings derived from 1,000 identities, while the query set was composed of a subset of these images, ensuring both positive and negative matches. This setup allows us to evaluate the effectiveness and accuracy of encrypted similarity search under realistic conditions.

To generate embeddings, we employed three different face recognition models with varying levels of complexity:

- **FaceNet-128d**: A lightweight model that outputs 128-dimensional embeddings.
- **FaceNet-512d**: A more expressive variant producing 512-dimensional feature vectors.
- **VGG-Face**: A deep convolutional model generating 4096-dimensional embeddings.

These models were chosen to assess how the system scales with increasing embedding dimensionality and to understand the trade-offs between accuracy and computational cost in the encrypted domain.

C. Cryptosystem Configurations

In our experiments, we focused exclusively on the CKKS homomorphic encryption scheme, applying it to evaluate the feasibility of privacy-preserving face search based on cosine similarity. All encryption parameters were selected to ensure a 128-bit security level, following the recommended guidelines in the SEAL library documentation [26].

Two parameter configurations were tested to study the trade-off between computational performance and encryption strength:

- **Lightweight Configuration**: Designed with a smaller polynomial modulus degree and reduced coefficient modulus sizes to improve efficiency, while still maintaining acceptable precision.
- **High-Precision Configuration**: Employs larger encryption parameters, including a higher polynomial modulus degree and larger coefficient moduli, allowing for greater numerical accuracy and deeper computations at the cost of increased runtime.

Both configurations were used to perform cosine similarity computations over encrypted embeddings with three different dimensionalities (128, 512, and 4096). This setup allowed us to assess the scalability and performance of CKKS-based encrypted search under varying levels of data complexity.

D. Experiment Results

To assess the computational efficiency of the proposed privacy-preserving face search system, we measured the execution times of three core operations under homomorphic encryption using the CKKS scheme: encryption, decryption, and encrypted cosine similarity computation (denoted as "matching"). The experiments were conducted across two cryptographic configurations—one lightweight and one more secure—while varying the embedding dimensionality based on three widely used face recognition models: FaceNet-128d, FaceNet-512d, and VGG-Face (4096d).

The detailed performance results are summarized in Table I. As shown, the smaller configuration ($n = 2^{13}, q = 200, g = 2^{40}$) achieves low latency for both encryption and decryption across all models, with matching times ranging from 24.14 ms (FaceNet-128d) to 39.11 ms (VGG-Face). These values are suitable for real-time or near-real-time applications, particularly when moderate embedding sizes are used.

In contrast, the larger configuration ($n = 2^{14}, q = 422, g = 2^{60}$) incurs higher overhead, with encryption and decryption

TABLE I
PERFORMANCE OF CKKS-BASED COSINE SIMILARITY (MS PER ITERATION)

n	q	g	Operation	FaceNet (128d)	FaceNet (512d)	VGG-Face (4096d)
2^{13}	200	2^{40}	Encryption	5.34	7.63	8.28
			Matching	24.14	30.62	39.11
			Decryption	0.92	1.26	1.12
2^{14}	422	2^{60}	Encryption	19.56	19.22	19.67
			Matching	198.38	271.42	308.34
			Decryption	8.23	8.89	9.16

times increasing approximately 2–3 times, and matching operations taking over 300 ms in the worst case (for VGG-Face). Nevertheless, this configuration provides a stronger security margin and higher precision, making it more appropriate for applications that prioritize data protection over latency.

Overall, the results confirm that the system can operate efficiently even under homomorphic encryption, with performance scalable across different embedding sizes and encryption strengths.

In addition to performance, we also evaluated the accuracy of the encrypted face matching process using cosine similarity under homomorphic encryption. To do this, we applied a fixed similarity threshold (empirically determined for each model) to decide whether a given query image matched an identity in the database. The evaluation was conducted on the LFW dataset, with 1,000 query images tested against the encrypted gallery set.

The results demonstrated that the use of CKKS-based encryption had no impact on matching accuracy when compared to plaintext computations. Since the cosine similarity is computed directly on normalized embeddings and preserved through homomorphic operations, the relative ranking and decision threshold remain unchanged. Across all three models, the encrypted matching process achieved identical true positive and true negative rates to the unencrypted baseline, confirming that the numerical precision of the CKKS scheme was sufficient to maintain fidelity for this task.

These findings validate that privacy-preserving face search using homomorphic encryption can maintain full utility, without sacrificing accuracy, while offering strong confidentiality guarantees. Therefore, the proposed method provides a practical and secure alternative to conventional face recognition systems in sensitive application domains.

V. CONCLUSION

This work presents Pri-Inv, a privacy-preserving framework for secure image matching in digital forensics using homomorphic encryption. By leveraging the CKKS scheme, our system enables encrypted face embeddings to be compared efficiently without revealing sensitive information. The proposed protocol supports cosine similarity and scales effectively across various embedding dimensions.

Experimental results on the LFW dataset confirm that Pri-Inv achieves accurate matching with minimal overhead, even on large datasets. By offloading computations to the cloud, the

framework maintains strong privacy guarantees while reducing the burden on local devices.

Overall, Pri-Investigato demonstrates the practical potential of homomorphic encryption for forensic investigations, balancing security, efficiency, and scalability. Future work may explore further optimizations and broader applications in privacy-preserving AI.

ACKNOWLEDGMENT

The article was completed with support from the Academy of Cryptography Techniques and Ministry of Public Security.

REFERENCES

- [1] T. B. Ogunseyi and O. M. Adedayo, "Cryptographic techniques for data privacy in digital forensics," *IEEE Access*, vol. 11, pp. 142 392–142 410, 2023.
- [2] A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, "A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions," *IEEE Access*, vol. 10, pp. 11 065–11 089, 2022.
- [3] F. Casino, T. K. Dasaklis, G. P. Spathoulas, M. Anagnostopoulos, A. Ghosal, I. Borocz, A. Solanas, M. Conti, and C. Patsakis, "Research trends, challenges, and emerging topics in digital forensics: A review of reviews," *IEEE Access*, vol. 10, pp. 25 464–25 493, 2022.
- [4] K. Sivanagireddy, S. Jagadeesh, and A. Narmada, "Identification of criminal & non-criminal faces using deep learning and optimization of image processing," *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 47 373–47 395, 2024.
- [5] S. Sandhya, A. Balasundaram, and A. Shaik, "Deep learning based face detection and identification of criminal suspects," *Computers, Materials & Continua*, vol. 74, no. 2, 2023.
- [6] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Computing Surveys (Csur)*, vol. 51, no. 4, pp. 1–35, 2018.
- [7] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.
- [8] M. Sharif, F. Naz, M. Yasmin, M. A. Shahid, and A. Rehman, "Face recognition: A survey," *Journal of Engineering Science & Technology Review*, vol. 10, no. 2, 2017.
- [9] E. V. Sousa and D. G. Macharet, "Structural reasoning for image-based social relation recognition," *Computer Vision and Image Understanding*, vol. 235, p. 103785, 2023.
- [10] S. M. Arman, T. Yang, S. Shahed, A. Al Mazroa, A. Attiah, and L. Mohaisen, "A comprehensive survey for privacy-preserving biometrics: Recent approaches, challenges, and future directions," *Computers, Materials & Continua*, vol. 78, no. 2, pp. 2087–2110, 2024.
- [11] S. Maurya, J. Bharti, and B. Roy, "A survey on cancellable biometrics techniques," in *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE, 2024, pp. 1–7.
- [12] V. K. Hahn and S. Marcel, "Biometric template protection for neural-network-based face recognition systems: A survey of methods and evaluation techniques," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 639–666, 2022.
- [13] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: a complete review," *Multimedia tools and applications*, vol. 80, no. 3, pp. 4825–4880, 2021.
- [14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
- [15] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.
- [16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
- [17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, X. Xie, M. W. Jones, and G. K. L. Tam, Eds. BMVA Press, 2015, pp. 41.1–41.12. [Online]. Available: <https://doi.org/10.5244/C.29.41>
- [18] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.
- [19] G. M. Penn, G. Pötzelsberger, M. Rohde, and A. Uhl, "Customisation of paillier homomorphic encryption for efficient binary biometric feature vector matching," in *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2014, pp. 1–6.
- [20] M. Upmanyu, A. M. Namboodiri, K. Srinathan, and C. Jawahar, "Blind authentication: a secure crypto-biometric verification protocol," *IEEE transactions on information forensics and security*, vol. 5, no. 2, pp. 255–268, 2010.
- [21] J. H. Cheon, H. Chung, M. Kim, and K.-W. Lee, "Ghostshell: Secure biometric authentication using integrity-based homomorphic evaluations," *Cryptology ePrint Archive*, 2016.
- [22] C. Gentry, S. Halevi, and N. P. Smart, "Fully homomorphic encryption with polylog overhead," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2012, pp. 465–482.
- [23] J. R. Troncoso-Pastoriza, D. González-Jiménez, and F. Pérez-González, "Fully private noninteractive face verification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 7, pp. 1101–1114, 2013.
- [24] J. Fan and F. Vercauteren, "Somewhat practical fully homomorphic encryption," *Cryptology ePrint Archive*, 2012.
- [25] J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *International conference on the theory and application of cryptology and information security*. Springer, 2017, pp. 409–437.
- [26] A. Benaissa, B. Retiat, B. Cebere, and A. E. Belfedhal, "Tenseal: A library for encrypted tensor operations using homomorphic encryption," 2021.

MAML AE-SAC: A Synergy framework of Meta-Learning and Adversarial Reinforcement Learning for Adaptive Intrusion Detection

Duy Nhat Dang*, Giang Nguyen Linh**, Tung Bui Trong*, Dong Le Van*, and Tuyen Ngoc Le[†]

*Hanoi University of Science and Technology, Hanoi, Vietnam

Email: duydang312@gmail.com, {giangnl, tungbt}@soict.hust.edu.vn, donglevan1@hust.edu.vn

[†]Department of Electronic Engineering, Ming Chi University of Technology, New Taipei, 24301 Taiwan, ROC

Center for Reliability Engineering, Ming Chi University of Technology, New Taipei, 24301 Taiwan, ROC

Email: tuyennl@mail.mcut.edu.tw

Abstract—Modern Intrusion Detection Systems (IDS) suffer enormous challenges dealing with sophisticated adversarial attacks and are challenged by their ability to cope with new attacks, especially under imbalanced conditions where sparse threats are often overlooked. This work introduces MAML AE-SAC, a novel framework that combines adversarial reinforcement learning with model-agnostic meta-learning (MAML) [1] to construct a robust and fast-adapting IDS. Training a Soft Actor-Critic (SAC) agent adversarially and adapting with MAML, our solution prepares the system to defensively counter both strong and infrequent threats. Test results with the NSL-KDD [2], AWID [3], and CICIDS2017 [4] data sets confirm that MAML AE-SAC sets a new record, achieving significant progress compared to existing benchmarks with regard to detection precision and generalizability.

Index Terms—Intrusion Detection, Reinforcement Learning, Meta-Learning, MAML, Adversarial Learning, Cybersecurity

I. INTRODUCTION

Conventional Intrusion Detection systems, which work based on known attack signatures, are useless against unknown or smartly manipulated threats. On the other hand, Anomaly detection IDS based on machine learning perform better but are still brittle when attackers manipulate input to deceive them. These vulnerabilities point to the need for a new, robust approach to intrusion detection.

Apart from the successful application of deep learning methods in this field, several challenges still exist that require attention. On the one hand, deep learning models show a high sensitivity to the datasets used during training. On the other hand, the network intrusion detection datasets obtained from real network environments usually include a large amount of normal behavior data along with a small amount of attack behavior data, leading to a highly unbalanced dataset. This imbalance leads to an inadequate detection ability for some intrusions in deep learning models. On the other hand, the development of effective models faces significant challenges. It is well known that deep learning models have a large number

of parameters, and the careful tuning of these parameters requires a large amount of both labor and time.

The emergence of deep reinforcement learning (DRL) has brought with it a new solution to solve the intrusion detection problem. Although these reinforcement learning techniques have achieved satisfactory performance in intrusion detection, there are still challenges in dealing with imbalanced data, multi-class cyberattacks and limited samples in certain types of attacks. This paper introduces a novel framework, MAML AE-SAC (Model-Agnostic Meta-Learning with Adversarial Environment Soft Actor-Critic), designed to overcome these multifaceted challenges. Our paper introduces a novel deep reinforcement learning model based on the Soft Actor-Critic (SAC) algorithm, enhanced by the integration of Model-Agnostic Meta-Learning (MAML), adversarial training, and the Synthetic Minority Over-sampling Technique (SMOTE).

In this work, we propose a novel architecture that synergizes adversarial RL with meta-learning to enhance rare attack detection, improve adversarial resilience, and enable rapid adaptation to new threats. And we also perform some experiments to evaluate the effectiveness of this MAML integrated framework to the intrusion detection domain.

The remainder of this paper is structured as follows. Section II reviews related literature. Section III details the MAML AE-SAC framework. Section IV describe our experimental setup and present a comprehensive analysis of the results. Finally, Section V concludes the paper and discusses future work.

II. RELATED WORK

Intrusion detection has seen significant developments with a shift from traditional approaches to data-driven ones. The application of Deep Learning (DL) architectures, such as using Autoencoders to learn unsupervised features [5] and applying CNN-LSTM architectures to identify spatio-temporal behaviors [6], has shown to be highly accurate. However, such architectures are inherently reactive and suffer from not being able to handle new attack varieties without undergoing rigorous retraining procedures.

This work is supported by the Ministry of Science and Technology of Vietnam (MOST) under Grant No. KC.01.04/21-30.

**Corresponding author

Meta-learning, or “learning to learn,” is an attractive solution to the problem of rapidly adapting to small datasets. The Model-Agnostic Meta-Learning (MAML) algorithm, as proposed by Finn et al. [1], is an important breakthrough in such areas. Instead of aiming for the optimization of a single optimal set of model parameters, MAML aims to learn a parameter initialization that can be rapidly adapted to new tasks using few examples and gradient updates. This line of work has been successful and spawned further advancements, leading to systems not only that meta-learn like MAML but also which meta-learn further aspects of the optimization process, even down to the updating rule itself [7], [8]. The core insight behind MAML—an versatile initialization well suited for rapid adaptation—serves as an underlying guiding principle behind our very own research work.

Research Gap. The existing body of work shows an important gap. The models based on deep learning lack proactivity, while the current reinforcement learning-based models are not adequately geared for prompt, few-shot adaptation when new threats are encountered. Meta-learning is a plausible remedy for enabling quick adaptation; but no effort is evident which combines this technique in an overarching and adversarial-based reinforcement learning architecture for intrusion detection framework. The present study fills the gap identified as well as establishing an intrusion detection framework which not just proves resilient against known types of attack but is also specifically trained for quick adaptation against new and dynamic threats.

III. METHODOLOGY

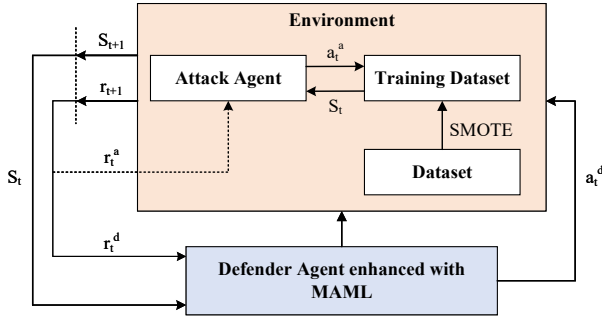


Fig. 1: Overview of the MAML AE-SAC Framework.

A. MAML AE-SAC

Our paper proposes a Defender Agent enhanced with the MAML [1] method for network intrusion detection, adapted from the adversarial reinforcement learning model in [9]. As illustrated in Figure 1, the framework consists of two core components: an adversarial environment for data generation and a MAML-enhanced Defender Agent. The environment features an Attack Agent and uses a training dataset pre-processed with SMOTE [10] to oversample minority classes.

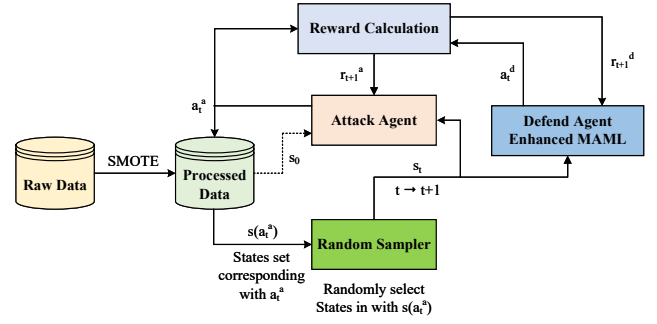


Fig. 2: The adversarial sampling process.

The system operates in an interactive loop. The Attack Agent (policy π_a) selects challenging data samples from the augmented training set. The MAML-enhanced Defender Agent (policy π_d) then classifies these samples. Both agents are implemented using reinforcement learning principles, receiving opposing rewards (r_t^a and r_t^d) to drive the adversarial dynamic. This process forces the Defender to learn a policy that is robust against difficult and rare attack types.

In this RL formulation, the *state* (s_t) represents a single network connection or packet, encapsulated as a feature vector containing attributes like protocol type, duration, and service. The *actions* are defined differently for each agent. For the Defender Agent, an action a_t^d is a classification decision from its action space, which includes all traffic categories (e.g., 'Normal', 'DoS', 'Probe'). For the Attack Agent, an action a_t^a is the selection of a specific attack type from the training set to challenge the Defender, aiming to find its weaknesses.

B. Reward Function

The adversarial nature of the framework is encoded in its reward functions, adopted from [9]. The functions are designed to incentivize the detection of difficult, low-frequency attacks.

$$r_{t+1}^d = \begin{cases} 0, & a_t^d \neq a_t, \\ 1, & a_t^d = a_t, a_t \in A_L, \\ 2, & a_t^d = a_t, a_t \in A_M, \end{cases} \quad (1)$$

$$r_{t+1}^a = \begin{cases} 0, & a_t^d = a_t, \\ 1, & a_t^d \neq a_t, a_t \in A_L, \\ 2, & a_t^d \neq a_t, a_t \in A_M, \end{cases} \quad (2)$$

Here, a_t^d is the Defender's action, a_t is the true label, A_L represents high-frequency (common) attack classes, and A_M represents low-frequency (minority) classes. The Defender is rewarded more for correctly identifying rare attacks, while the Attacker is rewarded more for finding rare attacks that the Defender misclassifies.

C. Training process

This research improved the training process in [9], which is illustrated in Fig 2. The environment agent selects an action label a_t^e from its policy π_e given the current state s_t . Based

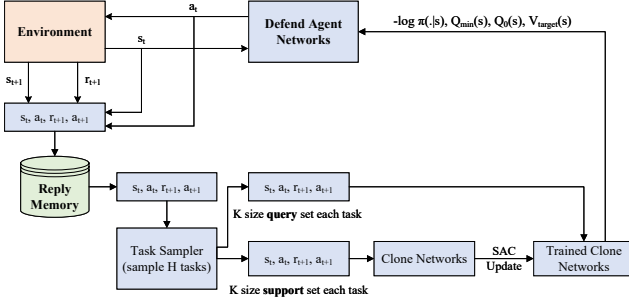


Fig. 3: The MAML-SAC meta-update procedure for the Defender Agent.

on this action label a_t^e , the next state s_{t+1} is selected from the training data, where the actual action label a_t^c is also received. Subsequently, the classifier agent's policy π_d determines the action label a_{t+1}^d given the state s_{t+1} . Following this, the environment agent receives a reward r_{t+1}^e , and the classifier agent receives a reward r_{t+1}^d .

Two experience tuples, $(s_t, a_t^d, r_{t+1}^d, s_{t+1})$ for the classifier agent and $(s_t, a_t^e, r_{t+1}^e, s_{t+1})$ for the environment agent, accumulate in their respective experience replay memories. After collecting sufficient empirical data, both agents are updated using the Soft Actor-Critic (SAC) [11]–[13] update procedure to obtain new policies π_a and π_d . Moreover, with Defend Agent, it's training alternates between a standard SAC update and a specialized MAML meta-update, triggered periodically (controlled by the P frequency). This Meta-Learning Update step is demonstrated in Fig. 3. The key difference between the two update mechanism starts with the Task Sampler which treats H most popular attack types from recent experiences as learning tasks, and then samples data from memory corresponding to these tasks. The sample set for each task includes a support set (D_{support}) and query data sets (D_{query}). Following this, a temporary copy of our Defender—the Clone Network—is trained on the support set using the standard SAC update.

Finally, in the meta-optimization step, the query set is fed into the trained clone networks, and then its output is used to update the parameters of the original Defender Agent, training it to generalize and learn effectively from small quantities of new task-specific data. The environment agent then continues data sampling using the updated policy π_a .

The initial state s_0 is randomly selected from the training data. The policies π_a and π_d assign probability values to all possible actions, and actions a_t^e and a_t^c are selected based on these probabilities. According to [14], the action sets A_a and A_d in the NSL-KDD dataset [2] are not necessarily equivalent. The classifier agent's action set is defined as $A_d \in \{0, 1, 2, 3, 4\}$, corresponding to the number of classifier classes. In contrast, the environment agent's action set is defined as $A_a \in \{0, 1, \dots, 22\}$, representing all possible attacks in the training dataset.

D. Technical challenges of Adversarial Meta-Reinforcement Learning

Integrating a meta-learning method like MAML into an adversarial training framework such as AE-SAC has several technical challenges that extend beyond a simple algorithmic combination. The primary difficulty stems from applying a meta-learner, which is designed for rapid adaptation across a distribution of stationary tasks to a highly non-stationary environment created from the evolving strategies of two agents.

The foremost challenge is how to maintain training stability. In a standard adversarial RL setting, the two agents create a "moving-target" problem for each other. MAML can potentially amplify this instability. For example, a successful meta-update can improve the Defender's ability to quickly adapt to a certain attack category and also can alter the reward landscape for the Attacker. If the Attacker's policy cannot adapt effectively in response, it may be driven towards unstable or even collapsed policies which affect the learning process for both agents. A delicate balance has to be achieved to ensure both the Defender's meta-learner and the Attacker's policy learner make steady progress.

Moreover, this framework is sensitivity to hyperparameters. The learning rate of the meta-optimizer (β), the inner-loop adaptation rate for tasks, and the learning rates of the underlying SAC agents must be carefully tuned. An improperly configured meta-optimizer can either cause catastrophic forgetting of the general policy or fail to learn a useful prior so that reducing the benefits of MAML in this adversarial context.

IV. EXPERIMENTAL RESULTS

A. Experiment scenarios

Hyperparameter Settings: In this paper, network architectures are based on [9]. This architecture for the environment agent and the classifier agent is provided in Tables I and Table II, respectively. The numbers in parentheses indicate the dimensions of the neurons for the input and output layers of the AWID dataset [3]. All networks listed in the tables are fully connected. The actor and critic neural network models of our agents have 3 hidden layers with 700 and 100 neurons (based on empirical testing) for the Defender and Attacker, respectively, and apply the ReLU activation function. Additional parameters for training the MAML AE-SAC model are provided in Table III.

TABLE I: Network Architecture for Attacker Agent

Name	Network Structure
Actor	122(46)-100-100-100-5(4)
Q Critic	122(46)-100-100-100-5(4)
V Critic	122(46)-100-100-100-1

Datasets: Our methodology's effectiveness is evaluated by applying our approach to three well-known benchmark datasets: NSL-KDD [2], a classical dataset with severe class imbalance; AWID [3], a modern and large-scale dataset capturing 802.11 network traffic; and CICIDS2017 [4], another popular large-scale benchmark.

TABLE II: Network Architecture for Defender Agent

Name	Network Structure
Actor	122(46)-700-700-700-23(4)
Q Critic	122(46)-700-700-700-23(4)
V Critic	122(46)-700-700-700-1

TABLE III: Key Training Hyperparameters

Parameter	Value
Number of epochs	60
Steps per epoch (Episode length)	180
Batch size (for SAC update)	180
Replay buffer size	1000
Discount factor (γ)	0.001
Entropy coefficient (α)	0.2
Target network update rate (τ)	0.1
MAML Parameters	
Meta-optimizer learning rate (β)	0.00025
Support set size per task (K)	50 (100)
Query set size per task (K)	50 (100)
Number of tasks per meta-batch (H)	4 (3)

Preprocessing: The preprocessing methods utilized include data cleaning, one-hot encoding for categorical data, and Min-Max normalization for numerical data. Notably, with respect to our NSL-KDD [2] training data, we use SMOTE [10] to over-sample minority attack classes with less than 60 instances, thus increasing their representation to 3,000 instances. The purpose of this is to provide a more balanced starting view for the learning agents.

Evaluation Metrics: The metrics used include Accuracy, Precision, Recall, and F1-Score. Due to the common presence of class imbalance in the datasets, the F1-Score is used as our primary metric of overall performance.

Baselines: We compare MAML AE-SAC against a comprehensive set of baselines, focusing on the top-performing models from traditional ML, DL, and state-of-the-art DRL.

B. Performance Evaluation

We evaluated our MAML AE-SAC architecture on three commonly accepted benchmark datasets. Across all experimental tests, our algorithm systematically achieved a new state-of-the-art accuracy, thus emphasizing the valuable benefits derived from combining meta-learning within an adversarial reinforcement learning framework.

In the established **NSL-KDD** data, where a severe imbalance exists amongst classes, MAML AE-SAC achieves a remarkable F1-Score of 83.23% (Table IV). This result, by virtue of exceeding not just the immediate predecessor’s performance, AE-SAC (80.39%), but even that of the specially developed AESMOTE technique’s optimised variant, 82.43%, points to a gain derived from the versatility cultivated by MAML, rather than by data augmentation methods alone. With regards to the complex and real-world **AWID** wireless network dataset, the proposed model proves to be better-performing, with an obtained F1-Score of 98.33% (see Table V). Notably, it shows a highest precision across all models tested, with

a reading of 98.55%. This higher precision has significant real-world relevance, since it corresponds to a lower rate of false alarm incidence, a major operational challenge faced by security operatives.

Finally, the comprehensive **CICIDS2017** [4] benchmark shows that MAML AE-SAC exhibits considerable scalability and robustness with an F1-Score of 99.11% (see Table VI). This very high score further supports its role as an excellent methodology, consistently outperforming powerful deep learning baselines and other reinforcement learning-based benchmarks, in a benchmark dataset covering a broad set of modern attacks. In conclusion, the experimental results on various benchmarks provide disparate yet strong evidence that highlights the effectiveness and improved generalizability of the MAML AE-SAC algorithm.

TABLE IV: Performance on the NSL-KDD Dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
<i>Traditional Machine Learning</i>				
RBF-SVM [9]	80.65	81.30	80.65	80.56
Linear-SVM [9]	75.60	76.22	75.60	72.95
<i>Deep Learning</i>				
DNN [9]	78.50	81.60	78.50	76.50
1D-CNN [9]	78.75	80.94	78.75	76.33
<i>Deep Reinforcement Learning</i>				
AESMOTE [9]	82.09	84.11	82.09	82.43
AE-SAC	81.01	81.23	81.01	80.39
MAML AE-SAC (Ours)	83.98	83.50	83.98	83.23

TABLE V: Performance on the AWID Dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
<i>Traditional Machine Learning</i>				
J48 (Decision Tree) [9]	96.26	96.20	96.30	96.30
Random Tree (RT) [9]	96.23	95.90	96.20	94.80
<i>Deep Learning</i>				
GRU [9]	95.27	93.65	95.27	93.67
1D-CNN [9]	95.37	94.24	95.37	93.57
<i>Deep Reinforcement Learning</i>				
SSDDQN [9]	98.19	98.40	98.19	98.22
AE-SAC	95.84	97.40	95.84	96.31
MAML AE-SAC (Ours)	98.29	98.55	98.29	98.33

TABLE VI: Performance on the CICIDS2017 Dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
<i>Traditional Machine Learning</i>				
SVM [15]	95.50	97.72	99.12	98.40
Stacking-based Ens. [16]	98.15	98.20	98.15	98.16
<i>Deep Learning</i>				
CNN+LSTM [15]	97.16	97.41	99.10	98.24
Attentive Transformer [17]	97.03	97.02	91.10	96.97
<i>Deep Reinforcement Learning</i>				
AE-SAC	97.79	99.42	97.79	98.53
MAML AE-SAC(Ours)	98.85	99.45	98.85	99.11

C. Discussion

The consistent improvement in performance compared to the baseline standard AE-SAC [9] across a number of datasets undeniably confirms benefits arising from its combination with MAML [1]. While robustness is improved with adversarial training, using a meta-learning approach combined with a loss

based on MAML leads to better generalization capabilities for the agent. Pre-training with MAML assists with "fast-learning". This leads to a better policy that shows robust generalization abilities across diverse scenarios. This reflects higher overall F1-scores.

The obtained results are promising, but research does come with a certain set of limitations. Our evaluation was primarily conducted on the NSL-KDD [2] dataset. Firstly, although NSL-KDD is a standard benchmark for comparing IDS performance, this dataset represents an older network traffic profile and may not fully capture the complexity and characteristics of modern attacks. To validate the robustness and generalizability of our approach, future efforts will extend to more recent, realistic and diverse datasets, including the Bot-IoT dataset which is designed for IoT ecosystems, and the UNSW-NB15. Secondly, The MAML AE-SAC algorithm is characterized by significant computational requirements, contains two agents and several stages included within its inner loop, leading to increased computational intensity during algorithm operation as well as practical implementation. In our case, concerning our own NSL-KDD dataset, our model's predictive performance falls short of predictive accuracy expectations for the U2R minor attack type (as shown in 5), the reason is due to restricted quantity of U2R instances which is only 52 samples included within our training data—with a less-than-ideal before implementation of the SMOTE technique [10]. In contrast, for the AWID dataset [3], our model shows improved performance metrics for the two lowest data classes of injection and impersonation when compared to the larger flooding data type, as seen in 4. Similarly, for the CIC-IDS2017 dataset, our model achieves relatively low performance for the Botnet ARES minor attack type (as shown in the 6), the reason is due to its traffic mimicking benign behavior, leading to a lower accuracy of 62% with 1,555 samples. In contrast, the Web Attack class with 1,694 samples achieves a high accuracy of 99% since it is more distinct anomalous patterns.

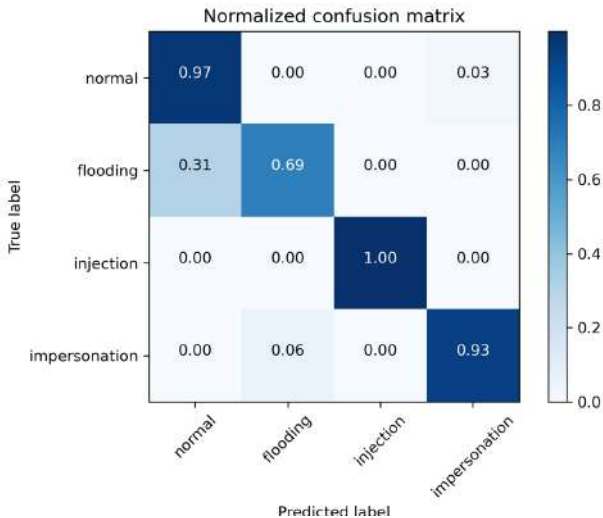


Fig. 4: Confusion matrix of MAML AE-SAC on AWID dataset.

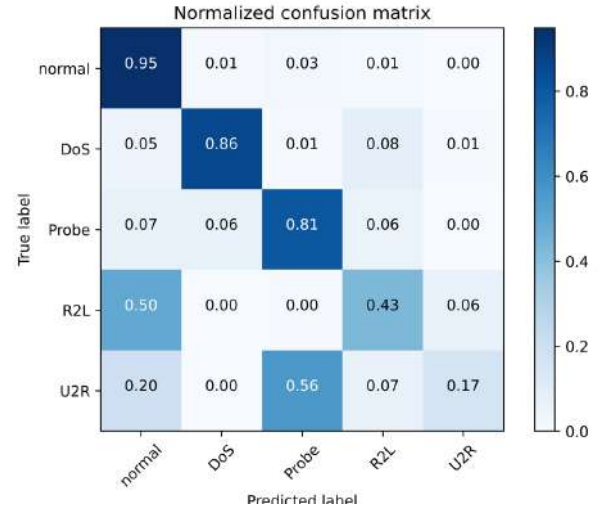


Fig. 5: Confusion matrix of MAML AE-SAC on NSL-KDD dataset.

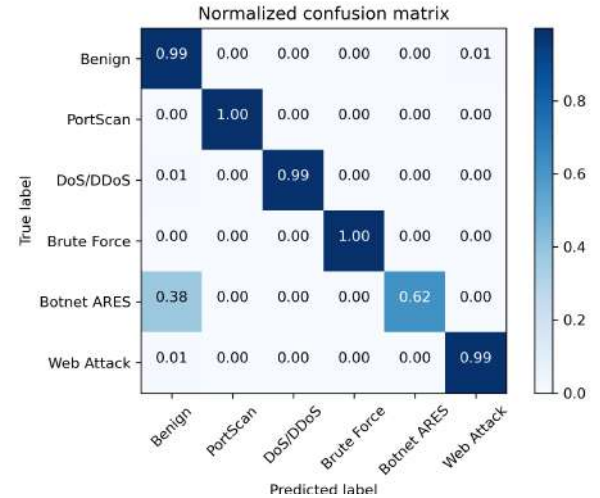


Fig. 6: Confusion matrix of MAML AE-SAC on CIC-IDS2017 dataset.

V. CONCLUSION

We present **MAML AE-SAC**, a new meta-learning intrusion detection framework, which combines adversarial reinforcement learning with ideas drawn from model-agnostic meta-learning [1]. While prioritizing goal-specific training with a view to supporting rapid adaptation, our proposed solution reveals a capability to effectively meet challenging threats with the ability to produce rapid reactions to novel and unexpected challenges. Extensive experiments on the NSL-KDD, AWID and CICIDS2017 datasets verified our approach's superiority compared to a wide array of state-of-the-art methods.

Despite promising results, the computational overhead of the dual-agent meta-learning framework presents a limitation. Future work will explore more efficient variants, such as Second-Order MAML for fast convergence but may require

more computation, and investigate alternative oversampling techniques to further improve performance on extremely rare attack classes. And to bridge the gap between research and practical application, we will focus on model optimization and compression in future work. For example, the process of quantization can reduce the weights of the network to lower precisions (i.e., change the value of floating-point 32 bits to 8-bit integers), which substantially shrinks the used memory and increases the speed of inference, particularly on specialized devices. Finally, network pruning offers another avenue, where redundant connections or neurons are systematically removed from the trained agents post-training. This research paves the way for a new generation of proactive and intelligent cybersecurity defenses.

ACKNOWLEDGMENT

This work is supported by the Ministry of Science and Technology of Vietnam (MOST) under Grant No. KC.01.04/21-30. We would like to thank Bach Khoa Cyber Security Center for their supports.

REFERENCES

- [1] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017.
- [2] M. Tavallaei, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, pp. 1–6, Ieee, 2009.
- [3] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 184–208, 2015.
- [4] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, *et al.*, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSp*, vol. 1, no. 2018, pp. 108–116, 2018.
- [5] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE transactions on emerging topics in computational intelligence*, vol. 2, no. 1, pp. 41–50, 2018.
- [6] J. Zhang, Y. Ling, X. Fu, X. Yang, G. Xiong, and R. Zhang, "Model of the intrusion detection system based on the integration of spatial-temporal features," *Computers & Security*, vol. 89, p. 101681, 2020.
- [7] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," *arXiv preprint arXiv:1707.09835*, 2017.
- [8] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *International conference on learning representations*, 2017.
- [9] Z. Li, C. Huang, S. Deng, W. Qiu, and X. Gao, "A soft actor-critic reinforcement learning algorithm for network intrusion detection," *Computers & Security*, vol. 135, p. 103502, 2023.
- [10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [11] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, pp. 1861–1870, Pmlr, 2018.
- [12] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.
- [13] P. Christodoulou, "Soft actor-critic for discrete action settings," *arXiv preprint arXiv:1910.07207*, 2019.
- [14] G. Caminero, M. Lopez-Martin, and B. Carro, "Adversarial environment reinforcement learning algorithm for intrusion detection," *Computer Networks*, vol. 159, pp. 96–109, 2019.
- [15] M. Roopak, G. Y. Tian, and J. Chambers, "Deep learning models for cyber security in iot networks," in *2019 IEEE 9th annual computing and communication workshop and conference (CCWC)*, pp. 0452–0457, IEEE, 2019.
- [16] M. Ali, M.-u. Haque, M. H. Durad, A. Usman, S. M. Mohsin, H. Mujlid, and C. Maple, "Effective network intrusion detection using stacking-based ensemble approach," *International Journal of Information Security*, vol. 22, no. 6, pp. 1781–1798, 2023.
- [17] D. Zegarra Rodriguez, O. Daniel Okey, S. S. Maidin, E. Umoren Udo, and J. H. Kleinschmidt, "Attentive transformer deep learning algorithm for intrusion detection on iot systems using automatic explainable feature selection," *Plos one*, vol. 18, no. 10, p. e0286652, 2023.

A Deep Self-Refining Approach for Fact Verification

1st Cuong Hoang

*Japan Advanced Institute of Science
and Technology*
Nomi, Ishikawa, Japan
0009-0005-2583-5332

2nd Vu Tran

*Japan Advanced Institute of Science
and Technology*
Nomi, Ishikawa, Japan

3rd Le-Minh Nguyen

*Japan Advanced Institute of Science
and Technology*
Nomi, Ishikawa, Japan

Abstract—In the digital age, the widespread dissemination of fake news and misinformation on social media and sensational tabloids has become a significant threat to public awareness and social stability. These deceptive claims, especially those involving misleading images coupled with manipulated textual contexts, are increasingly sophisticated and harder to detect. This situation calls for the urgent development of automatic fact-checking systems capable of verifying the consistency between visual evidence, contextual information, and the claims presented.

In this paper, we propose an automated multimodal fact-checking framework leveraging Multimodal Large Language Models (MLLMs) to verify claims based on image and contextual input. Our main idea is to design instruction prompts that guide the model through a multi-stage reasoning process, starting with an initial classification using a general-purpose prompt. Based on this preliminary prediction, we dynamically select tailored follow-up prompts and perform multiple inference rounds to gather diverse perspectives. Experimental results demonstrate that our method improves factuality prediction with a 0.6% gain in accuracy and a 3.04% increase in macro F1-score, highlighting its effectiveness in handling complex multimodal verification tasks.

Index Terms—Vision-language model, Fact-checking, Large Language Model.

I. INTRODUCTION

In today’s era of rapid technological advancement, particularly with the explosive growth of social media, individuals are exposed to an overwhelming volume of information and news on a daily basis. Exploiting this environment, numerous organizations and individuals deliberately disseminate false or misleading content on social media platforms, often driven by political agendas, subversive intentions, or the desire to incite confusion and hatred among users. To formalize this problem and provide a foundation for the global research community to address it, a substantial body of work has been devoted to fact-checking and misinformation detection in recent years.

In addressing Multimodal Fact-Checking, the problem is formulated as a multiple-choice question, where the goal is to select the correct answer as defined in Table I. Our objective is to verify the factual accuracy based on an image-context pair and choose the corresponding label. To effectively address this task, we adopt a multi-stage prompting strategy to enhance the semantic understanding of the image-context pair. We observed that the model often struggles to distinguish

between classes. This observation motivated us to integrate a self-refinement mechanism [1] into both LLMs and MLLMs, aiming to enhance the model’s ability to make more accurate label distinctions and improve overall performance.

TABLE I
ILLUSTRATION OF DATA INPUT IN MULTIMODAL FACT-CHECKING.

Image id	0599101.jpg
Context	The article discusses the resignation of Natalie Nougayrède, the female editor-in-chief of the leading French daily, Le Monde. She resigned after just a year in charge, citing a lack of means to run the paper with the necessary peace and serenity. Her resignation happened alongside the forced departure of Jill Abramson from the New York Times, highlighting issues faced by women in media within the same timeframe.
Claim	Natalie Nougayrède 46 last year in front of Le Monde’s offices MIGUEL MEDINA/AFP/Getty Images but last week was invited for an award for leadership.
Question	Based on the provided information, please determine whether the claim is factual.
Choice	A. True B. False C. Partially True D. Not Verifiable

II. RELATED WORK

Fact-checking is a crucial component in the news reporting process. Journalism has been defined by scholars as a “discipline of verification” to separate it from “entertainment, propaganda, fiction, or art” [2]. While fact-checking and verification are often used interchangeably, recent work has sought to define them as two distinct but complementary processes [3]. In particular, verification is defined as “a scientific-like approach of getting the facts and also the right facts” [4], which often involves verifying the source, date and the location of materials. Fact-checking on the other hand “addresses the claim’s logic, coherence and context” [5]. As a foundational component of the fact-checking framework, verification plays a critical initial role by systematically evaluating the credibility of the contextual evidence under consideration.

Although many works have been conducted on fact-checking from MLLMs [6], [7]; identifying the severity or nature of false claims, i.e., distinguishing between hoaxes, rumors and satire [8], [9]; detecting check-worthy claims [10],

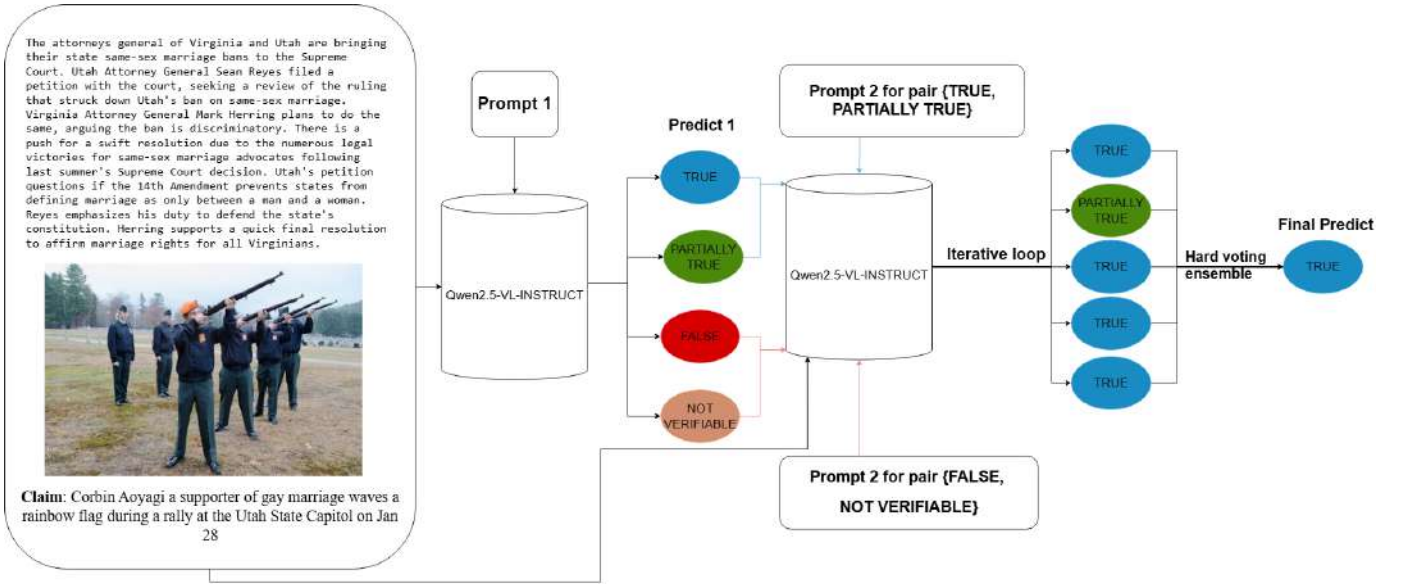


Fig. 1. **Pipeline for multi-turn claim verification with hard-vote ensemble.** The system takes as input an image, contextual information, and a claim to be verified. In the initial step, the input is passed through a model using Prompt 1 (illustrated in code IV-A), which outputs a preliminary label from four categories: "True", "False", "Partially True", or "Not Verifiable". Based on this initial prediction, a corresponding Prompt 2 (illustrated in code IV-B) is selected—tailored for the specific label pair—and the model is queried four additional times with variant formulations. The resulting outputs form a prediction vector, which is aggregated using a hard-voting ensemble strategy to determine the final verification label.

[11]; distinguishing between hyper-partisan and mainstream news [12]; stance detection and opinion analysis in fake news [13]; fauxtography, i.e., verifying the captions or claims which accompany published images [14]–[16].

These prior studies have inspired our research and guided the development of a dedicated architecture for this challenge. The system takes as input an image, contextual information, and a claim to be verified. In the initial step, the input is passed through a model using Prompt 1 (illustrated in code IV-A), which outputs a preliminary label from four categories: "True", "False", "Partially True", or "Not Verifiable". Based on this initial prediction, a corresponding Prompt 2 (illustrated in code IV-B) is selected—tailored for the specific label pair—and the model is queried four additional times with variant formulations. The resulting outputs form a prediction vector, which is aggregated using a hard-voting ensemble strategy to determine the final verification label.

III. DATASET

In this paper, we tackle the second task of the ACMMM 25: Multimodal Hallucination Detection and Fact Checking. The dataset is divided into two subsets: **Development (Dev)** and **Test**, each annotated with four categorical labels indicating the veracity of claims: *True*, *False*, *Partially True*, and *Not Verifiable*.

A. Development Set

The development set comprises a total of **200** instances, distributed across the four veracity labels as follows:

- **True:** 51 instances
- **False:** 59 instances

- **Partially True:** 31 instances
- **Not Verifiable:** 59 instances

This subset is primarily utilized for model training and hyperparameter tuning.

B. Test Set

The test set contains **1000** instances, serving as the primary benchmark for evaluating system performance. The label distribution in the test set is as follows:

- **True:** 253 instances
- **False:** 294 instances
- **Partially True:** 159 instances
- **Not Verifiable:** 294 instances

The relatively balanced distribution among the *False* and *Not Verifiable* categories ensures that the evaluation reflects model robustness in handling both falsified and unverifiable information.

Overall, the dataset is constructed to support the development and evaluation of fact verification systems, particularly in the context of misinformation detection and automatic claim validation.

IV. METHOD

In the context of multimodal fact-checking, the task is framed as a multiple-choice classification problem, where the system is required to determine the most appropriate veracity label for a given claim. The available options include: (A) True, (B) False, (C) Partially True, and (D) Not Verifiable. The primary objective is to assess the factual consistency of a given image and its associated textual context, and subsequently determine the appropriate classification label.

To tackle this challenge effectively, we proposed a multi-stage prompting framework designed to improve the model’s semantic comprehension of the multimodal input as shown in Fig 1.

A. Initial generation result

To guide the model in the multimodal fact-checking task, we design a structured prompt that clearly defines its role and decision-making procedure as follows:

```
[ROLE INSTRUCTION]
The CLAIM: {claim}
The CONTEXT: {context}
The IMAGE: {image}
Based on the provided information, please
determine whether the claim is factual.
A. True.
B. False.
C. Partially True.
D. Not verifiable.
[Reasoning] (optional)
The answer is:
```

The model is instructed to act as an effective fact-checking system capable of analyzing multimodal inputs, which include: (1) a textual claim (e.g., a news headline, a sentence from an article, or a social media post), (2) an associated image relevant to the claim, and (3) additional textual context, such as the full article, a related discussion thread, or other supplementary information. The task requires the model to assess the factual accuracy of the claim based on all provided inputs and assign one of four possible truthfulness labels.

To ensure that the model understands its specific task of selecting one of four answer options (A, B, C, or D), corresponding to "True", "False", "Partially True", and "Not Verifiable", respectively, we design the Role Instruction prompts and Reasoning as follows:

Long Role:

```
You are an effective multi-modal fact-checking
model. Your task is to verify the factual
accuracy of claims by analyzing multimodal
inputs. The given inputs include: a claim in
textual form, which can be a news headline, a
sentence from an article, or a social media
post; an accompanying image related to the
claim; and an additional context in textual
form, such as the full text of the news article
, a related social media discussion, or other
supplementary information. You must determine
the factual accuracy of the claim based on all
the provided inputs. They need to assign one of
four possible labels to the claim: \"True\",
\"False\", \"Partially True\", and \"Not
Verifiable\" by choosing one of the following
letters: A, B, C, D . Answer which must be in
the format of a single letter (A, B, C or D).
```

Short Role:

```
You are an effective multi-modal fact-checking
model. Answer which must be in the format of a
single letter (A, B, C or D). Do not include
any other information in your answer.
```

Reasoning:

```
Think step by step as follows:
Step 1: Check if the claim is verifiable or not.
If it is not verifiable, then the answer is D.
Step 2: If it is verifiable, then check if the
claim is true or false.
Step 3:
- Case 1: If it is true, check it Partially True
or True. Then if it is Partially True the
answer is C, if it is True the answer is A.
- Case 2: If it is false, check it Partially
True or False. Then if it is Partially True the
answer is C, if it is False the answer is B.
Step 4: Give back your answer which must be in
the format of a single letter (A, B, C or D)
inside <answer> tag like this: \"The answer is:
<answer>[ANSWER]</answer>\". Do not include
any other information in your answer.
```

B. Refined generation results

To guide the model in the multimodal fact-checking task, we design a structured prompt that clearly defines its role and decision-making procedure as follows:

```
[ROLE INSTRUCTION]
CLAIM: {claim}
CONTEXT: {context}
DEFINITION: [definition]
The IMAGE: {image}
Based on the provided information, please
determine whether the claim is factual.
[CHOICES]
The answer is:
```

At the refine stage of the multimodal fact-checking process, we instruct the model to select only one answer from two predefined pairs: ("True", "Partially True") and ("False", "Not Verifiable"), the following prompt similar to the other case:

```
You are an effective multi-modal fact-checking
model. Answer which must be in the format of a
single letter (A or C). Do not include any
other information in your answer
```

With the closest definition for each choice:

```
Definition:
- The choice is A - True if it is fully
supported by the CONTEXT. All critical facts
are correct, clearly stated or strongly implied
. There are no misleading or incorrect elements
- The choice is C - Partially True if it is
partially supported by the CONTEXT. Some parts
are correct, but others are missing context,
oversimplified, or inaccurate. It mixes both
correct and incorrect elements.
```


And for the other case:

Definition:

- The choice is B - False if it is clearly contradicted by the CONTEXT or includes factually incorrect information. It misrepresents the content or introduces inaccurate claims.
- The choice is D - Not Verifiable if it cannot be confirmed or refuted using only by the CONTEXT. It may be plausible, but the sample does not provide enough explicit evidence to support or deny it.

This approach enables the model to gain a more comprehensive perspective and deeper understanding of the task requirements, thereby enhancing its reasoning ability and improving the quality of the final prediction, as shown in Fig 2 and Fig 3.

C. Parameters

Firstly, the temperature parameter is adjusted to identify the optimal value that yields the highest performance, as illustrated in Fig 5. The line chart demonstrates that the model achieves the highest accuracy at *temperature* = 0.9, while the lowest accuracy is observed at *temperature* = 0.5 and 0.8. The results indicate that tuning this parameter has a noticeable impact on the model’s accuracy. Based on these findings, we select *temperature* = 0.9 as the optimal setting for our method.

During the experimental process, as illustrated in Fig 6, the model achieved the highest accuracy at *iteration* = 4. Specifically, the results consist of five predictions: one initial prediction and four subsequent predictions generated through iterative refinement using the proposed prompt refinement strategy IV-B. Finally, we applied the hard majority voting method to select the most frequently predicted label as the final output.

V. RESULT

A. Result in Development set

Based on the results observed in the development set, we chose the **Short Role** prompt to generate the best initial results and then applied our method by conducting experiments in a development set consisting of 200 candidates. As observed in Fig 2, the model exhibits a clear tendency to confuse the label "Partially True" with other classes. Specifically, the model struggles to effectively distinguish between the classes "True" and "Partially True." Moreover, it frequently misclassifies instances of "Not Verifiable" as "Partially True" or "False." Additionally, the model tends to confuse "False" with "Partially True."

These observations motivate a grouping strategy: we propose to cluster predictions into two groups based on the model’s initial output — Group 1 includes labels "A" (True) and "C" (Partially True), while Group 2 includes "B" (False) and "D" (Not Verifiable). The model is then restricted to selecting only between the two corresponding labels within

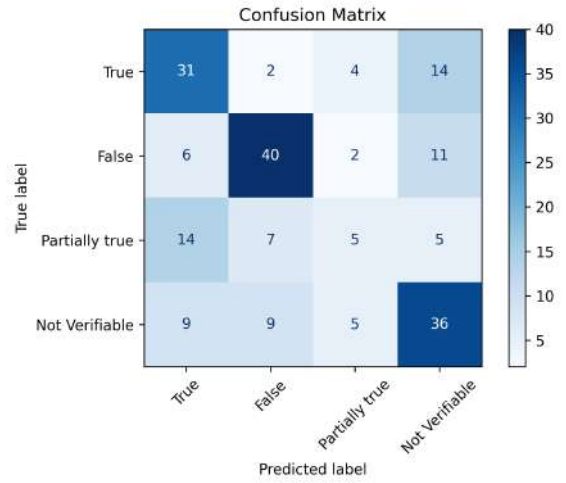


Fig. 2. Confusion Matrix on development set using Short Role prompt.

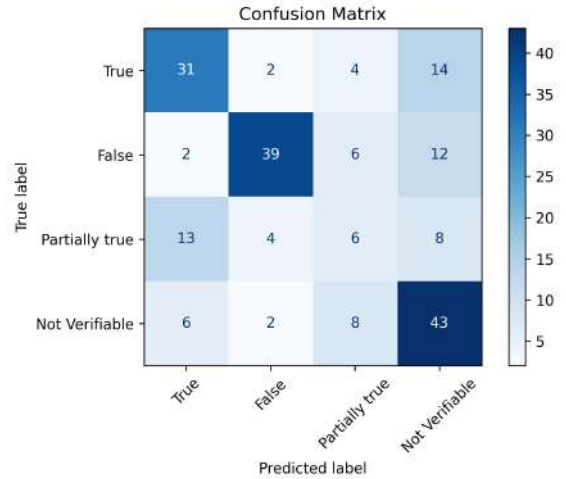


Fig. 3. Confusion matrix on development set when using our method.

each group. That is, if the prediction falls into Group 1, the model chooses between "A" and "C"; if into Group 2, between "B" and "D."

The goal of this approach is to improve class separability by reducing the confusion between these commonly misclassified categories. This constrained selection is further supported by a majority voting mechanism to enhance the consistency of the final decision, shown in Fig 3.

With this approach applied to the development set, we observed a significant improvement by comparing the effectiveness of our method with the initial result (baseline). Fig 4 shows an improvement of 3.5% in accuracy in Long Role and Short Role Instruction prompt, 7% in Long Role with Reasoning Instruction prompt, and 1.5% in Short Role with Reasoning Instruction prompt. The model demonstrated enhanced capability in distinguishing between both critical label pairs: ("True", "Partially True") and ("False", "Not Verifiable").

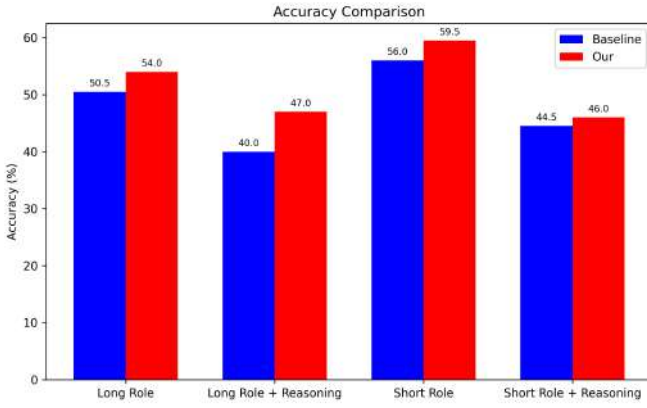


Fig. 4. Accuracy Comparison between baselines and applying our method

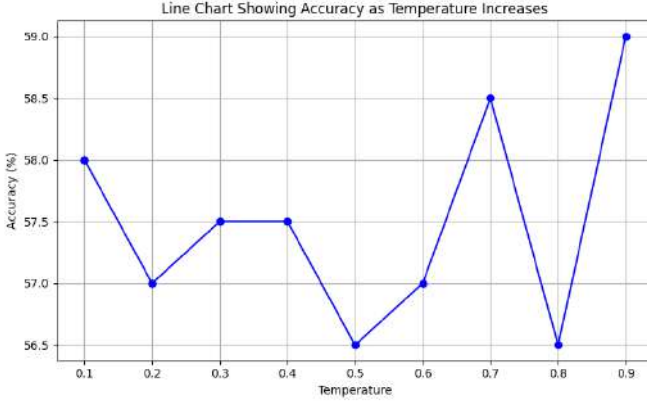


Fig. 5. Line Chart Showing Accuracy as Temperature Increases

B. Result in Test set

The test set contains 1,000 samples, and the results show patterns similar to those observed in the development set.

The results in Fig 7 indicate that the model is most effective at clearly distinguishing claims labeled "True" and "False." However, we also observe a persistent misclassification tendency between the two label pairs: ("True", "Partially True") and ("False", "Not Verifiable").

By applying our proposed method, the model demonstrates improved discrimination across both pairs—particularly in the first pair, where a more noticeable enhancement is observed as in Fig 8. Furthermore, we observed a notable shift in the model’s prediction tendencies, wherein instances previously classified as "Partially True" were increasingly labeled as "True", and those labeled as "False" were more frequently reclassified as "Not Verifiable".

With this approach applied to the test set, the Table II shows that our method results in a 0.6% decrease in overall accuracy but an increase of 1.7% in macro F1 when using the 3B version. On the other hand, the table shows that our approach decreased 2.5% in overall accuracy and decreased 2.1% in macro F1 with version 7B of the model. Also, with the 32B model version, our method improved 0.6% in overall

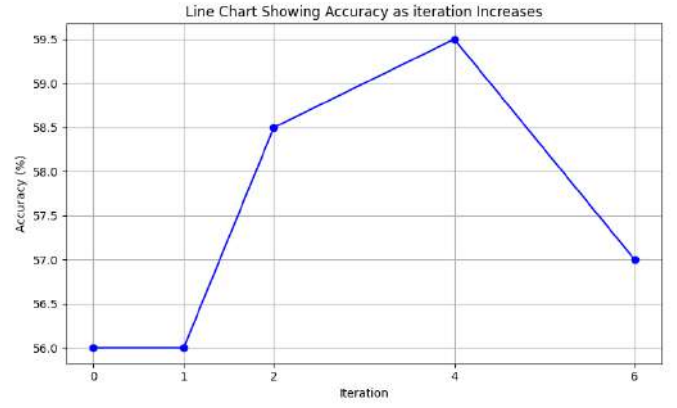


Fig. 6. Line Chart Showing Accuracy as Iteration Increases

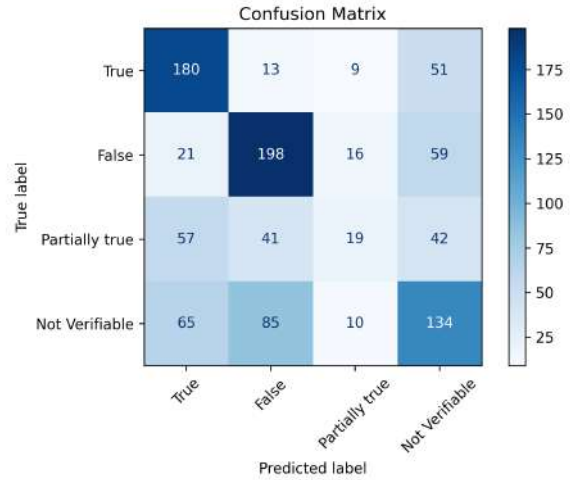


Fig. 7. Confusion Matrix on test set using Short Role prompt.

accuracy and 3.04% in macro F1. Based on the data in this table, it appears that the effectiveness of the method becomes more pronounced as the model size increases.

TABLE II
EXPERIMENTAL RESULTS ON THE TEST SET

Settings	Acc (%)	macro F1 (%)
Zero-shot		
Qwen2.5VL-3B-Instruct (baseline)	39.60	32.44
Qwen2.5VL-3B-Instruct (our)	39.00	34.10
Qwen2.5VL-7B-Instruct (baseline)	42.30	34.15
Qwen2.5VL-7B-Instruct (our)	39.80	32.05
Qwen2.5VL-32B-Instruct (baseline)	53.10	47.33
Qwen2.5VL-32B-Instruct (our)	53.70	50.37

VI. CONCLUSION

In this paper, we proposed a structured approach to improve model performance in the multimodal fact-checking classification task. Our method, which integrates guided prompts and a multi-stage reasoning process, has shown promising results, particularly with a significant increase in accuracy on

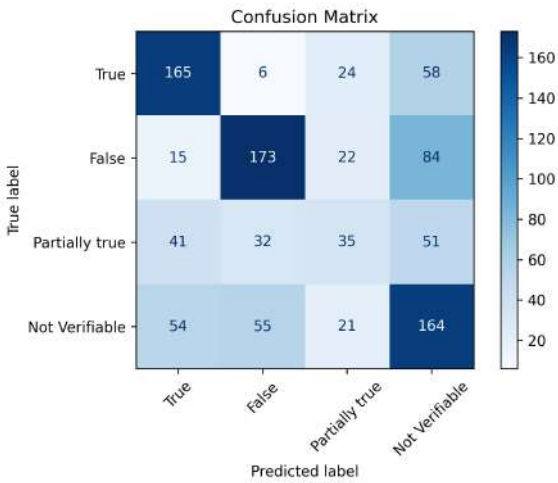


Fig. 8. Confusion matrix on test set when using our method.

the development set. This suggests that decomposing the task and incorporating semantic understanding through LLMs and MLLMs can contribute meaningfully to the model's ability to distinguish between fine-grained labels.

Despite these improvements, we also observed several limitations that highlight opportunities for further enhancement. In particular, the design of the instruction prompt and the model's ability to perform consistent and accurate reasoning still exhibit weaknesses. These issues occasionally lead to incorrect or inconsistent outputs, especially in borderline cases where label definitions are subtle or contextually complex.

Moreover, while our approach yielded a marked improvement on the development set, the gain on the unseen test set was only marginal. This discrepancy indicates potential overfitting to the development data or a lack of robustness in generalization, suggesting that our method, although effective in controlled settings, may still struggle with domain shifts or unanticipated context variations. It also implies that the reasoning framework and prompt construction strategy require further refinement to better align with the diversity of real-world claims.

Looking forward, our next steps include optimizing the model's reasoning process through more adaptive or dynamic prompt strategies, as well as exploring iterative refinement mechanisms in a more principled manner. Additionally, we plan to evaluate the scalability and generalizability of our approach by applying it to other benchmark datasets and integrating it with more advanced or specialized models. Through these efforts, we aim to better understand the limitations of current multimodal fact-checking systems and push toward more robust, interpretable, and accurate solutions.

REFERENCES

[1] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhunoye, Y. Yang, S. Gupta, B. Prasad Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, "Self-Refine: Iterative Refinement with Self-Feedback," *arXiv e-prints*, p. arXiv:2303.17651, Mar. 2023.

[2] J. Thorne and A. Vlachos, "Automated fact checking: Task formulations, methods and future directions," in *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3346–3359. [Online]. Available: <https://aclanthology.org/C18-1283/>

[3] C. Silverman, "Verification handbook: Additional materials," <http://verificationhandbook.com/additionalmaterial/>, 2016, accessed: 2018-09-30.

[4] B. Kovach, *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*, T. Rosenstiel, Ed., New York, 2014.

[5] A. Mantzarlis, "Will verification kill fact-checking?" <https://www.poynter.org/news/will-verification-kill-fact-checking/>, 2015, accessed: 2018-09-30.

[6] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 809–819. [Online]. Available: <https://aclanthology.org/N18-1074/>

[7] L. Tang, P. Laban, and G. Durrett, "MiniCheck: Efficient fact-checking of LLMs on grounding documents," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8818–8847. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.499/>

[8] V. L. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in *Proceedings of the second workshop on computational approaches to deception detection*, 2016, pp. 7–17.

[9] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.

[10] N. Hassan, C. Li, and M. Tremayne, "Detecting check-worthy factual claims in presidential debates," in *Proceedings of the 24th acm international on conference on information and knowledge management*, 2015, pp. 1835–1838.

[11] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1803–1812.

[12] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*, 2017.

[13] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, "A retrospective analysis of the fake news challenge stance detection task," *arXiv preprint arXiv:1806.05180*, 2018.

[14] D. Zlatkova, P. Nakov, and I. Koychev, "Fact-checking meets fauxtography: Verifying claims about images," *arXiv preprint arXiv:1908.11722*, 2019.

[15] N. Kotonya and F. Toni, "Explainable Automated Fact-Checking: A Survey," *arXiv e-prints*, p. arXiv:2011.03870, Nov. 2020.

[16] H. Chu, H. Chu, T.-M. Nguyen, S. T. Luu, C. Hoang, H. Nguyen, V. Tran, and L.-M. Nguyen, "DeepSix at acm mm 2025 grand challenge: Enhancing context text processing for multimodal hallucination detection and fact verification," in *Proceedings of the 33rd ACM International Conference on Multimedia*. Dublin, Ireland: Association for Computational Linguistics, Oct. 2025, p. 27–31. [Online]. Available: <https://doi.org/10.1145/3746027.3762061>