PROGRAM BOOK

THE 14TH INTERNATIONAL SYMPOSIUM ON INFORMATION AND COMMUNICATION TECHNOLOGY

Honoring 30 Years Of
SCHOOL OF INFORMATION AND COMMUNICATIONS TECHNOLOGY, HUST

12-14 December, 2025 | Nha Trang, Vietnam

TECHNICAL SPONSORS

FINANCIAL SPONSORS

TABLE OF CONTENT

SOICT 2025

FOREWORD

The 14th International Symposium on Information and Communication Technology (SOICT 2025) is held on December 12-14, 2025, in Nha Trang City, Vietnam. SOICT 2025 is an international academic forum for researchers and graduate students to share their latest research findings and to identify future challenges in computer science.

SOICT 2025 has received papers from 30 countries and regions in six major areas of research including Networking and Communication Technologies, AI Foundation and Big Data, AI Applications, Multimedia Processing, Software Engineering, and Recent Advances in Cyber Security, in addition to special sessions on Applied Operations Research and Optimization, Generative AI, Human Computer Interaction and Intelligent Interactive Systems, and Lifelog Event Retrieval. The Program Committee has followed a formal standard reviewing process; bidding, reviewing, and deliberating for selecting 118 papers for regular presentation, and 117 papers for poster presentation and publication in the proceedings. It is our great honor to receive world-class invited speakers: TUNG Kum Hoe Anthony (National University of Singapore, Singapore), Josiah Poon (University of Sydney, Australia), Vincent Wong (The University of British Columbia, Canada), and John C.S. Lui (The Chinese University of Hong Kong, Hong Kong).

We would like to thank the Program Committee members for their great responsibility in reviewing papers, and all the track chairs for actively monitoring the review and deliberation process and for proposing decisions on papers: Abdelhamid Mellouk (University of Paris-Est (UPEC), France), Mikael Gidlund (Mid Sweden University, Sweden), The Ngoc Dang (Posts & Telecommunications Institute of Technology, Vietnam), Thien Huynh-The (Ho Chi Minh City University of Technology and Education, Vietnam), Massimo Zancanaro (University of Trento, Italy), Katsumi Inoue (National Institute of Informatics, Japan), Thanh H. Nguyen (University of Oregon, United States), Tan Nguyen (National University of Singapore, Singapore), Cam-Tu Nguyen (Nanjing University, China), Xinyue Hao (Cardiff University, United Kingdom), Tae-Seong Kim (Kyung Hee University, Korea), Duc Thanh Nguyen (Deakin University, Australia), Thi Hoang Ngan Le (University of Arkansas, United States), Le duc Hau (Hanoi University of Science and Technology, Vietnam), Le Duy Dung (VinUniversity, Vietnam), Nguyen Van Tan (University of Dayton, United States), Duc Tien Dang Nguyen (University of Bergen, Norway), Habib Ullah (Norwegian University of Life Sciences, Norway), Mehdi Elahi (University of Bergen, Norway), Hongjo Kim (Yonsei University, Korea), Gerard Chalhoub (University of Clermont Auvergne, France), Sharif Abuadbba (CSIRO's Data61, Australia), Viet Vo (Swinburne University of Technology, Australia), Van-Hau Pham (University of Information Technology – VNU, Vietnam), Tran Hai Anh (Hanoi University of Science and Technology, Vietnam), Emrah Demir (Cardiff University, UK), Nguyen Viet Hung(Clermont Auvergne University, France), Nguyen Tuan Binh (Vin University, Vietnam), Vu Duc Minh (National Economics University, Vietnam), Pekka Abrahamsson (Tampere University, Finland), Quan Thanh Tho (Ho

FOREWORD

Chi Minh City University of Technology, Vietnam), Dinh Viet Sang (Hanoi University of Science and Technology, Vietnam), Shengdong Zhao (City University of Hong Kong, Hong Kong), Liting Zhou (Dublin City University, Ireland), Masitah Ghazali (Universiti Teknologi Malaysia, Malaysia), Chi-Thanh Vi (International University, VNUHCM, Vietnam), Vinh-Tiep Nguyen (University of Information Technology, VNUHCM, Vietnam), Trong-Le Do (University of Science, VNUHCM, Vietnam), Hai-Dang Nguyen (University of Science, VNUHCM, Vietnam), Tu V. Ninh (Dublin City University, Ireland), Tu-Khiem Le (Dublin City University, Ireland), Manabu Hagiwara (Chiba University, Japan), Martianus Frederic Ezerman (Nanyang Technological University, Singapore), Vu Van Khu (VinUni, Vietnam), Van-Duy Nguyen (Phenikaa University, Vietnam).

In particular, we would like to thank all Organizing Committee members, who have worked hard to ensure the best quality of the symposium.

We are grateful to Vingroup Innovation Foundation (VinIF) for financial support. Finally, we hope that the SOICT 2025 conference will provide an interesting and up-to-date scientific program. We would like to thank all authors and participants for making SOICT 2025 a memorable and enjoyable academic event in Nhatrang city, Vietnam.

Together, we make this SOICT 2025 Conference a successful event!

SOICT 2025 Program Co-Chairs

***Huynh Thi Thanh Binh**, Hanoi University of Science and Technology, Vietnam*

***Wray Buntine**, VinUniversity, Vietnam*

***Morten Fjeld**, University of Bergen, Norway & Chalmers University of Technology, Sweden*

***Klaus Schoffmann**, Klagenfurt University, Austria*

***Tran Minh Triet**, University of Science, VNUHCM, Vietnam*

***Tran The Truyen**, Deakin University, Australia*

SOICT 2025 General Co-Chairs

***Ho Tu Bao**, Vietnam Institute for Advanced Study in Mathematics, Vietnam*

***Cathal Gurrin**, Dublin City University, Ireland*

***Ichiro Ide**, Nagoya University, Japan*

***Ta Hai Tung**, Hanoi University of Science and Technology, Vietnam*

ORGANIZING COMMITTEE

HONORARY CHAIR:

Huynh Quyet Thang, Hanoi University of Science and Technology, Vietnam

GENERAL CHAIRS:

Cathal Gurrin, Dublin City University, Ireland
Ichiro Ide, Nagoya University, Japan
Ho Tu Bao, Vietnam Institute for Advanced Study in Mathematics, Vietnam
Ta Hai Tung, Hanoi University of Science and Technology, Vietnam

PROGRAM CHAIRS:

Wray Buntine, VinUniversity, Vietnam
Morten Fjeld, University of Bergen, Norway; Chalmers University of Technology, Sweden
Klaus Schöffmann, Klagenfurt University, Austria
Tran The Truyen, Deakin University, Australia
Tran Minh Triet, Vietnam National University Ho Chi Minh City, Vietnam
Huynh Thi Thanh Binh, Hanoi University of Science and Technology, Vietnam

TRACK CHAIRS:

Abdelhamid Mellouk, University of Paris-Est (UPEC), France
Mikael Gidlund, Mid Sweden University, Sweden
The Ngoc Dang, Posts & Telecommunications Institute of Technology, Vietnam
Thien Huynh-The, Ho Chi Minh City University of Technology and Education, Vietnam
Massimo Zancanaro, University of Trento, Italy
Katsumi Inoue, National Institute of Informatics, Japan
Thanh H. Nguyen, University of Oregon, United States
Tan Nguyen, National University of Singapore, Singapore
Cam-Tu Nguyen, Nanjing University, China
Xinyue Hao, Cardiff University, United Kingdom
Tae-Seong Kim, Kyung Hee University, Korea
Duc Thanh Nguyen, Deakin University, Australia
Thi Hoang Ngan Le, University of Arkansas, United States
Le Duc Hau, Hanoi University of Science and Technology, Vietnam
Le Duy Dung, VinUniversity, Vietnam
Nguyen Van Tam, University of Dayton, United States
Duc Tien Dang Nguyen, University of Bergen, Norway
Habib Ullah, Norwegian University of Life Sciences, Norway
Mehdi Elahi, University of Bergen, Norway
Hongjo Kim, Yonsei University, Korea
Gérard Chalhoub, University of Clermont Auvergne, France
Sharif Abuadbba, CSIRO's Data61, Australia
Viet Vo, Swinburne University of Technology, Australia
Van-Hau Pham, University of Information Technology – VNU, Vietnam
Tran Hai Anh, Hanoi University of Science and Technology, Vietnam
Emrah Demir, Cardiff University, UK
Viet Hung Nguyen, Clermont Auvergne University, France
Nguyen Tuan Binh, Vin University, Vietnam
Vu Duc Minh, National Economics University, Vietnam
Pekka Abrahamsson, Tampere University, Finland
Quan Thanh Tho, Ho Chi Minh City University of Technology, Vietnam
Dinh Viet Sang, Hanoi University of Science and Technology, Vietnam

ORGANIZING COMMITTEE

Shengdong Zhao, City University of Hong Kong
Liting Zhou, Dublin City University, Ireland
Masitah Ghazali, Universiti Teknologi Malaysia, Malaysia
Chi-Thanh Vi, International University, VNUHCM, Vietnam
Vinh-Tiep Nguyen, University of Information Technology, VNUHCM, Vietnam
Trong-Le Do, University of Science, VNUHCM, Vietnam
Hai-Dang Nguyen, University of Science, VNUHCM, Vietnam
Tu V. Ninh, Dublin City University, Ireland
Tu-Khiem Le, Dublin City University, Ireland
Manabu Hagiwara, Chiba University, Japan
Martianus Frederic Ezerman, Nanyang Technological University, Singapore
Vu Van Khu, Singapore University of Technology and Design, Singapore
Van-Duy Nguyen, Phenikaa University, Vietnam

TUTORIAL CHAIRS:

Ngo Duc Thanh, University of Information Technology, Vietnam
Vo Dinh Bay, HUTECH University of Technology, Vietnam
Than Quang Khoat, Hanoi University of Science and Technology, Vietnam

ORGANIZING CHAIRS:

Tran Quang Duc, Hanoi University of Science and Technology, Vietnam
Le Xuan Thanh, Hanoi University of Science and Technology, Vietnam
Tran Van Man, University of Science, VNU-HCM, Vietnam
Ngo Dai Nghiep, University of Science, VNU-HCM, Vietnam
Ngo Lam Trung, Hanoi University of Science and Technology, Vietnam
Pham Minh Phuong, University of Science, VNUHCM, Vietnam
Khanh-Duy Le, University of Science, VNUHCM, Vietnam

PUBLICATION CHAIRS:

Dinh Anh Dung, University of Sydney, Australia
Dang Tuan Linh, Hanoi University of Science and Technology, Vietnam
Dinh Thi Ha Ly, Hanoi University of Science and Technology, Vietnam
Nguyen Thi Oanh, Hanoi University of Science and Technology, Vietnam
Dam Quang Tuan, Hanoi University of Science and Technology, Vietnam
Nguyen Ngoc Thao, University of Science, VNUHCM, Vietnam

PUBLICITY CHAIRS:

Trinh Van Chien, Hanoi University of Science and Technology, Vietnam
Huynh Viet Tham, University of Science, VNU-HCM (VNUHCM-US), Vietnam
Trinh Thanh Trung, Hanoi University of Science and Technology, Vietnam

INDUSTRIAL SESSION CHAIRS:

Pham Ngoc Hung, Hanoi University of Science and Technology, Vietnam

WEB CHAIRS:

Nguyen Quoc Khanh, Hanoi University of Science and Technology, Vietnam
Hoang Viet Dung, Hanoi University of Science and Technology, Vietnam

SOICT 2025 INFORMATION

CONFERENCE VENUE

Sheraton Nha Trang Hotel & Spa
26-28 Tran Phu Street, Nha Trang City, Vietnam

REGISTRATION DESK OPENING TIME

Friday, 12 December 2025 | 07:30 - 18:00
Saturday, 13 November 2025 | 07:30 - 17:30
Location: Foyer, Level 2, Sheraton Nha Trang Hotel & Spa

FUNCTION ROOMS

Level 2

- Grand Ballroom - Grand Ballroom A - Grand Ballroom B
- Yersin Ballroom A - Yersin Ballroom B

REFRESHMENTS

Tea breaks are arranged in the foyer outside the meeting rooms
Buffet lunch is served at Feast Restaurant on the Level 1

INTERNET ACCESS

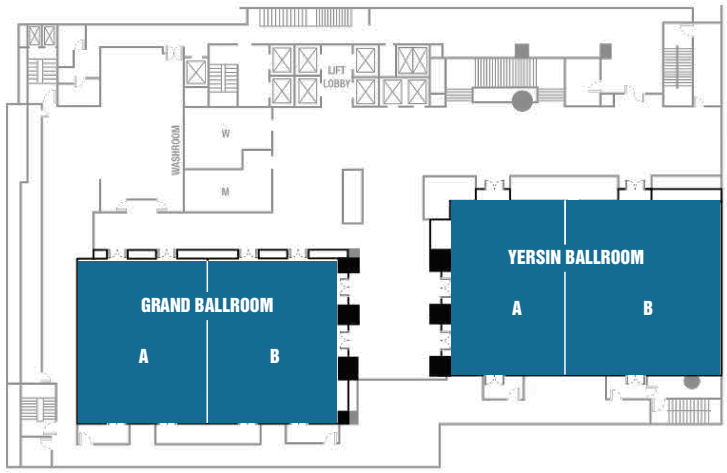
Complimentary Wi-Fi is available throughout the conference venue
Network Name (SSID): Marriott Bonvoy
Password: Not required

GALA DINNER

Venue: Lakshmi Hall 1 - Champa Island Nha Trang - Resort Hotel & Spa
Address: 304, 2/4 Road, Bac Nha Trang Ward, Khanh Hoa Province
Date & Time: Saturday, 13 November 2025 | 19:00 - 22:00
Bus pickup time at Sheraton Nha Trang Hotel & Spa: 18:30

SOICT 2025 LAYOUT

LEVEL 2



PROGRAM AT A GLANCE

DAY 1		Friday, 12 December 2025
07:45 - 18:00	Registration	Foyer, 2F
08:30 - 08:50	Conference Opening	Grand Balllroom, 2F
08:50 - 09:30	**Keynote I** Machine Learning for Integrated Sensing and Communication **Vincent Wong**	Grand Balllroom, 2F
09:30 - 10:10	**Keynote II** Quantum Internet: The Final Frontier **John C.S. Lui**	Grand Balllroom, 2F
10:10 - 10:40	Tea Break	Foyer, 2F
10:40 - 12:00	**SOICT Technical Session 1** Quantum Information	Grand Ballroom A, 2F
	SOICT Technical Session 2 AI Applications	Grand Ballroom B, 2F
	SOICT Technical Session 3 Software Engineering	Yersin Ballroom A, 2F
	OICT Technical Session 4 Networking and Communication Technologies	Yersin Ballroom B, 2F
	Poster Exhibition AI Applications	Foyer, 2F
12:00 - 13:30	Lunch	Feast Restaurant, 1F
13:30 - 18:00	**Poster Exhibition** AI Applications, AI Foundations and Big Data, Applied Operations Research and Optimization, Generative AI, Human Computer Interaction and Intelligent Interactive Systems	Foyer, 2F
13:30 - 15:30	**SOICT Technical Session 5** Generative AI	Grand Ballroom A, 2F
	SOICT Technical Session 6 AI Applications	Grand Ballroom B, 2F
	SOICT Technical Session 7 Applied Operations Research and Optimization	Yersin Ballroom A, 2F
	SOICT Technical Session 8 Multimedia Processing Chair: Khush Agarwal	Yersin Ballroom B, 2F
15:30 - 16:00	Tea Break	Foyer, 2F
16:00 - 18:00	**SOICT Technical Session 9** AI Applications/AI Foundations and Big Data	Grand Ballroom A, 2F
	SOICT Technical Session 10 AI Applications	Grand Ballroom B, 2F
	SOICT Technical Session 11 Applied Operations Research and Optimization	Yersin Ballroom A, 2F
	SOICT Technical Session 12 Networking and Communication Technologies / Software Engineering	Yersin Ballroom B, 2F

PROGRAM AT A GLANCE

DAY 2	Saturday, 13 December 2025	
08:00 - 17:40	Registration	Foyer, 2F
08:30 - 09:10	**Keynote III** Designing Multimodal Driven NLP for a Fluid World **Josiah Poon**	Grand Balllroom, 2F
09:10 - 09:50	**Keynote IV** Being Small in the Era of Large Models: Enabling Prudent AI with Lightweight, Just-in-Time AI Boxes **Tung Kum Hoe Anthony**	Grand Balllroom, 2F
09:50 - 10:20	Tea Break	Foyer, 2F
10:20 - 12:00	**SOICT Technical Session 13** Lifelog Event Retrieval	Grand Ballroom A, 2F
	SOICT Technical Session 14 AI Applications	Grand Ballroom B, 2F
	SOICT Technical Session 15 Human Computer Interaction and Intelligent	Yersin Ballroom A, 2F
	SOICT Technical Session 16 Multimedia Processing	Yersin Ballroom B, 2F
	Poster Exhibition Lifelog Event Retrieval	Foyer, 2F
12:00 - 13:30	Lunch	Feast Restaurant, 1F
13:30 - 17:00	**Poster Exhibition** Lifelog Event Retrieval, Multimedia Processing, Communication Technologies, Quantum Information, Recent Advances in Cyber Security, Software Engineering	Foyer, 2F
13:30 - 15:30	**SOICT Technical Session 17** Lifelog Event Retrieval	Grand Ballroom A, 2F
	SOICT Technical Session 18 AI Applications	Grand Ballroom B, 2F
	SOICT Technical Session 19 Recent Advances in Cyber Security	Yersin Ballroom A, 2F
	SOICT Technical Session 20 Multimedia Processing	Yersin Ballroom B, 2F
15:30 - 16:00	Tea Break	Foyer, 2F
16:00 - 17:00	**SOICT Technical Session 21** Lifelog Event Retrieval	Grand Ballroom A, 2F
	SOICT Technical Session 22 AI Applications	Grand Ballroom B, 2F
	SOICT Technical Session 23 Recent Advances in Cyber Security	Yersin Ballroom A, 2F
	SOICT Technical Session 24 Multimedia Processing	Yersin Ballroom B, 2F
19:00 - 21:30	Gala Dinner	Champa Island Nha Trang

08:50 - 09:30 | Grand Ballroom - 2F

KEYNOTE I:

Machine Learning for Integrated Sensing and Communication

Prof. VINCENT WONG

The University of British Columbia, Canada

ABSTRACT

Integrated sensing and communication (ISAC) is a key technology for the sixth-generation (6G) wireless networks, where the same spectral and hardware resources are used for both communication and environmental sensing. Many optimization problems in ISAC require accurate sensing and communication channel models, which are often difficult to obtain. Machine learning (ML) is a powerful tool for solving ISAC problems by enabling data-driven solutions that can bypass the reliance on explicit models. This talk will explore how ML techniques can improve ISAC performance beyond traditional optimization approaches. Two case studies will be discussed: sensing-assisted predictive beamforming and cooperative sensing through ML. These examples will demonstrate the potential of ML to enable end-to-end signal processing for ISAC in 6G wireless networks.

BIOGRAPHY

Vincent Wong is a Professor in the Department of Electrical and Computer Engineering at the University of British Columbia, Vancouver, Canada. His research areas include protocol design, optimization, and resource management of communication networks, with applications to the Internet, wireless networks, smart grid, mobile edge computing, and Internet of Things. Dr. Wong is the Editor-in-Chief of the IEEE Transactions on Wireless Communications. He is a Fellow of the IEEE, Canadian Academy of Engineering, and the Engineering Institute of Canada.

09:30 - 10:10 | Grand Ballroom - 2F

KEYNOTE II:

Quantum Internet: The Final Frontier

Prof. JOHN C.S. LUI

The Chinese University of Hong Kong, Hong Kong

ABSTRACT

In this talk, I will begin with a brief introduction to quantum computing, highlighting the importance and opportunities for pursuing fundamental research in the quantum Internet. In particular, I will discuss how quantum networks can enable quantum information transmission, parallel processing, and distributed processing. Next, I will introduce online learning theory and explain how it can help us explore compelling challenges in building quantum networks and the quantum Internet. To this end, I will delve into the quantum path selection problem, as well as the quantum border gateway protocol (QBGP) if time allows. Finally, I will outline several exciting open research problems at the intersection of quantum networks and quantum computing.

BIOGRAPHY

John C. S. Lui is the Choh-Ming Li Chair Professor in the Department of Computer Science and Engineering at The Chinese University of Hong Kong. His research interests include the quantum Internet and the theory and applications of online learning and optimization. He has served as a visiting professor at UCLA, Columbia University, Caltech, the University of Maryland, Purdue University, the University of Massachusetts Amherst, INRIA (France), and NII (Japan). He currently serves as a senior and associate editor for various IEEE and ACM Transactions and has received numerous best paper awards at IEEE and ACM conferences, as well as teaching awards from CUHK. He is a Fellow of the ACM and IEEE, and a Senior Research Fellow of the Hong Kong Research Grants Council (RGC) and the Croucher Foundation. His personal interests include films and general reading.

08:30 - 09:10 | Grand Ballroom - 2F

KEYNOTE III:

Designing Multimodal Driven NLP for a Fluid World

Assoc. Prof. JOSIAH POON

University of Sydney, Australia

ABSTRACT

Like water that adapts to any container, documents need Natural Language Processing (NLP) systems that adapt and move across modalities pages and scales to find verifiable evidence. In this talk, I will share a practical agenda to build NLP systems that ingest text, images, layout, tables and figures and produce traceable answers. We emphasise three pillars: integration, learning and retrieval. Integration: fuse multimodal features and layout aware encodings so text and visual content are interpreted together. Learning: train specialist teachers across modalities and distil their feature knowledge into compact deployable students for NLP tasks. Retrieval: adopt a retrieval first approach, using multipage and multimodal retrieval to find candidate passages, tables and figures, then chain those candidates into a clear evidence trail. I demonstrate how graph-based encodings and multiscale reasoning work together, and how multiteacher distillation compacts expert knowledge into deployable students. Then, with concise multimodal case studies and retrieval centric metrics, I show measurable gains in evidence grounding, generalisation and operational readiness. I conclude with practical measures to control complexity and annotation cost, and present simple experiments and evaluation criteria for different domains.

BIOGRAPHY

Associate Professor Josiah Poon leads a research group in the School of Computer Science at the University of Sydney. He specialises in machine learning, natural language processing and data driven decision support. The group investigates visually rich document understanding and multimodal AI and has produced widely used datasets, open toolkits and benchmarks that support academic and industrial projects. The work appears in leading conferences and journals across natural language processing, computer vision and health informatics and has influenced evaluation practices in document intelligence. The team focuses on building practical systems that deliver reliable evidence extraction, robust cross domain performance and efficient deployment. They collaborate with partners from diverse industries to translate research into operational solutions and to validate methods on real world tasks. In 2025 the group published a book Natural Language Understanding in Conversational AI with Deep Learning, which provides pragmatic guidance for building conversational systems in different domains.

09:10 - 09:50 | Grand Ballroom - 2F

KEYNOTE IV:

Being Small in the Era of Large Models:
Enabling Prudent AI with Lightweight, Just-in-Time AI Boxes

Prof. TUNG KUM HOE ANTHONY
National University of Singapore, Singapore

ABSTRACT

In an era dominated by massive foundation models, smaller players risk being left behind—unable to afford the scale, data, or manpower that large AI systems demand. This talk introduces the concept of Prudent AI—an approach that emphasizes right-sized, lightweight, and explainable intelligence delivered through just-in-time, Plug-and-Play AI Boxes. Focusing on applications like early anomaly detection in multivariate time series, we demonstrate how our AI Boxes use sparse data, minimal compute, and human-guided refinement to detect rare but critical events. The architecture integrates symbolic reasoning, data-driven refinement, and secure edge deployment, showing how being small can actually be a strength in resource-constrained settings. Through this, we reimagine how organizations can adopt AI that is transparent, agile, and sustainable.

BIOGRAPHY

Anthony K. H. Tung is currently a Professor in the Department of Computer Science, National University of Singapore (NUS). He received both his B.Sc.(2nd Class Honour) and M.Sc. in computer sciences from the National University of Singapore in 1997 and 1998 respectively. In 2001, he receives the Ph.D. in computer sciences from Simon Fraser University (SFU).

TECHNICAL PROGRAM

DAY 1 - FRIDAY, 12 DECEMBER 2025

07:45 - 18:00	Registration				
08:30 - 10:10	**GRAND BALLROOM**				
08:30 - 08:50	Conference Opening				
08:50 - 09:30	Keynote I: Vincent Wong (The University of British Columbia, Canada) Machine Learning for Integrated Sensing and Communication Chair: Eui-Nam Huh				
09:30 - 10:10	Keynote II: John C.S. Lui (The Chinese University of Hong Kong, Hong Kong) Quantum Internet: The Final Frontier Chair: Yukinobu Hoshino				
10:10 - 10:40	Tea Break				
07:45 - 18:00	**GRAND BALLROOM A**	**GRAND BALLROOM B**	**YERSIN BALLROOM A**	**YERSIN BALLROOM B**	**FOYER, 2F**
10:40 - 12:00	SOICT Technical Session 1 Quantum Information Chair: Nguyen Hung Son	SOICT Technical Session 2 AI Applications Chair: Namal Rathnayake	SOICT Technical Session 3 Software Engineering Chair: Jonathan Hoyin Chan	SOICT Technical Session 4 Networking and Communication Technologies Chair: Mellouk Abdelhamid	POSTER EXHIBITION AI Applications
10:40 - 11:00	*Thuy Dao Thi Thu, Dat Le Quocand Linh Hoang Dinh* Quantum Circuit Resource Assessment for ChaCha20 Stream Cipher	*Thai Anh Vo, Lan Anh Nguyen, Bao Do, Cong Thanh Ma, Son-Trung Doan and Son-Hong Ngo* TriFusion: GNN-Based Multimodal Fusion for 3D Object Detection in Autonomous Driving	*Kaung Myat Kyaw, Jonathan Chan and Udom Silparcha* CandleGen: Generating Synthetic OHLC Data for Different Market Trends using GANs	*Long Khanh Tran, Thiloan Bui, Viet Hieu Ha, Van Dat Nguyen, Dinh Dung Nguyen and Thi Ngoc Dung Kieu* Propagated Presence: A Bluetooth Propagation-Based Method for Automated Classroom Attendance on Mobile Devices	
11:00 - 11:20	*Trieu Nguyen, Minh Chung, Thanh-Dang Diep, and Nam Thoai* EDM4QS: An Emulator-Driven Model for Quantum Scheduling	*Nhat-Nam Duong, Trung-Kien Dao and Dinh-Van Nguyen* A Novel Approach for Sino-Vietnamese Text Transcription by Leveraging a Pre-trained BERT and Self-Attention Mechanism	*Thi-Ha Le, Quoc-Hung Pham, Huy-The Vu, Minh-Tien Nguyen and Xuan-Hieu Phan* Graph-based Multi-Agents for Text-to-SQL	*Quang Huy Duong and Brigitte Jaumard* An Evaluation on Defragmentation with CDC ROADMs in Elastic Optical Networks	
11:20 - 11:40	*Lan Anh Nguyen, Thai Anh Vo, Trang Nguyen, Son-Trung Doan, Toan-Duc Nguyen, Son-Hong Ngo and Eunsung Jung* Toward Acceleration of Variational Quantum Classifier Simulation on GPUs	*Duong Nguyen Vu Binh, Duy-Thanh Vu, Duy-Cat Can, Hai Dac Nguyen, Binh T. Nguyen, Oliver Y. Chén, and Huong Ha* A Comparison of Machine Learning Methods for Alzheimer's Disease Classification in Vietnamese Patients	*Long Nguyen, Quynh Vo, Thi Nguyen and Tho Quan* URAG 2.0: An Agentic Dual Retrieval Framework for Enhanced Reasoning in RAG-based QA Systems	*Thanh-Dat Tran, Son Truong, Hoc Phan, Hai-Trang Dang, and Thien Huynh-The* Fusing Gated Spatial-Channel Units and Fractal Cross- Scale Attention for Lightweight Waveform Classification	
11:40 - 12:00	*Hieu Nguyen Doan, Duc Nguyen Manh, Thu Ngo Tran Anh, Tung Nguyen Xuan, Chien Trinh Van and Hwang Won Joo* Performance Analysis of Quantum Federated Learning with Personalized Layer	*Binh-Dang Le, Quan Thi Khac, Duy Tran Ngoc Bao, and Thanh Van Le* CodeLit: A Skill-Based Framework for Automated Assessment of Code Comprehension	*Huynh Ngoc Khoa, Tang Nhat Hung, Dang Thien Binh, and Nguyen Thanh Binh* Boosting Test Smell Prediction Using Deep Learning	*Minh-Cat-Tuong Nguyen, Si-Thai Trang and Chi-Thanh Vi* A mobile-based attendance system using Bluetooth MAC address scanning	

12:00 - 13:30	LUNCH (Feast Restaurant, Level 1 - Sheraton Nha Trang Hotel & Spa)				
	GRAND BALLROOM A	**GRAND BALLROOM B**	**YERSIN BALLROOM A**	**YERSIN BALLROOM B**	**FOYER, 2F**
13:30 - 15:30	SOICT Technical Session 5 Generative AI Chair: Jian Gang Ngui	SOICT Technical Session 6 AI Applications Chair: Guy Nagels	SOICT Technical Session 7 Applied Operations Research and Optimization Chair: Markcus Westner	SOICT Technical Session 8 Multimedia Processing Chair: Khush Agarwal	POSTER EXHIBITION AI Applications, AI Foundations and Big Data, Applied Operations Research and Optimization, Generative AI, Human Computer Interaction and Intelligent Interactive Systems
13:30 - 13:50	*Industrial Talk* *Jian Gang Ngui (AI Singapore, Singapore)* SEA-LION: Southeast Asian Languages in One Network	*Vu Pham and Long Nguyen* Optimization Approaches for Language Models in the Task of Translating Sino-Vietnamese Texts into Modern Vietnamese	*Hoang Giang Pham and Thuy Anh Ta* Exponential Cone Reformulation for Scalable Estimation of Quantal Response and Multinomial Logit Models	*Thao Thi Phuong Dao, Tan-Cong Nguyen, Trong-Le Do, Mai-Khiem Tran, Minh-Khoi Pham, Trung-Nghia Le, Minh-Triet Tran and Thanh Dinh Le* DTD-Mamba: Dual Teacher Distillation for Mamba in Head and Neck Abscess Segmentation	
13:50 - 14:10	*Huyen Nguyen, Hieu Dam, Cong Tran, and Cuong Pham* AD-GENESIS: Anomaly Detection through Gradient-Guided Generative Synthesis	*Nam Nguyen Tu and Hiroki Takahashi* Motion-Gated Adaptive Filtering for Continuous Sign Language Recognition	*Ban Ha-Bang and Do Tuan-Anh* Reinforcement Learning-Enhanced GRASP for the Multiple Traveling Repairmen Problem with Workload Balance	*Trong-Hieu Nguyen-Mau, Minh-Nam Tran, Kim-Trang Phu-Thi, Minh-Triet Tran and Hai-Dang Nguyen* VietMed-VQA: A Novel Dataset and Benchmark for Vietnamese Medical Visual Question Answering	
14:10 - 14:30	*Huu Dung Nguyen, Tri Dung Do, Viet Cuong Nguyen, Oanh Thi Tran and Duc-Trong Le* PRADA-QA: Product QA with Multi-Agent Planning and Dynamic Knowledge Retrieval	*Khush Agarwal and Jonathan Hoyin Chan* Fine-Tuning Large Language Models for Automated English Speaking Proficiency Assessment Using Multimodal Linguistic and Prosodic Features	*Hue Tran, Thai Hoa Nguyen and Khanh Phuong Nguyen* The Min-makespan Vehicle Routing Problem with Drones under Multiple Trips and Visits	*Thao Thi Phuong Dao, Tan-Cong Nguyen, Nguyen Chi Thanh, Truong Hoang Viet, Trong-Le Do, Mai-Khiem Tran, Minh-Khoi Pham, Trung-Nghia Le, Minh-Triet Tran, and Thanh Dinh Le* MasHeNe: A Benchmark for Head and Neck CT Mass Segmentation using Window-Enhanced Mamba with Frequency-Domain Integration	
14:30 - 14:50	*Vu Tran and Long Nguyen* Enhancing RAFT with Knowledge Graphs for Question Answering on Vietnamese Legal Texts	*Minh Trinh The, Son Nguyen Van, Phuong Nguyen Nam and Hanh Nguyen Thi* DRONEs: Deep Reinforcement Optimization for Network k-Connectivity Restoration Enhancement in UAVs	*Van Quan La and Nguyen Hoang Phuong Tran* Grey Wolf Optimization with Entropy Control for Coverage in DSNs	*Minh Tri Ngo, Hieu Trung Dang, Hoang Trong Pham, Dinh Khoi Nguyen, Quyen Nguyen Huu and Duy Phan The* An Optimization-Driven Fusion Framework of Vision–Language Foundation Models for Large-Scale Video Retrieval	
14:50 - 15:10	*Tim Hallyburton, Ludovic Berset, Gernot A. Fink, Andreas Fischer and Anna Scius-Bertrand* Segmentation-Free Handwriting Recognition from Historical Handwritten Documents Using Large Vision-Language Models	*Toan Nguyen Khac and Ngoc Ly Quoc* XMedCLIP: A Multimodal Deep Neural Network for Bone Pathology Classification from X-ray Image	*Quoc-Trung Bui, Quang-Dung Pham, Van-Son Nguyen, and Minh Phan* Modeling and Solving the Bin Packing Problem with Relaxed Capacity Constraints: Applications in Agricultural Land Consolidation in Vietnam	*Ngoc-Thao Le, Cat-Thanh Hoang-Le and Quoc-Ngoc Ly* Text-Driven 3D Interior Scene Generation using 3D Gaussian Splatting	

	GRAND BALLROOM A	GRAND BALLROOM B	YERSIN BALLROOM A	YERSIN BALLROOM B	FOYER, 2F
15:10 - 15:30	*Kasper Lien Oftebro, Anh Nguyen Duc, Kai-Kristian Kemell and Anh Nguyen Quang* GenAI-Enabled Backlog Grooming in Agile Software Projects: An Empirical Study	*Thi Loan Bui, Thi Oanh Tran, Thị Kim Oanh Nguyễn, Chi Tho Luong and Vanha Tran* Automated ESG classification by using Natural Language Processing Techniques from Vietnamese Company Annual Reports		*Chinh Nguyen Minh, Long Tran Ngoc, Khoi Tran Man, Long Le Hoang Hien, Van Thai Hung, Duy-Dinh Le and Thanh Duc Ngo* When Events Speak: MLLM-Guided Video Retrieval with Temporal Reranking	POSTER EXHIBITION AI Applications, AI Foundations and Big Data, Applied Operations Research and Optimization, Generative AI, Human Computer Interaction and Intelligent Interactive Systems
15:30 - 16:00	Tea Break				
16:00 - 18:00	SOICT Technical Session 9 AI Applications/AI Foundations and Big Data Chair: Anna Scius-Bertrand	SOICT Technical Session 10 AI Applications Chair: Tran Cong	SOICT Technical Session 11 Applied Operations Research and Optimization Chair: Vi Chi Thanh	SOICT Technical Session 12 Networking and Communication Technologies/Software Engineering Chair: Huynh The Thien	
16:00 - 16:20	*Hai Anh Tran, Huy-Hieu Nguyen, Quoc-Trung Le, Abdelhamid Mellouk and Truong X. Tran* FedEABoost: A Client Entropy Adaptive Boosting Framework for Federated Learning	*Tan Hai Nguyen, Do Thanh Huyen Khong and Thi Quynh Hoa Nguyen* Enhancing Survey Efficiency: A Validated Vietnamese Short-Form of the MBTI Developed Through Machine Learning	*Quoc-Trung Bui, Minh Phan, Duy Vu Nguyen, Van Son Nguyen and Quang Dung Pham* Non-Parametric Feature Combination For Explainable Credit Scoring	*Thien Huynh-The, Minh-Thanh Le, Ngoc-Ha Truong, Van-Ca Phan and Truong-Thinh Le* A Lightweight and Robust Framework for Waveform Classification Using Dynamic Warping and State-Space Models	
16:20 - 16:40	*Teh-Jen Sun and Eui-Nam Huh* Entropy-Based Gradient Weighting and Batch-Size Adaptation for Virtual Data-Parallel Training	*Tien Dat Phan, Vy Anh Tran, Bao Long Hoang, Thi Minh Ngoc Truong and Thi Hau Nguyen* GRACE: A Knowledge Graph–Enhanced Conversational Recommendation System via Retrieval Augmented Generation	*Uyen Tran, Tan Tran and Canh Pham* Deterministic one-pass streaming algorithm for non-monotone DR-submodular maximization under a size constraint	*Tien H. Do, Thang V. Nguyen, Kien T. Phan, Hien T. T. Pham and Ngoc T. Dang* Channel-Aware Power and Rate Control for UOWC with DRL and HARQ Integration	
16:40 - 17:00	*Quang-Hung Bui, Bach Ngoc Pham, Anh-Minh Tran, Thanh Dat Le, The Phong Le, Tien Dung Nguyen and Anh Son Ta* AdaFRUGAL: Adaptive Memory-Efficient Training with Dynamic Control	*Yukinobu Hoshino, Namal Rathnayake, Tuan Linh Dang, and Upaka Rathnayake* Effectiveness of Rolling-Sum Preprocessing in River Mouth Water Depth Prediction Using Machine Learning	*Luan Thach, Phuong Do, Khoa Tan Vo, Thu Nguyen, Mong-Thy Nguyen-Thi and Tu-Anh Nguyen-Hoang* DESW: Reducing Concentration in Proof-of-Stake with Dynamic Exponential Stake Weighting	*Trung Vu-Thanh, Jing He, Xuan Quang Truong, Nga Phan Thi Thanh, Luong Nguyen Thi, Ninh Duong-Bao, and Khanh Nguyen-Huu,* Threshold-based AP Filtering and Distance Measure Analysis for K-means Clustering in WiFi Fingerprinting-based Indoor Localization System	
17:00 - 17:20	*Wassim Ghommidh and Mohamed Farah* Part-GNN: A partitioning-based graph neural network for efficient memory large scale data classification	*Tuan-Ngoc Nguyen, Hai-Dang Kieu, and Cam-Van Thi Nguyen* Enhance Sequential Recommendation via Linear Recurrent Units	*Bui Trong Duc, Ho Viet Duc Luong, Bui Xuan Son, Dang Quang Thang, Tran Van Tam, Nguyen Hai Dang, and Vu Van Tan* Balancing Efficiency and Fairness in the Integrated Truck–Drone Dispatching Problem with Dynamic Endurance via Pareto Front Grid Guided Multi-objective Optimization	*Nhat-Hoa Tran and Quang-Huy Vuong* A Bounded Model Checking Approach for Verifying OSEK/VDX Applications	

	GRAND BALLROOM A	GRAND BALLROOM B	YERSIN BALLROOM A	YERSIN BALLROOM B	FOYER, 2F
17:20 - 17:40	*Mark Jerome Santos, Andreas Luy, Kenneth Amurao, and Adriane Brent Castro* CITADEL: A Web-Based Faculty Performance Evaluation and Decision-Support System for Higher Education Institutions	*Dang Nhat Khuong, Nguyen Chi Kien, Truong Vinh Linh, Cao-Phan Khanh-Duy and Pham Minh Tuan* Aspect-Based Sentiment Analysis for Stock Price Movement Prediction	*Dung T.K. Ha, Anh T.N. Vu, Linh K. Duong and Phong T.D. Nguyen* Budgeted Object Detection via Online Submodular Approximation Algorithm	*Taejun Choi, Boyoon Kim and Hichan Moon* UAV-Based Target Terminal Search System for Emergency Rescue	POSTER EXHIBITION AI Applications, AI Foundations and Big Data, Applied Operations Research and Optimization, Generative AI, Human Computer Interaction and Intelligent Interactive Systems
17:40 - 18:00	*Rafael John Castro, Earl Gabriel Datu, Alex Gaebriel Limio, Julian Carlos Torno and Jenice Anne Marie Visperas* AUF iAssist: A Web-Based Helpdesk System for Efficient Support and Concern Resolution	*Abir Linoubli, Khairi Abidi and Mohamed Farah* Tokenization in Protein Language Models: Methods, Taxonomy, and Applications	COMPETITION SESSION Chair: Nguyen Hung Son		

DAY 2 - SATURDAY, 13 DECEMBER 2025

08:00 - 17:40	Registration
	GRAND BALLROOM
08:30 - 09:10	Keynote III: Josiah Poon (University of Sydney, Australia) Designing Multimodal Driven NLP for a Fluid World Chair: Cathal Gurrin
09:10 - 09:50	Keynote IV: Tung Kum Hoe Anthony (National University of Singapore, Singapore) Being Small in the Era of Large Models: Enabling Prudent AI with Lightweight, Just-in-Time AI Boxes Chair: Wray Buntine
09:50 - 10:20	Tea Break

	GRAND BALLROOM A	GRAND BALLROOM B	YERSIN BALLROOM A	YERSIN BALLROOM B	FOYER, 2F
10:20 - 12:00	SOICT Technical Session 13 Lifelog Event Retrieval Chair: Cathal Gurrin	SOICT Technical Session 14 AI Applications Chair: Nguyen Van Duy	SOICT Technical Session 15 Human Computer Interaction and Intelligent Interactive Systems Chairs: Dang Tuan Linh Le Duy Dung	SOICT Technical Session 16 Multimedia Processing Chair: Ide Ichiro	POSTER EXHIBITION Lifelog Event Retrieval
10:20 - 10:40	*Trong-Le Do, Viet-Tham Huynh, Hai-Dang Nguyen, Thuc Nguyen-Quang, Mai-Khiem Tran, Trong-Thuan Nguyen, Tu V. Ninh, Tu-Khiem Le, Thanh Duc Ngo, Duc-Tien Dang-Nguyen, Tu-Trinh Ngo, Klaus Schoffmann, Cathal Gurrin and Minh-Triet Tran* Toward Abstraction-Level Event Retrieval in Large Video Collections: Leveraging Human Knowledge and LLM-Based Reasoning in the Ho Chi Minh City AI Challenge 2025	*Ma Công Thành, Tran Duc Huy, Pham Ngoc Loc, Hoang Anh Vu and Nguyen Trong Khanh* FA-Net: A Dual-Branch Attention Architecture for Extracting Fine-Grained Anatomical Features of Wood	*Trong Tuan Do, Duc Minh Nguyen, Van Quyen Bui, and Dinh Duy Nguyen* A Co-Simulation Approach for UAV-Network-AI Interaction in Digital Twin Visual Context	*Trong-Thuan Nguyen and Minh-Triet Tran* LOGOS: Language-guided Oriented Object Detection in Aerial Scenes	

	GRAND BALLROOM A	GRAND BALLROOM B	YERSIN BALLROOM A	YERSIN BALLROOM B	FOYER, 2F
10:40 - 11:00	*Thanh Nhan Vo, Minh Doan Ngoc Binh, Pi-Dieu Sam, Hai Dang Nguyen and Khoi-Nguyen Nguyen* Real-Time Hybrid Multimodal Retrieval System for AI Challenge HCMC 2025	*Dung Tran, Huyen Tran, Hong Nguyen, Xuan-Vu Phan, and Nam-Phong Nguyen* Adaptive Rainfall Forecasting from Multiple Geographical Models Using Matrix Profile and Ensemble Learning	*Viet-Tham Huynh, Minh-Khang Nguyen, Nhut-Thanh Le-Hinh, Duy-Nam Ly, Tam V. Nguyen and Minh-Triet Tran* Fairy360VR: Immersive 360° Storytelling with Large Language Models and Generative Diffusion	*Nguyen Thanh Khoi* From Text to Thumbnail: A Unified Framework for Automated News Image Generation and Evaluation for Daily Activities	POSTER EXHIBITION Lifelog Event Retrieval
11:00 - 11:20	*Nguyen Mai Vinh, Khoi Nguyen Nam Anh, Duy Tran Khanh, Duy Do Quoc, Hung Bach Chan, Bao Tran Gia, Minh Vo, Tien Do and Thanh Duc Ngo* Towards Conversational Video Retrieval with an Intelligent Search Agent	*Nguyen Tran Minh Nhat, Pham Cong Hoang, Tran Phong Quan, Tran Le Dung, Nguyen Duy Long, Nguyen Duong Hung and Ho Viet Duc Luong* Toward a Culture-Aware Vietnamese Mental Health Support Chatbot with Large Language Models	*Phi Vu Vo Diep, Xuan Uyen Nguyen Vu, Chi Thanh Vi and Khanh-Duy Le* Enhancing VR Drink Taste Believability using Olfactory Stimulation	*Nguyen Duong Hung, Hoang Danh Quan, Ho Viet Duc Luong, Lang Hong Nguyet Anh, Nguyen Thi Thuy, and Dinh Viet Sang* Self-Supervised ViT for Endoscopy: I-JEPA Pretraining with Label-Free Diffusion Assessment	
11:20 - 11:40	*Minh-Quan Ho-Le, Duy-Khang Ho, Huy-Hoang Do-Huu, Nhut-Thanh Le Hinh, Hoa-Vien Vo-Hoang, Tu V. Ninh and Minh-Triet Tran* Applying Large Language Model (LLM) Agents for Automated Lifelog Retrieval	*Cuong Van Duc, Thai Tran Quoc, Minh Nguyen Dinh Tuan, Tam Vu Duc, Son Nguyen Van, and Hanh Nguyen Thi* MiRAGE: Misconception Detection with Retrieval-Guided Multi-Stage Reasoning and Ensemble Fusion	*Thi Quynh Hoa Nguyen, Duy Hai Nguyen, Thanh Ha Le, and Thi Duyen Ngo* An eye-tracking system for extracting and visualizing visual features of dyscalculia in children	*Hoang-Phuc Nguyen, Phuong-Linh Huynh-Ha, and Minh- Triet Tran* Generalizability Evaluation and Anchor-Guided Approach for Category-Agnostic Pose Estimation	
11:40 - 12:00	*Xuan Huy Manh, Anh Hao Kieu, Minh Hung Le, Duy Tung Nguyen, Thanh Tung Nguyen, Viet Hang Dao and Hai Vu* Estimating size of lesions in Endoscopic Images using depth model-based approaches	*An To Vinh, Nguyen Nguyen Hoang, Nguyen Luong Si, Tho Le Doan, Minh Vo, Tien Do and Thanh Duc Ngo* Leveraging Composed Image Retrieval Principles for Efficient Textual Feedback in Multimodal Retrieval	*Trung-Hau Nguyen-Tran, Dung-Minh Nguyen, My-Le Duong-Thi, Ngoc-Trinh Nguyen-Thi, Thi-Hai Vo, Hoai- Nam Do, and Khanh-Duy Le* MO-PO RM: A Collaborative Mixed Reality Board Game for Engaging Players and Audience in Learning through Playing	*Duc-Tuan Luu, Diem Nguyen, Anh-Khoa Nguyen Vu, and Vinh-Tiep Nguyen* RIOT: Robust Incremental Few-Shot Instance Segmentation via Synthetic Feature Generation with Optimal Transport	
12:00 - 13:30	LUNCH (Feast Restaurant, Level 1 - Sheraton Nha Trang Hotel & Spa)				
13:30 - 15:30	SOICT Technical Session 17 Lifelog Event Retrieval Chair: Cathal Gurrin	SOICT Technical Session 18 AI Applications Chairs: Than Quang Khoat Vu Van Khu	SOICT Technical Session 19 Recent Advances in Cyber Security Chair: Tran Hai Anh	SOICT Technical Session 20 Multimedia Processing Chair: Quan Thanh Tho	POSTER EXHIBITION Lifelog Event Retrieval, Multimedia Processing, Communication Technologies, Quantum Information, Recent Advances in Cyber Security, Software Engineering
13:30 - 13:50	*Duc-Nhuan Le, Hoang-Phuc Nguyen, Thanh-Duy Lam, Minh-Nhut Dang and Minh-Hoang Le* U-CESE: Unified Clip-based Event Search Engine for AI Challenge HCMC 2025	*Dam Thai Ninh, Nguyen Duc Kien, Trinh Ngoc Huynh, Dinh Tran Hiep, Nguyen Hai Anh, Bui Duc Manh, Miguel D'Haeseleer, Jeroen Van Schependom, Stijn Denissen, Tran Quoc Long, Nguyen Linh Trung and Guy Nagels* The Privacy–Utility Trade-off in Brain MRI Synthesis: A Comparative Framework for Generative Models	*Minh Tran Dang Quang, Tung Bui, Tran Dinh Kien Giang, Tran Quang Duc and Tuyen Ngoc Le* Robust Intrusion Detection and Classification in EVSE Using Ensemble Methods	*Quynh Vo, Dung Phan and Tho Quan* Scene Graph for Vietnamese Video Understanding: An Agentic Approach with Reasoning	

	GRAND BALLROOM A	GRAND BALLROOM B	YERSIN BALLROOM A	YERSIN BALLROOM B	FOYER, 2F
13:50 - 14:10	*Van-Loc Nguyen, Gia-Huy Vuong, Ngoc-Do Tran, Tien-Thanh Nguyen-Dang, Van-Son Ho, Van-Tu Ninh, and Minh-Triet Tran* Visionary: Optimized Temporal Video Retrieval via Large Language Model-Enhanced Query Processing	*Bao Bui-Quoc, Khang Nguyen-Vi, Hoa Nguyen-Phuong, and Nidal Kamel* Task-Aware Harmonization of Sentinel-2 for Canopy Height Mapping: A Deep Learning Application in the Ngoc Linh Mountains, Vietnam	*Hung Nguyen Tuan, Kiet Le Tuan, Nguyen Nguyen Trung, Luong Ho Nguyen, Duy Vu Ba and Hoa Nguyen Ngoc* FOAMI: Enhancing ICS Threat Detection via Feature Optimization, Realistic Augmentation, and Mutual Inference	*Quang-Linh Tran, Hoang-Bao Le, Tuong-Nghiem Diep, Binh Nguyen, Gareth J. F. Jones and Cathal Gurrin* OpenLifelogQA: An Open-Ended Multimodal Lifelog Question-Answering Dataset	POSTER EXHIBITION Lifelog Event Retrieval, Multimedia Processing, Communication Technologies, Quantum Information, Recent Advances in Cyber Security, Software Engineering
14:10 - 14:30	*Khoa Dinh Duc Anh, Duc-Tai Dinh, Trung Nguyen Le Hoang, and Nhan Nguyen Thanh* KPTER: K-Pointer for Temporal Event Retrieval	*Thanh Tam Tran, Ba Hung Ngo, Thu Thuan Pham, and Tae Jong Choi* Adaptive Multi-Level Attention for Effective Cross- Domain Brain Tumor Detection	*Vu Minh Manh, Nguyen Thanh Chung, and Cho Do Xuan* A Novel Framework for Android Malware Detection Based on Function Call Graph Pruning and Contrastive Learning	*Thanh-Son Nguyen, Van-Loc Nguyen, Cong-Luan Le, and Viet-Tham Huynh* EnAug: ENT Endoscopy Images Classification Using Ensemble and Augmentation Methods	
14:30 - 14:50	*Huu An Vu, Van Khanh Mai, Trong Tam Nguyen, Quang Duc Dam, Tien Huy Nguyen, and Thanh Huong Le* MADTempo: An Interactive System for Multi- Event Temporal Video Retrieval with Query Augmentation	*Leonhard Bürkner and Markus Westner* Critical Success Factors for AI Adoption: A Multivocal Literature Review and a Top Management Perspective	*Huynh Minh Hien, Ngo Trung Hieu, Nguyen Huu Quyen, Pham Van-Hau, Do Thi Thu Hien, and Phan The Duy* MPPO-GEM: Reinforcement Learning Approach for Generating Evasive Malware against Static and Dynamic Malware Detectors	*Minh-Khoa Le-Phan, Minh-Hoang Le, Minh-Triet Tran, and Trong-Le Do* EDGER: EDge-Guided with HEatmap Refinement for Generalizable Image Forgery Localization	
14:50 - 15:10	*Duong Nghia, Nguyen Tu, Pham Nhan, Le Truong, and Le Khoi* Althena-Vision: Adaptive Temporal Multimodal Event Retrieval with LLM-generated Multiperspective Fusion	*Thi Quynh Hoa Nguyen, Duy Hai Nguyen, Tuan Long Tran, Thi Trang Tran and Van Khoa Le* A Computational Framework for the Personalized Remediation of Reading Difficulties Using Dynamic Bayesian Networks	*Xuan Hung Truong, The Dung Luong and Anh Tu Tran* Pri-WeDec: A Private Deep Learning Approach for Weapon Detection in Digital Forensics	*Minh-Loi Nguyen, Xuan-Vu Le, Long-Bao Nguyen, Hoang-Bach Ngo and Trung-Nghia Le* Hierarchical Multi-Modal Retrieval for Knowledge-Grounded News Image Captioning	
15:10 - 15:30	*Duy Ho Khanh, Diep Tran Van, Sơn Nguyễn Hồng, Duy Nguyen, Hữu Trần Chí, and Binh Nguyen* Lucifer-TRACE: Dynamic Programming and LVLM-Aided Verification for Event-Based Video Retrieval	*Quang Nguyen and Khang Nguyen* Towards Reliable Oriented Surgical Instrument Detection: Benchmark and Evaluation	*Nhat Duy Dang, Linh Giang Nguyen, Dong Le Van, and Tuyen Ngoc Le* Few-Shot Intrusion Detection using Model-Agnostic Meta-Learning with Deep Neural Networks	*Ngoc Nguyen Tran, Duc-Duy Nguyen and Nhat-Duc Le* From Relative to Absolute: Monocular Depth Estimation in Aerial Imagery	
15:30 - 16:00	Tea Break				

	GRAND BALLROOM A	GRAND BALLROOM B	YERSIN BALLROOM A	YERSIN BALLROOM B	FOYER, 2F
16:00 - 17:20	SOICT Technical Session 21 Lifelog Event Retrieval Chair: Tran Minh Triet	SOICT Technical Session 22 AI Applications Chair: Ta Duy Hoang	SOICT Technical Session 23 Recent Advances in Cyber Security Chair: Dam Quang Tuan	SOICT Technical Session 24 Multimedia Processing Chair: Le Duc Hau	POSTER EXHIBITION Lifelog Event Retrieval, Multimedia Processing, Communication Technologies, Quantum Information, Recent Advances in Cyber Security, Software Engineering
16:00 - 16:20	*Minh Nguyen, Nga N.T. Nguyen, Cuong Dinh, Dang Nguyen, Dat Tien Nguyen and Huy M. Le* CLIPAR: Multimodal and Temporal-Aware Video Retrieval System	*Huyen Nguyen, Hieu Dam, Thuy Nguyen Thi Thu, and Viet Nguyen Kim* AuMoM: A Framework for Learning Discriminative Speaker Embeddings using a Mamba-based Mixture of Experts and Contrastive Loss	*Nghi Hoang Khoa, Vo Duc Chinh, Tu Chi Kien, Thai Hung Van and Phan The Duy* Password Generation Based on GenAI for Evaluating the Security of Password-Based Control Systems	*Thanh-Nhan Vo, Trong-Thuan Nguyen, Tam V. Nguyen and Minh-Triet Tran* SimGraph: A Unified Framework for Scene Graph-Based Image Generation and Editing	
16:20 - 16:40	*Duc-Tho Nguyen, Hieu-Hoc Tran-Minh, Khanh-Hoa Lam, Hoang-Nhut Ly, Huu-Phuc Huynh, Thanh-Tien Tran, and Trung-Nghia Le* Vortex: A Multi-Modal Fusion System for Intelligent Video Retrieval	*Anh Nguyen-Thi-Mai, Anh Nguyen-Thi-Van, Bao Doan-Quoc, Minh Tran-Duc, Tran Hung, Miroslav Voznak, Tu Dac Ho, Van Vo Nhan, Symeon Chatzinotas and Tran Dinh-Hieu* A Survey on Challenges and Emerging Frontiers of Multi-Agent Systems	*Giang Tran Dinh Kien, Duy Anh Hoang, Van Tong, Tung Bui, Tran Quang Duc and Tuyen Le Ngoc* FusionMalNet: A Hybrid Ensemble Architecture for Windows Malware Detection	*Duc-Manh Phan, Quoc-Duy Tran, Duy-Khang Do, Anh- Tuan Vo, Hai-Dang Nguyen, Trong Le Do, Mai-Khiem Tran, Vinh-Tiep Nguyen, Tam V. Nguyen, Isao Echizen, and Trung-Nghia Le* Forged Calamity: Benchmark for Cross-Domain Synthetic Disaster Detection in the Age of Diffusion	
19:00 - 21:30	GALA DINNER (Lakshmi Hall 1 - Champa Island Nha Trang - Resort Hotel & Spa)				

PROGRAM WITH ABSTRACT

DAY 1 - FRIDAY, 12 DECEMBER 2025

07:45-08:00	Registration
08:30-08:50	Conference Opening
08:50-09:30	**Session 1: Keynote I:** **Vincent Wong (The University of British Columbia, Canada)** CHAIR: *Eui-Nam Huh* LOCATION: Grand Ballroom, 2F
09:30-10:10	**Session 2: Keynote II:** **John C.S. Lui (The Chinese University of Hong Kong, Hong Kong)** CHAIR: *Yukinobu Hoshino* LOCATION: Grand Ballroom, 2F
10:10-10:40	Tea Break
10:40-12:00	**Session 3A: SOICT Technical Session I:** Quantum Information CHAIR: *Hung Son Nguyen* LOCATION: Grand Ballroom A, 2F

10:40 *Thuy Dao Thi Thu*, *Dat Le Quoc* and *Linh Hoang Dinh*

Quantum Circuit Resource Assessment for ChaCha20 Stream Cipher

ABSTRACT. The emergence of Grover's algorithm has significantly impacted the perceived security of symmetric-key cryptography in the quantum era. In response, NIST proposed three security levels for symmetric ciphers based on their resistance to quantum adversaries. This paper investigates the quantum implementation of the ChaCha stream cipher, focusing specifically on ChaCha20, which is the 20-round variant of ChaCha. We construct and simulate a quantum circuit for ChaCha20 using the ProjectQ framework, and evaluate its quantum resource requirements. Our implementation requires 1025 qubits, 64,512 CNOT gates, 21,504 Toffoli gates, and achieves a circuit depth of 511. Compared to existing designs, our circuit offers a significant depth reduction and uses only one ancillary qubit, making it more suitable for depth-constrained quantum environments. This result contributes to the broader understanding of quantum cost for stream ciphers, and provides a useful reference point for post-quantum cryptographic analysis.

11:00 *Trieu Nguyen*, *Minh Chung*, *Thanh-Dang Diep* and *Nam Thoai*

EDM4QS: An Emulator-Driven Model for Quantum Scheduling

ABSTRACT. Today, quantum processors have evolved from prototypes to real backends. Various platforms are offering quantum computing resources, including those provided by high performance computing centers and cloud providers. Essentially, quantum jobs, also known as input circuits submitted by users, can vary in terms of qubit requirements and complexity, relying on topology-aware or qubit connectivity. To manage this variability, quantum jobs go through a scheduling pipeline that maps, optimizes, and assigns tasks to the hardware under specific physical constraints. While numerous scheduling methods have been proposed, there are no proposed evaluation models or pipelines to address performance bottlenecks that occur when one phase impacts the others. Our paper presents EDM4QS – an emulator-driven model for quantum scheduling. This pipeline aims to synthesize quantum scheduling components that can evaluate various related techniques within a comprehensive framework, analyzing the interdependencies

between phases and their cumulative impact on system performance. As a long-term vision, this work facilitates holistic optimization and identifies the most essential steps in scheduling quantum jobs.

11:20 *Lan Anh Nguyen*, *Thai Anh Vo*, *Trang Nguyen*, *Son-Trung Doan*, *Toan-Duc Nguyen*, *Son-Hong Ngo* and *Eunsung Jung*

Toward Acceleration of Variational Quantum Classifier Simulation on GPUs

ABSTRACT. The Variational Quantum Classifier (VQC) is among the most widely studied models in Quantum Machine Learning (QML). However, due to the current limitations of quantum hardware, simulating QML algorithms on classical platforms such as CPUs, GPUs, or FPGAs has become a crucial step to assess performance and feasibility before a deployment on real quantum devices. In this work, we present an accelerated VQC simulation framework, referred to as A-VQC, which leverages the parallelism of classical hardwares, particularly GPUs, to achieve efficient simulation. Specifically, A-VQC introduces two complementary acceleration strategies: (1) data worker concurrency, which speeds up data transfer to GPUs by employing independent and asynchronous data-loading processes alongside VQC execution; (2) stream-wise concurrency, which exploits GPU parallel streams to train VQC on mini-batches concurrently. We implement A-VQC using a cross-platform integration of PennyLane and PyTorch. Our experiments demonstrate improvements in training speed (~10%) and GPU utilization (~30%) compared to conventional VQC simulations.

11:40 *Hieu Nguyen Doan*, *Duc Nguyen Manh*, *Thu Ngo Tran Anh*, *Tung Nguyen Xuan*, *Chien Trinh Van* and *Hwang Won Joo*

Performance Analysis of Quantum Federated Learning with Personalized Layer

ABSTRACT. Quantum computing has recently emerged as a groundbreaking field, promising unprecedented computational power and information processing capabilities. These unique advantages of quantum mechanics present a potential solution to the inherent challenges in personalized federated learning, including high communication costs due to the transmission of local updates and the limited computational capacity of classical devices. Inspired by this synergy, we propose a novel architecture named Quantum Federated Learning with Personalized Layer (QFL-PL). Our method significantly accelerates convergence, outperforming state-of-the-art approaches with 99.62% accuracy on MNIST and 86.23% on CIFAR-10.

10:40-12:00 Session 3B: SOICT Technical Session II: AI Applications

CHAIR: *Namal Rathnayake*

LOCATION: Grand Ballroom B, 2F

10:40 *Thai Anh Vo*, *Lan Anh Nguyen*, *Bao Do*, *Cong Thanh Ma*, *Son-Trung Doan* and *Son-Hong Ngo*

TriFusion: GNN-Based Multimodal Fusion for 3D Object Detection in Autonomous Driving

ABSTRACT. Reliable 3D object detection is critical for autonomous driving, yet LiDAR-only methods often fail under adverse weather, occlusion, or sensor degradation. We introduce Trifusion, a GNN-based multimodal fusion framework that integrates LiDAR, camera, and radar for robust 3D detection. Our approach builds a heterogeneous graph with nodes representing modality-specific features and edges encoding spatial and cross-modal correspondences, enabling attention based message passing across sensors. Evaluated on the nuScenes benchmark against leading baselines (e.g., PointPainting, MVX-Net, BEVFusion), Trifusion achieves superior accuracy and robustness in challenging conditions while maintaining efficiency. These results underscore the promise of graph-based fusion for reliable perception in autonomous driving.

11:00 *Nhat-Nam Duong*, *Trung-Kien Dao* and *Dinh-Van Nguyen*

A Novel Approach for Sino-Vietnamese Text Transcription by Leveraging a Pre-trained BERT and Self-Attention Mechanism

ABSTRACT. The Sino-Vietnamese (aka Hán Việt) vocabulary, derived from ancient Chinese characters (Han) but read with Vietnamese pronunciations, served as Vietnam's primary writing system until it was replaced by the modern script (chữ Quốc ngữ) in the 20th century. Today, most Vietnamese cannot read Han texts, making transcription tools crucial for preserving cultural heritage. However, the task of Sino-Vietnamese text transcription is challenging due to the presence of single-reading and multiple-reading characters, where the correct reading depends on meaning, sentence position, context, etc. Capturing contextualized information is therefore essential. This study proposes a neural network model based on a pre-trained BERT architecture, enhanced with specialized layers to capture contextual relationships in Han character sequences. Trained on expert-annotated data, the model achieved 96.08% accuracy and a 95.59% F1-score, outperforming existing approaches and providing a robust transcription solution.

11:20 *Duong Nguyen Vu Binh*, *Duy-Thanh Vu*, *Duy-Cat Can*, *Hai Dac Nguyen*, *Binh T. Nguyen*, *Oliver Y. Chén* and *Huong Ha*

A Comparison of Machine Learning Methods for Alzheimer's Disease Classification in Vietnamese Patients

ABSTRACT. Alzheimer's disease (AD) is a neurodegenerative disorder that poses an increasing burden in middle- and low-income countries. However, computational research on AD in these areas is limited, primarily due to resource constraints and small sample sizes. Traditional machine learning (ML) methods, designed for large datasets, thus may not perform optimally on smaller datasets under resource-limited settings. To evaluate which ML methods may be helpful for AD classification, given limited data, we present a modular framework for the analysis of a private Vietnamese AD dataset comprising 113 subjects. The framework incorporates a predictor module for model training and an explainer module for interpretation. We compared the classification performance, robustness, resilience, and reliability of the models. We also compared interpretable ML models with black-box models using post hoc explainability techniques. Our results indicated that the black-box XGBoost achieved the highest accuracy (81.4%), while the generalized additive model achieved a competitive performance (78.8% accuracy, 87.2% AUC). The generalized additive model also demonstrated greater robustness and resilience when compared to linear and tree-based models. Explainability analysis on inherently interpretable models and post hoc analysis on black-box models suggested the hippocampus as a common important brain region for AD classification, which aligns with previous medical findings. Overall, this study demonstrates the feasibility of using ML approaches for AD diagnosis using small datasets while balancing predictive performance and explainability.

11:40 *Binh-Dang Le*, *Quan Thi Khac*, *Duy Tran Ngoc Bao* and *Thanh Van Le*

CodeLit: A Skill-Based Framework for Automated Assessment of Code Comprehension

ABSTRACT. In programming education, verifying whether students genuinely comprehend the code they submit - especially in the age of AIgenerated solutions and peer imitation - poses a growing pedagogical challenge. This paper introduces CodeLit, a skill-based framework for automated assessment of code comprehension, bridging insights from story-based reading comprehension with Bloom's Taxonomy. CodeLit defines nine essential code comprehension skills - ranging from basic syntactic understanding to recognizing implicit logic and design abstractions—mapped across multiple cognitive levels.

Leveraging large language models (LLMs), CodeLit automatically generates targeted multiple-choice questions (MCQs) to assess these skills, applying a consistent prompt engineering strategy to both general-purpose and programming-oriented models. Our evaluations demonstrate that programming-oriented models significantly outperform general-purpose models in both completeness and quality, highlighting the value of domainadapted LLMs for code comprehension assessment. By aligning computational assessment with cognitive theory, CodeLit provides a novel pathway to reinforce academic integrity, personalize feedback, and deepen learning in programming courses

10:40-12:00 **Session 3C: SOICT Technical Session III: Software Engineering**

CHAIR: *Jonathan Chan*

LOCATION: Yersin Ballroom A, 2F

10:40 *Kaung Myat Kyaw*, *Jonathan Chan* and *Udom Silparcha*

CandleGen: Generating Synthetic OHLC Data for Different Market Trends using GANs

ABSTRACT. Financial data has been widely used in various applications, such as stock price prediction, algorithmic trading, and risk management. In algorithmic trading, for instance, traders often use historical OHLC financial data to develop and backtest trading strategies. This process allows traders to refine their approaches, understand risk, and ensure strategies are robust across different market conditions. However, relying on the historical financial data can be challenging as the historical data may not always be repeating in the future. In this project, we proposed generative adversarial network based system, CandleGen, for the synthetic OHLC data generation. Taking advantage of the strong generation ability of GAN, CandleGen creates OHLC data for the different market conditions (strong bull, bull, flat, bear, strong bear). We conduct both qualitative and quantitative experiments to evaluate the performance of CandleGen. The result shows that the generated OHLC data are aligned with the realistic data with a small margin of error.

11:00 *Thi-Ha Le*, *Quoc-Hung Pham*, *Huy-The Vu*, *Minh-Tien Nguyen* and *Xuan-Hieu Phan*

Graph-based Multi-Agents for Text-to-SQL

ABSTRACT. As a structured text generation task, Text-to-SQL translates natural language queries (NLQs) into executable SQL, enabling seamless database access. Despite advances in Large Language Models (LLMs), challenges persist with large relational databases and complex, multi-step queries requiring precise schema reasoning. We present GMA-SQL, a graph-based multi-agent framework that builds a three-layer schema graph and coordinates three LLM agents: a Graph Selector for schema pruning, a Graph CoT Decomposer for query reasoning, and a Reflexive Refiner for iterative validation. On Spider and BIRD benchmarks, GMA-SQL achieves higher execution accuracy than strong baselines, with notable gains on hard and extra-hard queries. Beyond SQL parsing, the framework supports bidirectional augmentation: graph reasoning generates synthetic NLQs from schemas for data enrichment, fostering advancements in natural language generation pipelines.

11:20 *Long Nguyen*, *Quynh Vo*, *Thi Nguyen* and *Tho Quan*

URAG 2.0: An Agentic Dual Retrieval Framework for Enhanced Reasoning in RAG-based QA Systems

ABSTRACT. Large Language Models (LLMs) have advanced Question-Answering (QA) systems but still suffer from factual errors and limited reasoning when relying solely on implicit knowledge. Retrieval-Augmented Generation (RAG) mitigates these issues by grounding responses in external corpora, yet existing pipelines often depend on a single retrieval channel, which hampers multi-hop reasoning and underutilizes heterogeneous evidence. Graph-based extensions attempt to capture structural relations but remain costly, noisy, and ultimately constrained to one stream. To

address these limitations, we propose URAG 2.0, an agentic dual-retrieval framework that extends our original URAG design. URAG 2.0 constructs two complementary indices: Frequently Asked Questions (FAQs) distilled and paraphrastically enriched from documents, and semantically chunked documents refined with context-aware rewriting. At inference, both indices are queried in parallel, and an orchestration layer fuses and ranks evidence before synthesis. Experiments across multiple QA benchmarks demonstrate that URAG 2.0 consistently outperforms advanced RAG baselines in both factual QA and multi-hop reasoning, establishing dual retrieval as a promising direction for building more accurate and explainable QA systems.

11:40 *Huynh Ngoc Khoa*, *Tang Nhat Hung*, *Dang Thien Binh* and *Nguyen Thanh Binh*

Boosting Test Smell Prediction Using Deep Learning

ABSTRACT. Test smells are indicative symptoms of poor design choices in test code, po-tentially reducing maintainability and compromising test effectiveness. While machine learning-based methods have been proposed to automate test smell detection, their predictive performance is still limited. Deep learning offers a promising solution due to its ability to earn complex con-text and patterns from data. However, its application to test smell predic-tion, particularly with sequence data extracted from test code, remains un-derexplored. To address these motivations, this study aims to present a deep learning-based approach for test smell prediction using input data in the form of sequences. The proposed method is experimentally evaluated on two popular test smells: Eager Test and Mystery Guest. The performance of all proposed models demonstrated significant improvement over baseline models, with the highest F1-score increase of approximately 24%. A com-parative evaluation of three deep learning models, including Convolutional Neural Network, Bidirectional Long Short-Term Memory, and Gated Recur-rent Unit, reveals that Bidirectional Long Short-Term Memory achieved the highest F1-score of 0.7475 for Eager Test, while Convolutional Neural Network performed best on Mystery Guest with F1-score of 0.6529. This work is considered the first effective application of deep learning for pre-dicting test smell on sequence data, highlighting the promising approach in the area.

10:40-12:00 Session 3D: SOICT Technical Session IV: Networking and Communication Technologies

CHAIR: *Abdelhamid Mellouk*

LOCATION: Yersin Ballroom B, 2F

10:40 *Long Khanh Tran*, *Thiloan Bui*, *Viet Hieu Ha*, *Van Dat Nguyen*, *Dinh Dung Nguyen* and *Thi Ngoc Dung Kieu*

Propagated Presence: A Bluetooth Propagation-Based Method for Automated Classroom Attendance on Mobile Devices

ABSTRACT. Intoday'sfast-moving,technology-drivenworld,manualrou- tines are steadily giving way to automated workows. Yet in many uni- versities, attendance systems remain outdated: students check in once at the start, with no location verication, no continuous tracking, and little defense against cheating. These gaps waste class time, invite er- rors, and make proxy check-ins easy. To tackle these shortcomings, we propose a solution that is fast, accurate, scalable, and cost-eective: an Android application that turns every student's smartphone into a Blue- tooth Low Energy (BLE) beacon. Instead of installing dedicated hard- ware, our approach reuses the radio already built into modern phones. Each device periodically broadcasts a unique identier linked to the user, along with session-specic data. At the same time, the lecturer's mobile device broadcasts its own beacon. After each round, every phone uploads its scan log to the server, where a propagation algorithm determines the set of students actually in attendance. Using a round-based cyclescan, upload, repeatprevents data gaps and continuously conrms that stu- dents remain in the room for the entire session. We also embed real-time validity checks to block impersonation

attempts. Because no extra hard- ware is needed, campus-wide deployment is inexpensive, and the algo- rithm's design minimizes calibration headaches even when phone models dier. Our experiments show that the method is reliable, highly accu- rate, and ready to scale. In future work, we plan to adapt the system to other settings, such as workplaces, and to develop a companion iOS app.

11:00 *Quang Huy Duong* and *Brigitte Jaumard*

An Evaluation on Defragmentation with CDC ROADMs in Elastic Optical Networks

ABSTRACT. The anticipated widespread adoption of Colorless, Directionless, and Contentionless (CDC) ROADMs in Elastic Optical Networks (EONs) is driven by their ability to eliminate spectrum contention. However, a significant gap exists in the literature regarding a quantifiable evaluation of their benefits, particularly in the context of network defragmentation. To address this, we present a new exact solution and two efficient heuristic algorithms for the make-before-break defragmentation problem in EONs with CDC ROADMs. Our exact algorithm marks a significant advance over existing solutions, and our heuristics enable us to perform a large-scale evaluation on a realistic 24-node, 86-link network. The results from our experiments show that CDC ROADMs provide a 2\% improvement in the blocking rate, offering crucial data for their cost-benefit analysis and strategic deployment.

11:20 *Thanh-Dat Tran*, *Son Truong*, *Hoc Phan*, *Hai-Trang Dang* and *Thien Huynh-The*

Fusing Gated Spatial-Channel Units and Fractal Cross-Scale Attention for Lightweight Waveform Classification

ABSTRACT. Accurately classifying radar-communication signals remains a fundamental yet challenging task, as these waveforms often exhibit high variability and are easily distorted by noise, channel fading, and spectrum overlap. Conventional deep learning approaches, despite their remarkable progress, tend to capture either local spatial cues or long-range dependencies in isolation, thereby limiting their ability to achieve robust recognition in complex environments. To address these challenges, we propose SynerNet, a lightweight deep neural architecture that effectively integrates spatial diversity and cross-scale attention mechanisms. Specifically, SynerNet leverages the smoothed pseudo Wigner-Ville distribution to generate informative time-frequency representations, which retain high resolution while mitigating undesired cross-term interference. Building upon these representations, the network is enhanced by two key components: the Gated Spatial-Channel Unit module, which jointly models spatial and channel dependencies to selectively emphasize salient features while suppressing noise, and the Fractal Cross-Scale Attention module, which employs a hierarchical fractal-inspired attention scheme to preserve fine-grained details across multiple scales while ensuring global consistency. Simulation results on 12 waveform classes encompassing both radar and communication signals demonstrate that SynerNet achieves an average classification accuracy of 90.61%, with only 47K parameters and an inference latency of 0.552 ms, outperforming existing deep learning approaches. These results highlight the strong potential of SynerNet for real-world deployment in intelligent sensing and wireless communication systems under resource-constrained environments.

11:40 *Minh-Cat-Tuong Nguyen*, *Si-Thai Trang* and *Chi-Thanh Vi*

A mobile-based attendance system using Bluetooth MAC address scanning

ABSTRACT. The development of automated attendance systems is a mature area of research, with a wide body of literature exploring various technologies to overcome the limitations of manual methods. Traditional paper-based roll calls or sign-in sheets are widely considered inefficient, consuming valuable instructional time, and are highly susceptible to inaccuracies and academic dishonesty, such as proxy attendance. In response, various technological paradigms have been proposed and implemented, each presenting a unique set of trade-offs between security, cost,

usability, and precision. This paper presents the design, implementation, and evaluation of a mobile-based attendance management system for academic institutions. The system leverages the classic Bluetooth device discovery process, capitalizing on the observation that modern smartphones broadcast their static, non-randomized MAC address when placed in a user-initiated discoverable mode. It leverages a robust Android application and a cloud backend to provide an automated, contactless attendance verification mechanism that is both efficient and scalable.

10:40-12:00 **Session 3E: Poster Exhibition**

CHAIRS: *Hai Anh Tran*, *Hoang Ta* and *Hung Son Nguyen*

Quoc-Viet Hoang and *Trung-Hieu Le*

Feature Optimization for Improving Locust Detection

ABSTRACT. Locusts are a major contributor to economic losses in global agriculture. Early detection of these insects enables growers to implement effective control strategies, thereby enhancing crop yield and quality. However, current object detection models have achieved suboptimal results in locust detection, primarily due to they are often small size, occlusion, and camouflage. In this work, we propose an effective approach to address this challenge, dubbed FO-YOLO. A Feature Optimization module is developed and integrated into the head of YOLOv10 to improve feature extraction for detection. The proposed module concurrently receives features from multiple levels and establishes a shorter pathway between low-level and high-level features through feature fusion. In addition, we add an extra prediction layer to enhance the detection performance for small locusts. Experimental results verify that our FO-YOLO achieves state-of-the-art performance, surpassing competitive models.

Chinh Tran Duc, *Nghia Nguyen Tri*, *Ha Chu Hai* and *Hanh Nguyen Thi*

HMCT: A Hybrid Multi-Scale CNNs- Transformer encoder for Fault Diagnosis in WSNs

ABSTRACT. Fault diagnosis in wireless sensor networks (WSNs) is a critical task to ensure the reliability and precision of the collected data. Sensor nodes are often deployed in harsh and unattended environments, making them prone to faults such as bias shifts, drifts, spikes, erratic fluctuations, and stuck failures. If not detected, these faults can reduce network performance and cause wrong decisions. This paper proposes a hybrid deep learning framework that integrates multiscale convolutional neural networks (CNNs) with a Transformer-based attention mechanism to extract temporal and spatial features from collected sensor data in order to maximize the ability to diagnosis faults. To validate our approach, we construct various experiments on a realistic dataset that includes temperature, humidity, and surface pressure measurements. The experimental results show that our proposed framework reaches strong performance in fault classification, significantly outperforming conventional machine learning baselines. Compared to the strongest baseline, the proposed approach achieves an accuracy improvement of more than X%.

Dang Phuong Nam, *Nguyen Kieu Linh* and *Pham Thanh Hieu*

SafeGen: Embedding Ethical Safeguards in Text-to-Image Generation

ABSTRACT. Generative Artificial Intelligence (AI) has created unprecedented opportunities for creative expression, education, and research. Text-to-image systems such as DALL·E, Stable Diffusion, and Midjourney can now convert ideas into visuals within seconds, but they also present a dual-use dilemma, raising critical ethical concerns: amplifying societal biases, producing high-fidelity disinformation, and violating intellectual property. This paper introduces SafeGen, a framework that embeds ethical safeguards directly into the text-to-image generation pipeline, grounding its design in established principles for Trustworthy AI. SafeGen integrates two complementary components: BGE-M3, a fine-tuned text classifier that filters

harmful or misleading prompts, and Hyper-SD, an optimized diffusion model that produces highfidelity, semantically aligned images. Built on a curated multilingual (English-Vietnamese) dataset and a fairness-aware training process, SafeGen demonstrates that creative freedom and ethical responsibility can be reconciled within a single workflow. Quantitative evaluations confirm its effectiveness, with Hyper-SD achieving IS = 3.52, FID = 22.08, and SSIM = 0.79, while BGE-M3 reaches an F1-Score of 0.81. An ablation study further validates the importance of domain-specific fine-tuning for both modules. Case studies illustrate SafeGen's practical impact in blocking unsafe prompts, generating inclusive teaching materials, and reinforcing academic integrity.

Ngoc-Bich Nguyen Thi, *Tam Vo Minh*, *Phuong Thao Nguyen Ho*, *Khoa Nguyen* and *Vinh-Tiep Nguyen*

Evaluating Syllabus via Sub-Criteria: A Comparative Study of LLM and Experts

ABSTRACT. Syllabus design and evaluation are critical tasks for high school teachers but remain time-consuming, resource-intensive, and difficult to scale in Vietnam's education reforms. Although the Ministry of Education and Training (MOET) provides official criteria, these standards are often vague, leading to inconsistent application. Large language models (LLMs) offer efficiency and scalability for educational tasks, yet their role in aligning with national evaluation criteria remains underexplored. In this study, we investigate the use of LLMs to automatically decompose MOET's criteria into fine-grained sub-criteria and compare them with human-expert decompositions. We then evaluate the same syllabus against both sets of sub-criteria to assess consistency, reliability, and scalability. Our findings reveal that LLMs can effectively transform broad standards into actionable components, reducing workload and enabling replication, while also presenting limitations in accuracy and contextual alignment. This study contributes to the growing discourse on AI in education by highlighting how LLMs can complement human expertise in curriculum evaluation within the Vietnamese context.

Huong Chau, *Ly Thi Le*, *Phuc Chanh Lam*, *Anh Thi-Hoang Nguyen*, *Trong-Hop Do* and *Huy M. Le*

ViTrustKOL: A Vietnamese Dataset for Consumer Trust Classification toward Key Opinion Leaders

ABSTRACT. We present a novel annotated Vietnamese dataset for the study of consumer trust toward Key Opinion Leaders (KOLs) on the TikTok platform. The corpus comprises 16,000 user comments manually labeled into three trust categories—Positive, Neutral, and Negative—reflecting authentic consumer–KOL interactions and a wide spectrum of linguistic expressions of trust and distrust. To demonstrate the dataset's utility, we benchmark two modeling paradigms: a Vietnamese-specific pre-trained encoder (PhoBERT) and a large language model (LLM)-based pipeline. Experimental results indicate that the LLM-based approach substantially outperforms PhoBERT, achieving 69.2% accuracy and a 68.5% macro F1-score versus PhoBERT's 60.7% accuracy and 59.3% macro F1-score. The primary contributions of this work are threefold: (1) the introduction of a large, manually curated Vietnamese dataset tailored for trust classification in social media, (2) a systematic benchmarking of both language-specific and LLM-based methods on this resource, and (3) empirical evidence that LLM-based pipelines can provide notable performance gains for trust analysis in Vietnamese. This dataset and the accompanying benchmarks establish a foundation for future research on context-aware trust modeling and practical applications in Vietnamese natural language processing.

Tung Luong Nguyen, *Hong Nhat Tran* and *Van Hai Do*

Efficient Caching for Conditional Flow Matching in Vietnamese Zero-Shot TTS

ABSTRACT. Zero-shot text-to-speech (ZS-TTS) has advanced rapidly in high-resource

languages, but Vietnamese remains challenging due to its complex phonology and the mismatch between orthography and pronunciation. We investigate Conditional Flow Matching (CFM) for Vietnamese ZS-TTS and find that naive multilingual fine-tuning fails to close the quality gap. To address this, we propose two components: a phoneme-based input representation that better aligns linguistic and acoustic units for Vietnamese, and a cache-based sampler that reuses intermediate computations to reduce inference time without retraining. Implemented on F5-TTS, our system achieves strong perceptual quality and speaker similarity on Vietnamese (MOS 4.42, SIM-o 0.8093) with competitive intelligibility, and generalizes well to cross-lingual synthesis (MOS 3.84, WER 2.94\%). Ablation results reveal a clear balance between quality and efficiency: moderate caching retains most perceptual quality while significantly improving synthesis speed. These findings demonstrate that phoneme-level modeling and caching together offer a simple and effective path toward high-quality, efficient CFM-based Vietnamese ZS-TTS.

Truong Xuan Hung, *Luong The Dung*, *Tran Anh Tu*, *Vo Dinh Quyet* and *Nguyen Anh Khoi*

A Robust Multi-Modal Framework for Explicit Content Detection in Digital Forensics via Adversarial-Resilient Ensemble Learning and Homomorphic Encryption

ABSTRACT. The rapid expansion of digital media, driven by generative AI developments as of 2025, has posed formidable challenges in digital forensics, especially for identifying explicit materials such as child sexual abuse material (CSAM). Conventional detection systems, often based on skin-tone heuristics or basic convolutional neural networks (CNNs), remain susceptible to adversarial distortions and synthetic deepfakes, resulting in elevated false-positive rates in practical settings. In this study, we propose an integrated multi-modal architecture that combines CNNs with adversarial hardening, feature extraction in alternative color spaces (YCbCr, HSV), and ensemble methods (SVM and RNN) to deliver enhanced durability and precision. To mitigate privacy risks associated with sensitive CSAM handling, we incorporate the Cheon-KimKim-Song (CKKS) homomorphic encryption scheme, facilitating operations on ciphertexts without revealing plaintext, thereby upholding investigator confidentiality and forensic integrity. Tested on benchmarks including NPDI, UTKFace, and bespoke adversarial datasets, our approach yields 98.5% accuracy, reducing false positives by 15-20% relative to references like NudeNet and DeepPornDetection, while sustaining efficacy in encrypted computations with negligible performance penalties. Key innovations encompass a specialized adversarial perturbation generator for forensic contexts, CKKS-enabled secure ensemble inference, and compatibility with platforms such as Autopsy. This research fills notable voids in the 2025 scholarly landscape, where emphases on deepfake resilience and privacy-preserving machine learning (PPML) prevail, yet the synergy of explicit-content-targeted adversarial defenses with homomorphic encryption is largely uncharted, presenting a methodologically fresh strategy to bolster forensic inquiries while prioritizing data security.

Thi-Huong Nguyen, *Anh-Loi Nguyen*, *Doan-Tung Duong* and *Van-Toi Nguyen*

A multimodal framework for Vietnamese Sign Language Recognition

ABSTRACT. In this paper, we propose a multimodal framework that integrates multiple input modalities, namely RGB video, optical flow, and keypoint information, after each modality is processed through deep learning architectures. This design is based on the idea that various input sources complement features of sign language: RGB frames offer appearance cues, optical flow encodes motion dynamics, and keypoints highlight skeletal structures. By performing a late fusion, our method leverages the strengths of each modality. We evaluated the proposed framework on the ViSL120 dataset of isolated Vietnamese Sign Language (ViSL) and performed

systematic comparisons with single-modal and other fusions. The results demonstrate that our multimodal approach significantly outperforms the recognition accuracy of the baseline models.

Luyen Vo, *Phuoc-Sang Tran* and *Anh-Cuong Le*

Addressing Data Scarcity and Imbalance in Depression Screening with Persona-Driven Synthetic Data

ABSTRACT. Data scarcity and privacy concerns are significant barriers to developing machine learning models for depression screening. This research introduces and validates a novel framework for generating high-quality synthetic data to address this challenge. The core of the methodology is a three-stage pipeline where a Large Language Model (LLM), guided by clinical scales (PHQ-8) and diverse user personas, generates realistic narrative synopses from clinical interviews. Empirical validation demonstrates the framework's efficacy: a model trained exclusively on our synthetic data achieved a high F1-score of 0.84 on the DAIC test set, indicating a more stable and reliable training process compared to prior methods while approaching the performance of models trained on real data. Further analysis, including embedding space visualization, confirms that our synthetic data closely mirrors the semantic features and label distribution of real-world data, thereby mitigating the risks of label imbalance and enhancing model generalization. This study contributes (1) a validated framework for synthetic data generation and (2) a high-quality dataset released to the research community. Our work lays the groundwork for developing more robust and equitable AI-driven tools for mental health screening, underscoring that the quality of synthetic data is paramount for building reliable applications.

Quoc-Viet Hoang, *Trung-Hieu Le*, *Hong-Son Vu* and *Van-Quyet Nguyen*

Fish-Net: an Effective Model for Underwater Fish Detection

ABSTRACT. The automated detection of fish is growing in demand for different applications, such as aquaculture monitoring and oceanographic research. Nevertheless, the performance of existing object detectors is often limited by changing illumination and low-light conditions underwater. In this study, we propose an effective model to improve the accuracy of fish detection, dubbed Fish-Net. Our approach—a feature-based learning method is developed by attaching a proposed feature improvement (FI) module before the backbone section of YOLOv10. The FI module is responsible for enhancing low-light images and generating sharp features of fish to provide to the baseline detector. Comprehensive experimental results on publicly available datasets validate that our approach outperforms existing object detection models.

Cuong Nguyen, *Dung Tran*, *Hong Nguyen*, *Xuan-Vu Phan* and *Nam-Phong Nguyen*

VRAE: Vertical Residual Autoencoder for License Plate Denoising and Deblurring

ABSTRACT. In real-world traffic surveillance, vehicle images captured under adverse weather, poor lighting, or high-speed motion often suffer from severe noise and blur. Such degradations significantly reduce the accuracy of license plate recognition systems, especially when the plate occupies only a small region within the full vehicle image. Restoring these degraded images in a fast real-time manner is thus a crucial pre-processing step to enhance recognition performance. In this work, we propose a Vertical Residual Autoencoder (VRAE) architecture designed for the image enhancement task in traffic surveillance. The method incorporates an enhancement strategy that employs an auxiliary block, which injects input-aware features at each encoding stage to guide the representation learning process, enabling better general information preservation throughout the network compared to conventional autoencoders. Experiments on a vehicle image dataset with visible license plates demonstrate that our method consistently outperforms Autoencoder (AE), Generative Adversarial Network (GAN), and Flow-Based (FB) approaches. Compared with AE at

the same depth, it improves PSNR by about 20%, reduces NMSE by around 50%, and enhances SSIM by 1%, while requiring only a marginal increase of roughly 1% in parameters.

Phuong Thao Vu, *Thi Hanh Nguyen*, *Hai Long Nguyen* and *Hai Chau Le*

A Hybrid Quantum-Classical Machine Learning Framework for Robust Sepsis Detection Utilizing Immune Gene Signatures

ABSTRACT. Sepsis remains a critical medical condition demanding rapid and accurate diagnosis to improve patient outcomes. While gene expression data offers a promising avenue for precise diagnostics, its high dimensionality and inherent platform-specific bias pose significant challenges for conventional machine learning models. This paper introduces a novel hybrid quantum-classical machine learning framework for robust, cross-platform sepsis detection. Our methodology leverages a carefully curated set of public gene expression datasets (Affymetrix U133, AgilentV2, AffyU219) partitioned into distinct training and testing cohorts to ensure generalizability. We employ a rigorous feature engineering pipeline focused on immune-related genes, involving differential expression analysis and Random Forest-based selection, to identify a minimal, high-impact gene signature. The core of our approach is the development of a Quantum Support Vector Machine (QSVM) model, where gene expression data is encoded into a quantum feature space using a parameterized quantum circuit, enabling the calculation of a complex, high-dimensional kernel. We benchmark our developed QSVM-based method against a state-of-the-art XGBoost model and a Classical SVM. Results demonstrate that our proposed model consistently outperforms these classical counterparts in multiple independent test sets, achieving superior accuracy (up to 99.42%), sensitivity (up to 99.79%), and F1-score (up to 99.69%).

Hong-Viet Tran, *Xuan-Son Quan*, *Tien-Khoi Nguyen* and *Lam-Quan Tran*

ViFin-MARS: A Question-Answering System for Financial News Dataset integrating User Intent Identification and Multi-Agent RAG Systems

ABSTRACT. Large Language Models (LLMs) have significantly advanced Natural Language Processing, demonstrating remarkable capabilities in general tasks such as text generation, summarization, and question answering. However, when generating authentic questions in the rapidly changing financial news domain, they are prone to producing hallucinated responses. Additionally, retraining LLMs on new data has many considerations such as computational cost and data quality. This paper introduces ViFin-MARS, a question-answering system. Our approach uniquely integrates User Intent Identification with a Multi-Agent Retrieval Augmented Generation (RAG) framework. The system first classifies user queries into specific financial intents. This classification then directs the query to a specialized agent within the multi-agent system, which is optimized to retrieve relevant context from our financial news dataset and generate a answer. Experimental results on our evaluation dataset show that ViFin-MARS achieves 76% accuracy, demonstrating the effectiveness of this integrated architecture in enhancing the reliability and precision of financial question-answering.

Van Nang Hung Nguyen, *Nam Van Hoang*, *The Thang Cao* and *Truc Thi Kim Nguyen*

Polynomial-Augmented Instant Neural Graphics Primitives

ABSTRACT. Neural Radiance Fields (NeRFs) provide high-quality view synthesis and 3D reconstruction. However, training is still expensive. Instant Neural Graphics Primitives (Instant-NGP) reduce this cost by combining hash encoding with shallow MLPs. This design enables real-time training but keeps the dense layers linear, which limits expressiveness and slows convergence. We propose MPLP (Multi-degree Polynomial Layer Perceptrons), a lightweight extension of Instant-NGP. Each layer augments its input with second-degree polynomial terms before projection and

normalization. This expansion improves non-linear modeling without increasing depth and with only a minor parameter increase.. On the NeRF-Synthetic Lego dataset, MPLP-NeRF improves PSNR from 24.34 dB to 24.93 dB (+0.59 dB). Reconstruction loss drops by 12%. The model also converges 40% faster. Training overhead is modest: only +15% runtime and negligible memory use. These results show that MPLP is a practical and scalable upgrade to Instant-NGP. It balances speed and quality, making polynomial-enhanced NeRFs a strong candidate for real-time view synthesis and 3D reconstruction.

Hong Nguyen Thi, *Viet Anh Nguyen* and *Kien Do Trung*

Improve the Effectiveness of Predicting Student Learning Outcomes using a MoE Networks with LSTM Routing

ABSTRACT. In online or blended courses, learners must self-manage their learning progress. Predicting learning outcomes is therefore essential to provide timely warnings to learners and help lecturers adjust teaching plans to enhance training quality. Previous studies have shown that LSTM models using time-series data are effective. Still, they often rely on multi-layer deep networks, which increase computational costs and make it difficult to model the behaviors of different learner groups. This study proposes a hybrid expert network architecture to predict whether learners will pass or fail. Each expert comprises an LSTM layer combined with an Attention mechanism, while a routing mechanism—implemented as an LSTM network—selects the appropriate expert based on time-series data. The number of experts is determined by clustering learners using the Fuzzy C-Means. Experiments on the OULAD dataset demonstrate that the proposed architecture outperforms traditional stacked multi-layer LSTM models.

Thi-Thuy-Duong Trinh, *Ngoc-Hai Truong*, *Trong-Huy Nguyen* and *Thanh-Huong Le*

Contrastive Preference Optimization for Low-Resource Vietnamese to Khmer Neural Machine Translation

ABSTRACT. With the rapid development of multilingual large language models (LLMs), machine translation has made significant progress. However, translating low-resource languages remains challenging because LLMs are often trained on limited data for these languages. In addition, the available bilingual datasets for such language pairs are typically small and may contain noise, leading to suboptimal translation quality. In this paper, we address this problem in the context of Vietnamese to Khmer machine translation. We propose a three-stage training pipeline that effectively leverages both monolingual and bilingual data while keeping computational costs low. First, we continually pre-train the LLM on Vietnamese and Khmer monolingual data to improve its language understanding. Next, we apply parameter-efficient fine-tuning using LoRA to align the model with bilingual data. Finally, we train the model using the Contrastive Preference Optimization (CPO) method, which helps it better distinguish between high- and low-quality translations, thereby producing more accurate and natural outputs. Experimental results show that our approach outperforms existing models and few-shot prompting with GPT-OSS-120b across both traditional metrics (BLEU, METEOR) and semantic metrics (COMET, KIWI-COMET).

12:00-13:30	Lunch LOCATION: Feast Restaurant, 1F
13:30-15:30	**Session 4A - SOICT Technical Session V: Generative AI** CHAIR: *Jian Gang Ngui* LOCATION: Grand Ballroom A, 2F
13:30	*Jian Gang Ngui* **SEA-LION: Southeast Asian Languages in One Network**

13:50 *Huyen Nguyen*, *Hieu Dam*, *Cong Tran* and *Cuong Pham*

AD-GENESIS: Anomaly Detection through Gradient-Guided Generative Synthesis

ABSTRACT. The generalization capacity of anomaly detection models when confronted with previously unseen anomalies continues to pose a substantial challenge, especially within mission-critical systems. To tackle this fundamental problem, we present AD-GENESIS - an innovative framework that leverages generative artificial intelligence in conjunction with gradient-based optimization for training anomaly detection models. The framework performs direct optimization of latent variables within the generative model to synthesize diverse and novel anomalous samples that can effectively evade existing detection models. These synthetically generated samples are subsequently employed to enhance the detection model, thereby improving its capability to process complex anomalous patterns. Additionally, AD-GENESIS incorporates a state-of-the-art Mamba-based architecture for the detection model. This choice leverages Mamba's linear-time complexity and ability to capture long-range dependencies, ensuring high accuracy and computational efficiency during inference. Experimental evaluation conducted on the ADBench benchmark reveals that AD-GENESIS achieves superior performance compared to contemporary state-of-the-art models while maintaining minimal inference latency.

14:10 *Huu Dung Nguyen*, *Tri Dung Do*, *Viet Cuong Nguyen*, *Oanh Thi Tran* and *Duc-Trong Le*

PRADA-QA: Product QA with Multi-Agent Planning and Dynamic Knowledge Retrieval

ABSTRACT. Large Language Model (LLM)-based autonomous agents have demonstrated strong capabilities in decision-making and handling complex tasks. However, there remains a notable gap in public research on leveraging multi-agent systems for Product Question Answering (PQA), a crucial area in modern e-commerce. In this work, we introduce PRADA-QA, a framework that enhances the user experience through multi-agent collaboration, enabling dynamic information retrieval from diverse sources to respond to user queries accurately. In addition, we propose a planning module that adaptively guides the agents' objectives, improving task fulfillment efficiency while minimizing redundant steps and operational costs. For evaluation, we employ a reward model--an indispensable component in reinforcement learning-based LLM post-training--as a proxy for human preferences. This approach was designed to capture user-centric quality and may also be generalizable to other open-ended QA scenarios. Leveraging a reward model-based evaluation strategy, we conduct extensive experiments across three distinct domains to assess the effectiveness of PRADA-QA. The experimental results demonstrate its superiority compared to traditional approaches, highlighting its enhanced ability to generate accurate and contextually appropriate responses for the product question-answering task.

14:30 *Vu Tran* and *Long Nguyen*

Enhancing RAFT with Knowledge Graphs for Question Answering on Vietnamese Legal Texts

ABSTRACT. Vietnamese Legal Question Answering (Legal QA) is an emerging field with the potential to enhance access to legal information, yet it faces challenges such as limited datasets, insufficient reasoning capabilities, and a lack of unified benchmarks. This study addresses these gaps by constructing an expert-verified dataset in two key domains—Labor Law and Enterprise Law—and by advancing Retrieval-Augmented Generation (RAG) methods for legal QA. We implement multiple approaches, including Naive RAG, RAG with query expansion, RAFT with a fine-tuned reranker, and RAFT with knowledge graph integration (RAFT-KG). Experimental evaluations compare open-source models (LLaMA-based) with GPT-4.1 using automatic LLM-based metrics (RAGAs) across dimensions of context recall,

faithfulness, answer relevancy, and factual correctness. Results show that while GPT-4.1 achieved the highest factual correctness, RAFT-KG and query expansion methods improved context recall and faithfulness, and the RAFT reranker provided balanced, domain-adapted performance. These findings demonstrate that integrating knowledge graphs and domain-tuned rerankers enables opensource models to approach proprietary LLM performance, paving the way for transparent, robust, and scalable legal QA systems tailored to the Vietnamese context

14:50 *Tim Hallyburton*, *Ludovic Berset*, *Gernot A. Fink*, *Andreas Fischer* and *Anna Scius-Bertrand*

Segmentation-Free Handwriting Recognition from Historical Handwritten Documents Using Large Vision-Language Models

ABSTRACT. Handwriting recognition and information extraction are valuable tasks for preserving cultural heritage. Traditional deep learning approaches depend on two principal success factors, which are often difficult to ensure. First, an explicit segmentation of scanned pages into text lines to facilitate handwriting recognition. Secondly, a large amount of training samples for fine-tuning to specific manuscripts. In contrast, pre- trained Large Vision-Language Models (LVLMs) do not depend on these factors, as they are able to extract handwritten information from entire pages by means of a prompt. In this paper, we contribute an experimental benchmark that aims to assess to what degree explicit segmentation and/or fine-tuning still are necessary for LVLMs. Our findings on the Washington, IAM, and CM1 datasets indicate that both traditional success factors are becoming less important for the best performing LVLMs and optimized prompts.

15:10 *Kasper Lien Oftebro*, *Anh Nguyen Duc*, *Kai-Kristian Kemell* and *Anh Nguyen Quang*

GenAI-Enabled Backlog Grooming in Agile Software Projects: An Empirical Study

ABSTRACT. Effective backlog management is critical for ensuring that development teams remain aligned with evolving requirements and stakeholder expectations. However, as product backlogs consistently grow in scale and complexity, they tend to become cluttered with redundant, outdated, or poorly defined tasks, complicating prioritization and decision-making processes. This study investigates whether a generative- AI (GenAI) assistant can automate backlog grooming in Agile software projects without sacrificing accuracy or transparency. Through Design Science cicles, we developed a Jira plug-in that embeds backlog issues with the vector database, detects duplicates via cosine similarity, and leverage GPT-4o model to propose merges, deletions, or new issues. We found that AI-assisted backlog grooming achieved 100% precision while reducing the time-to-completion by 45%. The findings demonstrated the tool's potential to streamline backlog refinement processes while improving user experiences

13:30-15:30 Session 4B: SOICT Technical Session VI: AI Applications

CHAIR: *Guy Nagels*

LOCATION: Grand Ballroom B, 2F

13:30 *Vu Pham* and *Long Nguyen*

Optimization Approaches for Language Models in the Task of Translating Sino-Vietnamese Texts into Modern Vietnamese

ABSTRACT. This study proposes a method for developing compact language models in the task of translating Sino-Vietnamese texts into Modern Vietnamese. In the field of language modeling, developing models from scratch and deploying them is costly due to the large amount of data and computational resources required. To address this challenge, we introduce the KD-QLoRA method, which combines Knowledge Distillation (KD) with Quantized Low-Rank Adapter (QLoRA). This approach inherits and extends existing techniques by integrating multiple optimization methods to improve the development and deployment of language models. Experiments

conducted on open-source pretrained models such as LLaMA 3, Qwen 3, and Phi 3 demonstrate that KD-QLoRA enables the development of smaller language models that outperform QLoRA. Moreover, Qwen 3 1.7B, when fine-tuned with KD-QLoRA, achieves performance comparable to or better than larger models in the task of translating Sino-Vietnamese texts into Modern Vietnamese.

13:50 *Nam Nguyen Tu* and *Hiroki Takahashi*

Motion-Gated Adaptive Filtering for Continuous Sign Language Recognition

ABSTRACT. Continuous Sign Language Recognition (CSLR) is challenged by the complex spatio-temporal dynamics inherent in sign language videos. Existing methods often rely on uniform processing strategies, computationally expensive external cues like optical flow, or struggle with undertrained feature extractors. To address these limitations, we propose Motion-Gated Adaptive Spatio-Temporal Filtering (MG-ASTF), a novel plug-and-play module for deep CSLR networks. Crucially, MG-ASTF computes motion estimations directly from intermediate feature maps, eliminating the need for external data. It uses these internal motion cues to dynamically gate two parallel, specialized filtering pathways: one prioritizing temporal dynamics for high-motion segments and another emphasizing spatial detail for static or low-motion regions. We integrate MG-ASTF into a standard ResNet-based architecture and demonstrate its efficacy on the PHOENIX14 and PHOENIX14-T benchmarks. Our RGB-only approach achieves highly competitive results, notably matching the performance of complex multi-modal systems, thereby showcasing a more efficient path to robust feature learning in CSLR.

14:10 *Khush Agarwal* and *Jonathan Hoyin Chan*

Fine-Tuning Large Language Models for Automated English Speaking Proficiency Assessment Using Multimodal Linguistic and Prosodic Features

ABSTRACT. Automated Spoken Language Assessment (ASLA) presents a scalable solution for evaluating English as a Second Language (ESL) learners, yet requires robust and accurate systems. This paper proposes a novel approach by fine-tuning several Large Language Model (LLM) variants from the Qwen family on a rich, multimodal dataset incorporating both prosodic and linguistic features. Our systematic evaluation demonstrates that a fine-tuned Qwen 2.5 7b model achieves the best overall performance in predicting CEFR proficiency levels, outperforming even larger, newer models. The results confirm that fine-tuning is essential for predictive accuracy and that both prosodic and linguistic features provide complementary information that enhances performance. Furthermore, we show that post-processing with isotonic regression substantially improves score calibration. However, a detailed per-score analysis reveals a primary limitation: the models exhibit a systematic bias, overestimating low-proficiency speakers and underestimating high-proficiency speakers, largely due to data imbalance. While performance is strong in the mid-range proficiency levels, this study highlights the critical challenge of ensuring accuracy at the extremes. Overall, this work charts a practical path for integrating advanced LLMs into ASLA systems and provides a clear roadmap for future research to address model bias and enhance discriminative power across the full proficiency spectrum.

14:30 *Minh Trinh The*, *Son Nguyen Van*, *Phuong Nguyen Nam* and *Hanh Nguyen Thi*

DRONEs: Deep Reinforcement Optimization for Network k-Connectivity Restoration Enhancement in UAVs

ABSTRACT. Maintaining k-connectivity is essential for resilient multi-hop UAV communication, yet mixed-integer and heuristic approaches scale poorly with swarm size and operating area. We present a deep reinforcement learning framework with a permutation-invariant, cross-attention encoder that captures inter-agent interactions and generalizes zero-shot from small training swarms (5–50 UAVs) to much larger

fleets (e.g., 100) without retraining. The policy is trained under centralized training with decentralized execution (CTDE), enabling fast, feed-forward decisions at test time. We further apply a Hungarian post-assignment refinement to map goals to vehicles, minimizing total displacement without altering the learned topology. Across benchmarks, our method matches or surpasses strong heuristics on small instances and, at larger scales, achieves real-time performance with markedly lower latency and equal or better solution quality. These results highlight a practical path to scalable k-connectivity restoration in UAV swarms.

14:50 *Toan Nguyen Khac* and *Ngoc Ly Quoc*

XMedCLIP: A Multimodal Deep Neural Network for Bone Pathology Classification from X-ray Image

ABSTRACT. This study proposes a two-stage framework for bone pathology classification under limited data. The pretraining stage aligns a ViT image encoder and a PubMedBERT text encoder in a shared space using bidirectional contrastive learning on paired X-rays and physician diagnoses. The fine-tuning stage then freezes both encoders and trains crossattention fusion and a classifier on paired X-rays with patient self-reports and initial examination notes. The model employs a cross-attention fusion head to combine image and text features before a linear classifier. This framework simulates reasoning of clinicians. With a cross-attention fusion head (Ours-CrsAtt), it reaches 72.78% accuracy in the full-shot setting that outperforms strong unimodal and multimodal baselines. In addition, with a cosine classification head (Ours-Cosine), the model achieves 51.85% accuracy in the one-shot setting and 66.30% in few-shot. Overall, the proposed multimodal architecture learns effectively even with few examples and delivers higher accuracy than single-modality baselines. The method offers a practical and scalable solution for medical imaging workflows.

15:10 *Thiloan Bui*, *Thi Oanh Tran*, *Thị Kim Oanh Nguyễn*, *Chi Tho Luong* and *Vanha Tran*

Automated ESG classification by using Natural Language Processing Techniques from Vietnamese Company Annual Reports

ABSTRACT. As sustainable development gains increasing attention, more and more companies and investors are using environmental, social, and governance (ESG) performance, i.e., non-financial activities, as evalua- tion indicators. Currently, ESG classification and ratings are performed by numerous institutions, and these assessments are subject to human bias, resulting in varying ESG classifications and ratings for the same company. Although many automated ESG classification and rating mod- els employing natural language processing (NLP) techniques have been developed to address this shortcoming, these models are primarily based on the English language. For Vietnam, a developing country, such mod- els are currently lacking. Therefore, in this work, we first construct a Vietnamese-language ESG dataset, collected from the annual sustainabil- ity reports of listed companies in Vietnam. We then employ fine-tuning techniques to fine-tune the bidirectional encoder representations from transformers (BERT) model on this dataset, resulting in an ESG classi- fication model tailored for the Vietnamese market. This model achieves 81.88% accuracy on this dataset. The trained model improves trans- parency in ESG classification and ratings and reduces human bias, pro- viding Vietnamese companies and investors with a reliable tool for as- sessing corporate ESG performance.

13:30-15:30 Session 4C: SOICT Technical Session VII: Applied Operations Research and Optimization

CHAIR: *Markus Westner*

LOCATION: Yersin Ballroom A, 2F

13:30 *Hoang Giang Pham* and *Thuy Anh Ta*

Exponential Cone Reformulation for Scalable Estimation of Quantal Response

and Multinomial Logit Models

ABSTRACT. Quantal response (QR) and multinomial logit (MNL) models are fundamental in behavioral and discrete choice modeling, providing probabilistic frameworks that capture bounded rationality in strategic environments and heterogeneous preferences in individual decision-making. Traditional parameter estimation relies on maximum-likelihood methods solved via gradient-based algorithms, which are sensitive to step-size choices and often scale poorly in high-dimensional or large-scale settings. In this work, we revisit the estimation of both QR and MNL models through convex conic optimization. We show that their maximum-likelihood problems admit exact reformulations as exponential cone programs (ECPs), enabling the replacement of log-sum-exp terms in the likelihood with convex conic constraints. This reformulation allows efficient solution by modern conic solvers using interior-point algorithms with polynomial-time complexity guarantees, thereby ensuring robustness and stable convergence. Numerical experiments on synthetic datasets demonstrate that the ECP approach consistently outperforms gradient-based methods in both runtime and solution quality, highlighting exponential cone programming as a practical and scalable alternative for estimating MNL and QR models.

13:50 *Ban Ha-Bang* and *Do Tuan-Anh*

Reinforcement Learning-Enhanced GRASP for the Multiple Traveling Repairmen Problem with Workload Balance

ABSTRACT. The Multiple Traveling Repairmen Problem with Balanced Workloads (mTRP-WB) is a novel variant of the classical Multiple Traveling Repairmen Problem (mTRP), motivated by real-world applications where fairness in workload distribution is as important as service efficiency. Typical scenarios include urban logistics, preventive maintenance, and healthcare services, where tasks must not only minimize customer waiting times but also be allocated equitably among repairmen to avoid overburdening certain agents. In mTRP-WB, the objective is to minimize a weighted combination of total customer latency and workload imbalance, making the problem computationally challenging as it generalizes both mTRP and load balancing. To address this challenge, we propose a hybrid metaheuristic called GRASP-RL, which combines the Greedy Randomized Adaptive Search Procedure (GRASP) with Reinforcement Learning (RL). In GRASP, the RL-based adaptive strategy dynamically guides the restricted candidate list (RCL) selection, allowing the algorithm to gradually learn which choices lead to a higher-quality initial solution, thereby balancing exploration and exploitation during the construction phase. In the improvement phase, Variable Neighborhood Search (VNS) is employed to exploit promising solution spaces and further enhance solution quality. Extensive computational experiments on benchmark instances demonstrate the impressive efficiency of GRASP-RL across many cases. Compared with state-of-the-art metaheuristics for related problems, GRASP-RL achieves competitive performance despite not being specifically designed for them, highlighting its robustness and scalability.

14:10 *Hue Tran*, *Thai Hoa Nguyen* and *Khanh Phuong Nguyen*

The Min-makespan Vehicle Routing Problem with Drones under Multiple Trips and Visits

ABSTRACT. This paper introduces a novel hybrid pickup routing problem that integrates drones and ground vehicles operating in parallel and independently to minimize the system makespan, while ensuring that the time from each customer's pickup to depot arrival does not exceed a predefined waiting time threshold. The model incorporates practical constraints, including vehicle capacity and drone endurance limits, as well as support for multiple trips with multiple customer visits per trip. We formalize this problem as the Min-makespan Vehicle Routing Problem

with Drones under Multiple Trips and Visits, which has applications in epidemic response and reverse logistics. To the best of our knowledge, this represents the first comprehensive study of this particular problem formulation.

We develop a Mixed Integer Linear Programming model to capture the problem structure and propose an adaptive tabu search algorithm that incorporates multiple neighborhood structures and memory mechanisms to enable effective exploration of the solution space. Extensive computational experiments on a newly developed benchmark dataset with up to 1,000 customers demonstrate the algorithm's efficiency. Furthermore, we show the algorithm's versatility by adapting it to solve related routing problem variants, with comparative results confirming its competitiveness against state-of-the-art methods.

14:30 *Van Quan La* and *Nguyen Hoang Phuong Tran*

Grey Wolf Optimization with Entropy Control for Coverage in DSNs

ABSTRACT. Ensuring reliable and efficient q-coverage in directional sensor networks (DSNs) is a crucial yet challenging task due to heterogeneous coverage requirements, directional limitations, and the NP-hard nature of deployment optimization. To tackle this problem, we propose AACGWO (Adaptive A/C Dynamics Grey Wolf Optimizer), a novel metaheuristic algorithm designed for the target-oriented q-coverage problem. Unlike conventional Grey Wolf Optimizer (GWO) variants that adopt linearly decreasing parameters, AACGWO introduces a cosine-based decay mechanism to ensure smoother phase transitions. In addition, it employs entropy-guided adaptation to dynamically balance exploration and exploitation by monitoring population diversity. The algorithm further incorporates an angle-based encoding strategy to optimize sensor orientations in DSNs. Extensive experiments are carried out under multiple deployment scenarios, covering variations in sensor density, target distribution, sensing range, and spatial configuration. The results consistently demonstrate that AACGWO outperforms recent GWO-based approaches, including DEGWO, IGWO2, ACGWO, and CGGWO, with respect to coverage satisfaction rate, fairness, resource efficiency, and redundancy minimization. These findings confirm the robustness, scalability, and adaptability of AACGWO for real-world DSN applications, particularly in complex and dynamic deployment environments.

14:50 *Quoc-Trung Bui*, *Quang-Dung Pham*, *Van-Son Nguyen* and *Minh Phan*

Modeling and Solving the Bin Packing Problem with Relaxed Capacity Constraints: Applications in Agricultural Land Consolidation in Vietnam

ABSTRACT. This paper introduces the Bin Packing Problem with Relaxed Capacity Constraints (\textbf{BPRC}), a novel variant of the well-known Bin Packing Problem arising from Agricultural Land Consolidation in Vietnam. \textbf{BPRC} aims to assign farming households, each with an expected land area, to agricultural fields of fixed areas such that the total absolute difference between each field's area and the sum of assigned households' expected areas is minimized. In this paper, we prove \textbf{BPRC} is NP-hard via reduction from the 2-Partition problem. To solve it, we propose a Mixed Integer Programming model for exact method using solvers like CPLEX, incorporating optimality properties to enhance efficiency, and a Tabu Search metaheuristic for large-scale instances. Extensive experiments on a real-world instance from Thai Binh province and 27 synthetic instances demonstrate that our methods outperform Vietnam's government guidelines, eliminating added parcels and reducing area deviations significantly. The exact method solves small instances optimally within 30 minutes, while the metaheuristic yields near-optimal or optimal solutions in under 2 minutes, proving practical for real-world applications.

13:30-15:30 Session 4D: SOICT Technical Session VIII: Multimedia Processing

CHAIR: *Khush Agarwal*

LOCATION: Yersin Ballroom B, 2F

13:30 *Thao Thi Phuong Dao*, *Tan-Cong Nguyen*, *Trong-Le Do*, *Mai-Khiem Tran*, *Minh-Khoi Pham*, *Trung-Nghia Le*, *Minh-Triet Tran* and *Thanh Dinh Le*

DTD-Mamba: Dual Teacher Distillation for Mamba in Head and Neck Abscess Segmentation

ABSTRACT. Accurate delineation of head and neck abscess boundaries on contrast-enhanced CT is essential for diagnosis and treatment planning. However, it remains difficult due to ambiguous lesion margins and this region's complex anatomy. In this study, we present Dual Teacher Distillation from Mamba and CNN-based models (DTD-Mamba). This efficient segmentation framework trains a compact Mamba student network under guidance from two heterogeneous teachers. A CNN teacher emphasizes local textures and sharp edges, while a Mamba-based teacher captures long-range dependencies and the global context of related anatomy. Our distillation objective jointly transfers knowledge at multiple levels, allowing the student to inherit both fine-grained details and holistic structure without incurring the computational cost of large models. We evaluate DTD-Mamba on our head and neck abscess dataset, characterized by heterogeneous appearance and intricate anatomy. Our proposed model achieves a Dice Similarity Coefficient of 0.44, an Intersection-over-Union of 0.31, and a Normalized Surface Distance of 0.74. Moreover, this architecture substantially reduces computation and memory requirements compared to its predecessors. These results highlight both the intrinsic challenge of precise boundary delineation in this clinical setting and the benefit of combining local and global supervision during training. DTD-Mamba provides a practical approach to deploying resource-efficient segmentation for complex neck infections. Our code will be released upon paper acceptance.

13:50 *Trong-Hieu Nguyen-Mau*, *Minh-Nam Tran*, *Kim-Trang Phu-Thi*, *Minh-Triet Tran* and *Hai-Dang Nguyen*

VietMed-VQA: A Novel Dataset and Benchmark for Vietnamese Medical Visual Question Answering

ABSTRACT. Medical visual question answering systems hold significant promise for assisting healthcare through automated image analysis. However, most existing systems are English-centric, limiting their accessibility in non-English-speaking regions such as Vietnam. To bridge this gap in multilingual medical AI, we introduce VietMed-VQA, a novel Vietnamese dataset for medical visual question answering, derived from translating established English benchmarks including PathVQA, SLAKE, and VQA-RAD. Leveraging Llama-3.1-70B-Instruct with domain-tailored prompts, we ensure translation accuracy via back-translation, embedding-based semantic checks, and lightweight filtering that removes only the bottom 1\% of pairs. LoRA fine-tuning on three state-of-the-art vision-language models Llama-3.2-11B-Vision-Instruct, Qwen2-VL-7B-Instruct, and LLaVA-1.5-7B, achieves strong results, including 82.9\% accuracy for Llama-3.2-11B on SLAKE. Our ablation studies reveal that subtle filtering enhances performance without sacrificing data volume. Furthermore, evaluations on unseen datasets like MedVQA and WorldMedQA-V demonstrate robust generalization, with accuracies exceeding 64\% on challenging out-of-domain samples. These resources lay the groundwork for AI-driven diagnostics and training in Vietnamese healthcare, paving the way for broader adoption of multilingual medical AI in underserved regions.

14:10 *Thao Thi Phuong Dao*, *Tan-Cong Nguyen*, *Nguyen Chi Thanh*, *Truong Hoang Viet*, *Trong-Le Do*, *Mai-Khiem Tran*, *Minh-Khoi Pham*, *Trung-Nghia Le*, *Minh-Triet Tran* and *Thanh Dinh Le*

MasHeNe: A Benchmark for Head and Neck CT Mass Segmentation using Window-Enhanced Mamba with Frequency-Domain Integration

ABSTRACT. Head and neck masses are space-occupying lesions that can compress

the airway and esophagus and may affect nerves and blood vessels. Available public datasets primarily focus on malignant lesions and often overlook other space-occupying conditions in this region. To address this gap, we introduce MasHeNe, an initial dataset of 3,779 contrast-enhanced CT slices that includes both tumors and cysts with pixel-level annotations. We also establish a benchmark using standard segmentation baselines and report common metrics to enable fair comparison. In addition, we propose the Windowing-Enhanced Mamba with Frequency integration (WEMF) model. WEMF applies tri-window enhancement to enrich the input appearance before feature extraction. It further uses multi-frequency attention to fuse information across skip connections within a U-shaped Mamba backbone. On MasHeNe, WEMF attains the best performance among evaluated methods, with a Dice of 70.45 %, IoU of 66.89 %, NSD of 72.33 %, and HD95 of 5.12 mm. This model indicates stable and strong results on this challenging task. MasHeNe provides a benchmark for head-and-neck mass segmentation beyond malignancy-only datasets. The observed error patterns also suggest that this task remains challenging and requires further research. Our dataset and code will be made publicly available upon acceptance of the paper.

14:30 *Minh Tri Ngo*, *Hieu Trung Dang*, Hoang *Trong Pham*, *Dinh Khoi Nguyen*, *Quyen Nguyen Huu* and *Duy Phan The*

An Optimization-Driven Fusion Framework of Vision–Language Foundation Models for Large-Scale Video Retrieval

ABSTRACT. Video retrieval from large-scale datasets has become increasingly vital as the demand for efficient access to multimodal information continues to grow. Yet, existing approaches often fall short when confronted with complex queries that require understanding both global semantics and fine-grained temporal or contextual dependencies. To overcome these limitations, we propose a hybrid multimodal retrieval framework that unifies the complementary strengths of CLIP and BEiT-3. CLIP offers robust global alignment between visual and textual representations, while BEiT-3 captures nuanced contextual relationships within frames. The framework further integrates LLM-based per-frame captioning, OCR, and ASR modules to enrich multimodal semantics. A temporally aware multi-channel re-ranking mechanism then fuses these representations to accurately retrieve sequential and multi-stage events. Evaluated on the 2025 Ho Chi Minh City AI Challenge dataset, our system achieves notable gains in retrieval accuracy for complex and context-dependent video queries.

14:50 *Ngoc-Thao Le*, *Cat-Thanh Hoang-Le* and *Quoc-Ngoc Ly*

Text-Driven 3D Interior Scene Generation using 3D Gaussian Splatting

ABSTRACT. Text-driven generation of 3D(3D) indoor environments underpins applications in AI-assisted interior design, e-commerce visualization, and immersive XR. Yet practical systems still struggle to turn natural language into coherent, editable 3D scenes with high visual and geometric fidelity. We present a pipeline focused on robust panorama initialization, geometry-preserving optimization, and efficient view completion. Concretely, we introduce (i) a Best-of-N panorama selection guided by a composite CLIP–Discontinuity score to enforce semantic alignment and global wrap-around consistency, (ii) engineering optimizations to Progressive Novel View Inpainting (PNVI) that improve throughput, and (iii) depth-regularized 3D Gaussian Splatting (3DGS) to stabilize geometry during optimization. We further curate a bilingual (EN/VI) evaluation set of 50 prompts spanning 34 room types to stress diversity and linguistic coverage. Empirically, our components jointly yield more faithful, structurally consistent, and editable indoor scenes from text: Best-of-N selection improves semantic alignment (CLIP: 55.69 to 60.03) and reduces discontinuities (111.48 to 45.24), depth-regularized 3DGS enhances geometric fidelity (PSNR: +1.01 dB, SSIM: +0.0193), and PNVI optimizations achieve 35% runtime reduction (122s to 80s) while maintaining practical efficiency.

15:10 *Chinh Nguyen Minh*, *Long Tran Ngoc*, *Khoi Tran Man*, *Long Le Hoang Hien*, *Van Thai Hung*, *Duy-Dinh Le* and *Thanh Duc Ngo*

When Events Speak: MLLM-Guided Video Retrieval with Temporal Reranking

ABSTRACT. The rapid expansion of online video content has intensified the need for efficient retrieval of specific moments. However, most existing video retrieval methods fail to capture short-lived events that occur within only a few frames and overlook dynamic cues such as motion, audio, and dialogue. This lack of event awareness prevents the system from understanding how actions evolve, leading to fragmented interpretations of event sequences. To address these challenges, we propose an event-aware video retrieval system that integrates MLLM-based event capturing with a temporal reranking mechanism. The MLLM captures both coarse-grained and fine-grained events, providing semantically rich and temporally aligned descriptions. The temporal reranking module then re-assesses the entire event sequence, dynamically relinking results based on recomputed sequence scores to maintain temporal coherence. This design allows the system to retrieve results that are both semantically meaningful and temporally coherent. Evaluated on the Ho Chi Minh AI Challenge 2025, our system achieved 85 out of 88 correct results with accuracy above ninety-five percent across three evaluation stages, demonstrating strong robustness in complex event-sequence retrieval.

13:30-18:00 Session 4E: Poster Exhibition

CHAIRS: *Nguyen Linh Trung*, *Khoat Than* and *Duc-Hau Le*

Đăng Lê

VisionCare: Compute-Aware Hybrid CNN–Transformer Heads for Multi-Disease Retinal Diagnosis with Explainable AI

ABSTRACT. Population-scale retinal screening demands models that are accurate, interpretable, and computationally efficient. We present VisionCare, a unified and reproducible framework for multi-disease diagnosis from color fundus photographs, trained end-to-end using a single, fixed recipe at a standard input resolution of 224×224. Built on a ConvNeXt-Base backbone, VisionCare explores three progressively more expressive classification heads: (A) a top-down feature pyramid with generalized mean pooling (GeM) and squeeze-and-excitation (SE); (B) a compute-aware reverse-pyramid design that compresses multi-scale features to the deepest stride with dilated refinement; and (C) a dual-path fusion module that incorporates pooled-key/value self-attention with a learned gating mechanism to balance convolutional and non-local representations. VisionCare supports case-level interpretability via Grad-CAM and Grad-CAM++, and includes lightweight integration points for tele-ophthalmology deployment. On a benchmark dataset with 11 diagnostic categories, the attention-gated head (C) achieves a test accuracy of 0.9138, macro-F1 of 0.9289, Cohen's kappa of 0.9014, and a macro-AUROC of ≈0.996 without increasing input resolution or modifying the training procedure. By standardizing the pipeline and holding optimization constant, we isolate architectural contributions and maintain practical feasibility for deployment on commodity GPUs.

Truc Thi-Thanh Le, *Mai Thi-Ngoc Vu* and *Luan Thanh Nguyen*

An In-Depth Investigation into Vietnamese LexicalText Normalization on Social Media

ABSTRACT. Informal Vietnamese on social media, often rich in abbreviations and misspellings, poses significant challenges for natural language processing (NLP). This study presents an in-depth investigation of Vietnamese lexical normalization through evaluating a broad range of approaches, including deep learning models, transformer-based transfer learning, and prompting-based large language models. The experimental results show that ViT5, a Vietnamese-specific T5 variant, delivers the highest performance, achieving a 63.06% error reduction rate (ERR) over the Leave-

As-Is baseline. At the same time, LLM-based prompting approaches yield only moderate gains and, in some cases, even underperform due to over-normalization and loss of context. We then assess the impact of normalization on five downstream Vietnamese social media NLP tasks. Normalization markedly boosts performance in Constructive Speech and Toxic Speech Detection (F1 improvements from 69.29% to 83.82% and from 68.25% to 79.77%, respectively), while showing limited or even negative effects on tasks that require the preservation of subtle linguistic cues, such as Emotion Classification and Hate Speech Detection. These findings underscore the value of targeted normalization in improving task performance where non-standard text forms impede model understanding, while cautioning against its blanket application in contexts reliant on nuanced linguistic features.

Le Van-Vinh, *Nguyen Nhat-Hoang*, *Nguyen Duc-Quyen* and *Dang Duc-Hanh*

A Method for Composing Concerns into a Unified Domain Model in Domain-Driven Design

ABSTRACT. Domain-Driven Design (DDD) emphasizes iterative development around a rich domain model to align developers and domain experts. While ubiquitous language and Domain-Specific Languages (DSLs) improve expressiveness and maintainability, modern systems often require multiple heterogeneous DSLs to cover diverse concerns. Existing DDD approaches, however, lack systematic methods to compose such DSLs, resulting in fragmented models and limited automation. Although meta-modeling offers a standard way to define DSLs, it is often rigid and framework-dependent. This paper introduces a novel method for composing heterogeneous concern DSLs into a unified domain model within DDD. Each DSL is defined with consistent syntax and formal semantics, and integrated via an annotation-based composition mechanism at the abstract syntax tree (AST) level. This ensures concern orthogonality, model cohesion, and supports consistency checking, automated code generation, and traceability. The approach is implemented using JetBrains MPS and the JDA framework, and validated through representative case studies, advancing modular and executable domain modeling for complex systems.

Tram Doan Thi, *Vy Vu Ngoc Thao*, *Thanh Le Tuan Minh*, *Lam Nguyen Luu Phuong Ngoc*, *Hoang Nguyen Tran Le* and *Binh Nguyen Thanh*

MedPRS: Scientific Paper Submission Recommendation System for Medical Research

ABSTRACT. A paper submission recommendation system aim to assist researchers in choosing suitable journals or conferences for their work. This research topic has been extensively studied during the last five years. The need for such systems is especially critical in the medical field, where the rapid expansion of biomedical literature makes selecting appropriate venues increasingly challenging. In this study, we propose three approaches developed on a newly constructed dataset comprising 1.2million biomedical articles from 1,406 journals. The dataset is enriched with metadata such as titles, abstracts, keywords, journal Aims & Scope, and a newly introduced feature, Categories. By leveraging domain-specific transformer models, including BioBERT and BioMedBERT, our system achieves strong performance, with Top-1 to Top-10 accuracies ranging from 0.6865 to 0.9582.

Khanh Linh Tran, *Minh Nguyen Dang*, *Hung Nguyen Quoc*, *Thien Nguyen Trong* and *Linh Nguyen Kieu*

Enhancing YOLOv11n for Reliable Child Detection in Noisy Surveillance Footage

ABSTRACT. This paper presents a practical and lightweight solution for enhancing child detection in low-quality surveillance footage, a critical component in real-world missing child alert and daycare monitoring systems. Building upon the efficient YOLOv11n architecture, we propose a deployment-ready pipeline that improves detection under challenging conditions including occlusion, small object size, low

resolution, motion blur, and poor lighting, common in existing CCTV infrastructures. Our approach introduces a domain-specific augmentation strategy that synthesizes realistic child placements using spatial perturbations (e.g., partial visibility, truncation, and overlaps) combined with photometric degradations (e.g., lighting variation and noise). To improve recall of small and partially occluded instances, we integrate Slicing Aided Hyper Inference (SAHI) at inference time. All components are trained and evaluated on a filtered, child-only subset of the Roboflow Daycare dataset. Compared to the baseline YOLOv11n, our enhanced system achieves a mAP@0.5 of 0.967 and mAP@0.5:0.95 of 0.783, yielding absolute improvements of 0.7% and 2.3% respectively, without architectural changes. Importantly, the entire pipeline maintains compatibility with low-power edge devices and supports real-time performance, making it highly suitable for cost-sensitive deployments in industrial surveillance applications. The example augmented dataset and the source code used to generate it are available at: https://github.com/html-ptit/Data-Augmentation-YOLOv11n-child-detection

Dang Thi Thu Ha, *Nguyen Dinh Van* and *Nguyen Duc Hoa*

Accurate Mixed-Gas Concentration Prediction in Electronic Nose Using Image-Guided Autoencoder–TCN Hybrid Model

ABSTRACT. Reliable estimation of gas mixture concentrations is fundamental for advancing intelligent electronic-nose technologies. Nevertheless, the accurate prediction of gas mixture concentrations remains a challenge because of the complex interactions between the gases, drift, and noise from the sensors. Many of the existing methods are limited in terms of the generalizability of the model. To address this challenge, this study presents an image-based Autoencoder–TCN hybrid model to predict gas concentrations in mixtures using an electronic nose sensor array. Raw sensor signals are converted into images, allowing the autoencoder to extract deep nonlinear features that reflect the complex interactions between different gases. These features are then fed into the temporal convolutional network (TCN) regressor, resulting in an accurate estimation of multicomponent gas concentrations. The experiment was conducted on both a public dataset consisting of a mixture of methane and ethylene and a private dataset consisting of a mixture of ammonia and hydrogen sulfide. The proposed method achieved a high accuracy and demonstrated noise tolerance and generalizability. Under five-fold cross-validation, our model achieved superior performance compared to baseline models, such as MLP, CNN, RNN, and XGBoost. This combination opens up the potential for real-time gas monitoring in environmental and industrial applications.

Cao Doanh Bui, *Hoai Luan Pham*, *Vu Trung Duong Le*, *K. Maï Nguyen* and *Yasuhiko Nakashima*

Merging-based Federated Learning for Lifelong Whole Slide Image Analysis with Vision-Language Models

ABSTRACT. Whole Slide Images (WSIs) are gigapixel-scale pathology images essential for accurate cancer diagnosis and prognosis. However, current WSI analysis methods typically require training a separate model for each cancer type, leading to scalability issues, high computational cost, and practical limitations in data sharing across institutions. Recent continual learning approaches aim to build unified multi-task models but remain constrained by fixed class/task assumptions and the requirement of centralized training data, which is often impractical due to privacy concerns and the massive size of WSIs. In this study, we propose an efficient federated learning framework for WSI analysis that aggregates knowledge from distributed models into a single multi-purpose model without requiring raw data sharing. Our method leverages pathology vision–language models (VLMs) as the backbone, fine-tunes them on local tasks using class prompts, and merges the resulting weights through a model-merging

strategy. We evaluate the approach on six TCGA cancer subtyping tasks under both class-incremental (CLASS-IL) and task-incremental (TASK-IL) settings. Experimental results demonstrate that our method consistently outperforms continual learning and zero-shot baselines, achieving the best trade-off between accuracy and forgetting. These findings highlight the potential of federated model merging to enable scalable, privacy-preserving, and clinically useful WSI analysis.

Du Doan and *Khanh-Duy Nguyen*

Domain-Incremental Learning for UAV Traffic Video Anomaly Detection

ABSTRACT. Anomaly detection in video is a crucial research area, but traditional methods face challenges when environmental conditions change, particularly in real-world scenarios. This paper addresses this issue by applying Domain-Incremental Learning (DIL) to anomaly detection in UAV-based traffic surveillance, enabling the model to better adapt to different weather conditions while mitigating catastrophic forgetting. We experiment with four prominent anomaly detection methods: Future-Frame Prediction (FFP), Spatio-Temporal Dissociation (STD), Margin Learning Embedded Prediction (MLEP), and Memory-guided Normality for Anomaly Detection (MNAD). These methods are combined with forgetting mitigation strategies such as Elastic Weight Consolidation (EWC) and Experience Replay (ER). The experiments are conducted across three weather domains: clear, snow, and fog. The results show that FFP outperforms the domain-specific models, achieving 1-4\% higher performance, demonstrating good generalization across domains. STD performs well in foggy conditions (AUC = 56.54), while MLEP and MNAD struggle with forgetting knowledge from previous domains, showing unstable performance (AUC = 51.63 and 51.14 on the Original domain). These results highlight the potential of DIL in developing flexible and efficient anomaly detection systems, and also point to the need for improving forgetting mitigation strategies to optimize model generalization, particularly with methods such as MLEP and MNAD.

Ly-Huynh Phan, *Chi Minh Hieu Nguyen*, *Dinh-Dat Nguyen*, *Quang-Dung Dang* and *Truong-Giang Nguyen*

A Dual-Path approach for Time Series Anomaly Detection in Building Environmental Sensors

ABSTRACT. Anomaly detection within environmental time series data plays a crucial role in modern monitoring systems, yet it continues to pose challenges due to the inherently complex and nonlinear nature of sensor-generated signals. This study proposes a dual-path approach for time series anomaly detection that combines the expressive capabilities of deep learning with the transparency of classic machine learning techniques. The approach integrates a bidirectional Long Short-term memory (LSTM) autoencoder for extracting temporal features with density-based outlier detection algorithms, specifically Local Outlier Factor (LOF) and Isolation Forest. This methodology effectively models time-dependent patterns while maintaining a balance between interpretability and computational cost. The proposed approach shows significant improvements compared to standalone deep learning and conventional statistical approaches, across various evaluation metrics through extensive testing on indoor environmental sensor datasets. The results analyze the different impacts of components on the anomaly detection process: Isolation Forest (49.08%), Reconstruction Error (39.27%), and LOF (11.65%). Using synthetic data with differing noise intensities improved the model's resilience across diverse anomaly categories—point, contextual, and collective-achieving detection rates above 86%. These findings highlight the approach's practical value in real-world environmental monitoring by balancing high accuracy and interpretability.

Yuri Seo, *Hyeon-Ki Jo* and *Eui-Nam Huh*

FLoRA-KD: Efficient Communication in Federated Learning for Multi-Organ Segmentation through LoRA Knowledge Distillation

ABSTRACT. Models for medical image analysis require large-scale datasets with expert-annotated labeling. However, most datasets are either partially labeled or collected from multiple institutions, leading to issues such as data inconsistency and quality problems. Additionally, medical data sharing is restricted by privacy regulations. Federated learning is a well-established approach to address these challenges. It ensures data privacy by sharing model parameters instead of raw data, mitigating privacy concerns. However, previous studies on multi-organ CT segmentation and federated learning did not consider client scalability and communication overhead. Due to the heterogeneity of partially labeled data, repetitive parameter sharing between the server and clients can lead to issues including increased communication overhead, scalability limitations, and potential delays. To address these limitations, we propose "FLoRA-KD" an efficient federated learning method for reducing communication costs on multi-organ segmentation, which uses LoRA aggregation and LoRA knowledge distillation with partially labeled datasets. FLoRA-KD initializes and trains each client with the global LoRA, allowing efficient fine-tuning on local private datasets with minimal parameter updates. Moreover, sharing each client's LoRA adapter enables knowledge transfer from the latest updated parameters on other datasets. Our proposed method was evaluated on three publicly available abdominal CT datasets. Experimental results demonstrated that FLoRA-KD outperformed state-of-the-art methods in communication efficiency while achieving high accuracy.

Do Minh Duc, *Quan Xuan Truong*, *Nguyen Tat Dat* and *Nguyen Van Vinh*

Auto-Prompting with Retrieval Guidance for Frame Detection in Logistics

ABSTRACT. Prompt engineering plays a critical role in adapting large language models (LLMs) to complex reasoning and labeling tasks without the need for extensive fine-tuning. In this paper, we propose a novel prompt optimization pipeline for frame detection in logistics texts, combining retrieval-augmented generation (RAG), few-shot prompting, chain-of-thought (CoT) reasoning, and automatic CoT synthesis (Auto-CoT) to generate highly effective task-specific prompts. Central to our approach is an LLM-based prompt optimizer agent that iteratively refines the prompts using retrieved examples, performance feedback, and internal self-evaluation. Our framework is evaluated on a real-world logistics text annotation task, where reasoning accuracy and labeling efficiency are critical. Experimental results show that the optimized prompts—particularly those enhanced via Auto-CoT and RAG—improve real-world inference accuracy by up to 15\% compared to baseline zero-shot or static prompts. The system demonstrates consistent improvements across multiple LLMs, including GPT-4o, Qwen 2.5 (72B), and LLaMA 3.1 (70B), validating its generalizability and practical value. These findings suggest that structured prompt optimization is a viable alternative to full fine-tuning, offering scalable solutions for deploying LLMs in domain-specific NLP applications such as logistics.

Dac Trung Huy Nguyen, *Thi Xuan Ly Nguyen* and *Dinh Ngoc Tram Pham*

Factors Influencing the Actual Use of AI-Enabled Chatbots in Digital Wallets for Personal Financial Management Among Vietnamese Online Users

ABSTRACT. As digital wallets gain popularity in Vietnam, AI-enabled chatbots are powerful tools for personal financial management, yet their actual usage remains limited and underexplored. This study investigates the key factors influencing their adoption by integrating the Information Systems Success model, Unified Theory of Acceptance and Use of Technology 2, and Innovation Resistance Theory into a unified framework. Survey data from 543 Vietnamese users reveal that social influence, effort

expectancy, performance expectancy, hedonic motivation, and AI self-efficacy positively impact behavioral intention, which in turn strongly drives actual usage. Meanwhile, the tradition barrier, image barrier, and risk barrier negatively affect intention. Quality factors like system quality, information quality, and anthropomorphism also contribute to shaping performance and effort expectations. Interestingly, service quality did not show a significant direct effect on performance expectancy. These findings suggest that to foster adoption, digital wallet providers should focus on enhancing user trust, reducing perceived barriers, and delivering intuitive, personalized, and socially supported AI experiences.

Hien Do Hoang, *Phan Ba Cuong*, *Pham Do Thanh*, *Hien Do Thi Thu*, *Nguyen Tan Cam* and *Van-Hau Pham*

Toward Adaptive Web Application Honeypots: Fine-Tuned Large Language Models for Realistic Response Emulation

ABSTRACT. With the rise of complex attacks, traditional web honeypots struggle to maintain authenticity due to static and easily fingerprinted responses. This paper presents a fine-tuned Large Language Model (LLM)-powered honeypot framework that generates dynamic, context-aware responses closely aligned with real web application behaviors. To achieve this, we collect realistic requests and responses from target applications, preprocess them by extracting essential information from requests and normalizing responses, and fine-tune the LLM on these request-response pairs. Experimental results demonstrate that the proposed method consistently outperforms both raw-training baselines and existing LLM-based honeypots, namely Galah and VelLMes. Specifically, it achieves higher similarity scores across all metrics, with Cosine similarity reaching 0.9396 compared to 0.4506 for Galah and 0.7357 for VelLMes. Moreover, it yields a substantially lower Levenshtein distance at 329.35 compared to 564.98 for the baseline, 2940.25 for Galah, and 2340.99 for VelLMes. These improvements confirm the model's ability to generate highly realistic, structurally valid, and functionally robust responses, thereby enhancing attacker engagement and deception effectiveness.

Hieu Pham, *Trung Pham*, *Vi Nguyen*, *Long Nguyen*, *Truong Tran*, *Duc Nguyen*, *Tuong Nguyen*, *Long Nguyen*, *Huong Ha* and *Tho Quan*

GAFB-MKL: Adaptive Filter Banks via Genetic Algorithm and Sparse Multiple Kernel Learning for EEG-based Motor Imagery Classification

ABSTRACT. Motor–imagery EEG (MI–EEG) decoding is difficult due to non-stationarity, inter-subject variability, and the concentration of discriminative information in narrow, person-specific sub-bands. We present GAFB–MKL, a compact and interpretable pipeline that adapts continuous, subject-specific filter banks with a Genetic Algorithm (GA), weights bands via sparse, nonnegative Multiple Kernel Learning (MKL) on FBCSP features, and classifies with a precomputed-kernel SVM. Unlike fixed-bank designs with post hoc feature selection, GAFB–MKL unifies supervised band adaptation and label-aligned fusion, yielding bandlevel attribution with few hyperparameters and no heavy training. On BCI Competition IV-2b, GAFB–MKL attains 79.65% average accuracy and peaks at 96.56% on Subject 4, surpassing fixed-bank FBCSP variants and slightly exceeding EEGNet (79.44%) under a unified preprocessing/evaluation pipeline. On the HCMIU hand–binary dataset, it reaches 63.21% on average, outperforming FA-GPNet (57.44%), CNN baselines, and classical Riemannian/FBCSP pipelines. Ablations corroborate the design principle discover, then distill: removing GA (fixed bank + MKL) reduces performance to 75.55%, while removing MKL (GA + SVM) yields 74.44%. Beyond accuracy, learned band weights consistently reveal subject-specific μ/β peaks, supporting fast calibration and transparent deployment in resource-constrained BCI.

Huong Dinh, Anh Phan, Truong Vu, Ngoc Vi and Anh Tran

Linguistic and Semantic Graph-based Neural Networks for Hate Speech Detection

ABSTRACT. Hate speech detection remains a challenging problem due to the diversity of expressions, implicit abuse, and frequent use of slang or coded language. Although pretrained and large language models (LLMs) have demonstrated strong capabilities, they often struggle with these issues. This paper proposes a graph-based framework that models relationships between text samples using both syntactic and semantic similarities. Each dataset is represented as a graph where nodes are text samples and edges are constructed based on token matching and cosine similarity of pretrained embeddings. Two graph neural network architectures—Graph Attention Networks (GAT) and the Unified Message Passing model (UniMP)—are employed to capture structural and contextual dependencies. Experiments on four real datasets show that our models consistently outperform LLMs on multi-class tasks. These findings highlight the effectiveness of GNNs as a resource-efficient alternative to LLMs for robust hate speech detection.

Lu Le Phuc, Trung Duong Chi, Vinh Do Le, Thanh Tran Tan, Khang Nguyen Phu Bao, Doan Vo Nam Thuc and Hai Nguyen Son

A deep learning model for drug-target interactions prediction in drug discovery

ABSTRACT. Accurately predicting interactions between drugs and proteins is a crucial step in the drug discovery process. While most current research focuses on regression-based prediction of binding affinity values, many real-world applications only require identifying whether a significant interaction exists. This study proposes an effective model for predicting binding affinity values in drug–target interaction (DTI) tasks by leveraging a Long Short-Term Memory (LSTM) network to extract sequential features from SMILES strings of drugs and amino acid sequences of proteins. The learned representations are then concatenated and passed through a regression layer to estimate the binding affinity. The training data is constructed from benchmark datasets, where binary interaction labels are derived based on predefined affinity thresholds. Experimental results demonstrate that the LSTM-based model achieves high predictive accuracy while reducing model complexity compared to regression-based approaches, highlighting its feasibility and effectiveness for virtual screening of drug–target interactions.

Seung-Woo Jeong, Hacksung Boo, Hoang-Hai Tran and Eui-Nam Huh

Optimization of Resource Allocation Using SLA Violation Penalty and Workload Prediction in Cloud Datacenters

ABSTRACT. This paper proposes a Service Level Agreement (SLA) Violation Penalty (SVP) mechanism as a new determinant for optimizing resource allocation in cloud datacenters. The proposed approach leverages time series workload prediction models to comprehensively consider prediction accuracy, inference latency, and node workload, effectively reducing SLA violations while improving system availability and operational efficiency. We evaluate ARIMA, LSTM, Bi-LSTM, LSTM-Attention, and LSTM-Autoencoder models on real-world datasets from Alibaba, Google Cluster, and Bitbrains, simulating diverse workload patterns. Results show that SVP-based allocation significantly improves node availability and lowers violation rates compared to prediction-based and minimum load methods. The mechanism also maintains stable performance in both resource-limited and abundant environments. Overall, SVP provides an effective strategy for minimizing SLA-related costs and improving datacenter performance.

Minh Lu, An Trinh, Vinh Nguyen, Nhan Nguyen, Minh Nguyen, Duc Bui and Quang Tran

Lightweight Multi-Trait IELTS Essay Scoring with Prompt- and Topic-Awareness

ABSTRACT. Automated essay scoring (AES) systems have been increasingly developed

to assist in evaluation of IELTS Writing tasks. However, most existing systems typically predict a single overall score and are often too computationally expensive for low-end servers. In this paper, we introduce a lightweight, multi-trait AES model that scores essays on the four IELTS criteria–Task Response (TR), Coherence and Cohesion (CC); Lexical Resource (LR), and Grammatical Range and Accuracy (GRA)–by jointly encoding the prompt and essay with BERT, augmenting representations with topic-distribution features from FASTopic, and modeling inter-trait interactions with an attention mechanism and a trait-similarity loss. Each trait is derived from the shared prompt–essay representation and informed by topic features, interacting through attention mechanism, and producing interpretable trait-level scores. To enable low-resource deployment we distill the model into a more compact model. Evaluated on the chillies IELTS writing dataset (10,000+ essays), our model raises quadratic weighted kappa (QWK) by 2.78% over base BERT and outperforms prompting-based LLMs; the distilled model is approximately 5× faster and 4× smaller while maintaining competitive accuracy–offering a practical solution for real-world EdTech applications. This work provides a practical and interpretable AES solution, bridging research models and scalable deployment. Our prototype is available at https://study.engonow.com/ielts_writing_scoring

Quang Dũng Nguyễn, *Đức Dũng Nguyễn*, *Hữu Trí Dũng Võ* and *Thanh Hương Lê*

ViLexCPO: A Multi-Task and Preference-Aligned Framework for Legal Question Answering

ABSTRACT. In legal Question Answering (QA), especially for Vietnamese, building reliable systems is challenging because there is not enough high- quality training data, and the tasks often require complex logical reason- ing and accurate legal citations. Previous studies have focused on large language models (LLMs), but these models are difficult to use in practice due to high infrastructure requirements. This creates a need for small- to-medium language models (SLMs) that are specially trained and optimized for the legal domain. To address this need, we propose a two-stage training framework. In the first stage, we use multi-task supervised fine- tuning (SFT) to train the model to perform three tasks at the same time: (1) determine the usefulness of legal citations, (2) answer multiple-choice legal questions, and (3) predict the most relevant citations. In the second stage, we apply Contrastive Preference Optimization (CPO) to further align the model with high-quality human feedback, improving both the accuracy and legal soundness of its responses. Experiments on the VLSP LegalSLM dataset show that our approach improves citation accuracy by up to 20% compared to using SFT alone. It also performs well across all three tasks, with especially strong results in syllogism-based QA.

Ngoc-Bich Nguyen Thi, *Tam Vo Minh*, *Bich Van Nguyen*, *Khoa Nguyen* and *Vinh-Tiep Nguyen*

MSA: Breaking Down MOET Criteria into Sub-Criteria for Education

ABSTRACT. At the global level, efforts to decompose curricula and educational standards have applied techniques such as natural language processing (NLP) and adaptive syllabus generation. While effective in certain contexts, these approaches struggle with vague or high-level criteria and often lack the flexibility required for practical classroom use. Recent studies suggest that large language models (LLMs) such as GPT-4 can support tasks like exercise generation, syllabus development, and course mapping. However, they frequently exhibit overgeneralization, omissions, and cultural bias, indicating the need for structured decomposition strategies and human oversight. To address this gap, we propose MOET Sub-Criteria Analytic (MSA) a method that leverages in-context learning with LLMs to decompose the Ministry of Education and Training (MOET) criteria into actionable sub-criteria. To ensure the quality and pedagogical soundness of the generated outputs, we introduce EduQual-

5, a systematic evaluation framework grounded in educational measurement theory and policy research. EduQual-5 operationalizes five interrelated dimensions—relevance, validity, clarity, feasibility, and fairness—into a transparent Likert-scale assessment conducted through both expert judgment and structured analysis. Experiments indicate that MSA produces sub-criteria rated at competitive Likert results compared to real sub-criteria of MOET.

Ngoc Minh Nguyen Thi, *Thanh Van Tran Thi*, *Nang Diu Dao* and *Diep Thi Hoang*

XGPhy: A Machine Learning Framework for Predicting Optimization Difficulty in Maximum Likelihood Phylogenetic Inference

ABSTRACT. In the task of inferring phylogenetic trees from multiple se- quence alignments, widely-used maximum likelihood tools such as IQ- TREE and RAxML rely on heuristic optimization techniques to approx- imate globally optimal solutions. When the solution space exhibits a high density of near-optimal trees—indicative of a flat or rugged combinato- rial landscape—search effectiveness can be compromised. A key open problem is how to automatically characterize the hardness of a phylogenetic inference instance before performing the search, enabling adaptive selection of search heuristics and hyperparameters. In this study, we present XGPhy, a machine learning framework that predicts instance-level difficulty in maximum likelihood phylogenetic in- ference. The model is trained using features and training data derived through a procedure adapted from Pythia (Haag, 2022), and leverages the XGBoost algorithm for supervised prediction. Preliminary results show that XGPhy's difficulty scores are positively correlated with Pythia's, indicating its promise as an automated pre-search diagnostic tool.

Linh Nguyen, *Lan Phan* and *Hiep Huynh*

Enhancing User-Based Context-Aware Collaborative Filtering Using Energy Distance with Post-Filtering Contextual Features

ABSTRACT. This paper introduces an enhanced framework for user-based context-aware collaborative filtering (CACF-EDPF) that incorporates energy distance (ED) as a robust metric to capture distributional discrepancies in user–item interactions, alongside a post-filtering (PF) mechanism to refine recommendations based on contextual features. Unlike conventional CACF methods, which often struggle to effectively integrate contextual information, the proposed CACF-EDPF approach models user preferences with greater fidelity by jointly leveraging statistical distance and contextual relevance. Comprehensive experiments on three benchmark datasets—MovieLens, Amazon, and Yelp—demonstrate that CACFEDPF consistently outperforms both standard CACF and energy-based models (EBM) in terms of prediction accuracy and adaptability to context. These findings highlight the effectiveness of combining distribution-aware similarity with contextual post-filtering, pointing toward a promising direction for developing more accurate, flexible, and context-sensitive recommender systems for real-world applications.

Dat Tran, *Nguyet Anh Le* and *Hoai Nam Vu*

TI-FS: Text and Images Mutual Support for Improving Few-Shot Learning in Cross-Device Image Recapture Detection

ABSTRACT. Text-image joint training has been extensively employed in computer vision tasks, particularly in cross-device image recapture detection. Moreover, utilizing text-image combined models for few-shot deep learning model fine-tuning has shown promising performance improvements. However, the integration of these two modalities presents significant challenges, and issues related to the quality of cross-device image recapture detection are often difficult to distinguish. To address these challenges, we develop a mutual guidance mechanism that enables the text-image joint training model to guide the few-shot deep learning model (TI-FS) through image representations and textual guidance components. Our extensive experiments

demonstrate that our TI-FS model significantly outperforms current state-of-the-art methods in both general image recognition tasks and specifically in cross-device image recapture detection.

Tan Khang Huynh, *Ha Dung Nguyen* and *Thanh Binh Nguyen*

ViConBERT: Context-Gloss Aligned Vietnamese Word Embedding for Polysemous and Sense-Aware Representations

ABSTRACT. Recent advances in contextualized word embeddings have greatly improved semantic tasks such as Word Sense Disambiguation (WSD) and contextual similarity, but most progress has been limited to high-resource languages like English. Vietnamese, in contrast, still lacks robust models and evaluation resources for fine-grained semantic understanding. In this paper, we present ViConBERT, a novel framework for learning Vietnamese contextualized embeddings that integrates contrastive learning (SimCLR) and gloss-based distillation to better capture word meaning. We also introduce ViConWSD, the first large-scale synthetic dataset for evaluating semantic understanding in Vietnamese, covering both WSD and contextual similarity. Experimental results show that ViConBERT outperforms strong baselines on WSD (F1 = 0.87) and achieves competitive performance on ViCon (AP = 0.88) and ViSim-400 (Spearman's rank correlation = 0.60), demonstrating its effectiveness in modeling both discrete senses and graded semantic relations. Our code, models, and data are available at https://github.com/tkhangg0910/ViConBERT.

Man Nguyen, *Thu Nguyen*, *Khoa Tan Vo*, *Phuc Nguyen* and *Tu-Anh Nguyen-Hoang*

LoDiBi: Automated Course Quality Evaluation Framework with LOQCA, DeepIFSA, and BiLSTM

ABSTRACT. Course quality is a critical factor shaping learner experience and institutional success. High-quality courses help learners achieve goals and ensure compliance with education standards. Traditional evaluation relies on expert review after course completion, which is slow, costly, and delays improvement. In Massive Open Online Courses (MOOCs), challenges are greater due to sparse and fragmented learning data. We propose LoDiBi (LOQCA, DeepIFSA, BiLSTM), a framework with three integrated modules. LOQCA automatically labels course quality from learner behavior. DeepIFSA imputes missing values using attention, CutMix, and contrastive learning, making it effective in sparse settings. BiLSTM captures temporal learning patterns to enhance prediction accuracy. Combined, these modules enable early prediction of course quality and provide instructional designers with actionable evidence for timely adjustments. Experiments on real-world MOOC datasets show that LoDiBi outperforms existing methods. Data quality was maximum (Completeness and Consistency reached 1). Balanced classification was achieved (Macro-F1, Balanced Accuracy greater than 0.9). Strong agreement with ground-truth labels was confirmed (MCC and Kappa greater than 0.9). Predictive performance was also high (Accuracy, Precision, and Recall between 0.93 and 0.94). LoDiBi provides a scalable solution for automated course evaluation, helping institutions make faster, data-driven decisions to enhance learning outcomes.

Thi Ha Dinh, *Thi Lich Nghiem* and *Ngoc Hoa Nguyen*

Exploring Consumer Behavior in Clean Food Consumption using Positive–Negative Association Rule Mining: A case study in Vietnam

ABSTRACT. The growing demand for clean food, including safe and organic products, reflects increasing global concerns about food safety, health, and sustainability. While developed markets such as Europe, North America, and parts of Asia have experienced steady growth in organic food consumption, Vietnam has more recently witnessed a rapid rise in demand, driven by greater health awareness and concerns over food contamination. However, the domestic market still faces challenges such as high prices, supply chain constraints, and limited consumer trust in certification.

Previous studies mainly applied econometric methods to analyze purchasing intentions and willingness to pay, whereas recent advances in machine learning have enabled deeper insights into consumer behavior. Among these, association rule mining is particularly effective in uncovering hidden consumption patterns. This study employs positive–negative association rule mining to identify both frequent and infrequent purchasing habits in the clean food sector. The analysis was conducted using the dataset of a clean food retail chain in Hanoi. The experimental results provide implications for supply planning, inventory management, and the promotion of sustainable consumption.

Phong Huy Nguyen, *Xuan Phuc Nguyen* and *Ngoc Hoang Luong*

Optimization of Kolmogorov–Arnold Networks for Reinforcement Learning via NeuroEvolution of Augmenting Topologies

ABSTRACT. The Kolmogorov–Arnold Network (KAN) emerges as an alternative neural architecture that replaces fixed activation functions with learnable basis functions, enabling improved interpretability and efficiency in function approximation. In this work, we propose a novel method for evolving KAN architectures using the NeuroEvolution of Augmenting Topologies (NEAT) algorithm. Unlike conventional gradient-based optimization, our approach explores both the structure and functional composition of networks through evolutionary search, allowing the discovery of compact and expressive models without backpropagation. We modified NEAT to incorporate the KANs formulation at each node, enabling the evolution of not only connectivity patterns but also node-specific functional mappings. Experimental results demonstrate that evolved KANs can achieve competitive performance in reinforcement learning tasks. This study highlights the potential of combining KANs with evolutionary computation to develop interpretable and gradient-free learning systems.

The Nguyen Huu, *Minh Thu Vu Thi*, *Hai Anh Le Thi* and *Tien Nguyen Huu*

Towards Regional AQI Mapping in Northern Vietnam: Multi-Source Data Fusion and Ensemble Learning

ABSTRACT. Air pollution, particularly fine particulate matter (PM2.5), poses a growing threat to both the environment and public health in rapidly urbanizing regions. In Northern Vietnam, this challenge is amplified by dense population, concentrated industrial activities, and seasonal meteorological factors, highlighting the need for scalable and reliable forecasting solutions. We propose a multi-source data fusion framework that integrates ground monitoring data, meteorological variables from GFS, remote sensing data from MODIS/Terra and Sentinel-5P, and geographical indicators for PM2.5–AQI prediction and regional AQI mapping. Leveraging tree-based models (Random Forest, LightGBM, Extra Trees, Gradient Boosting) and the advanced ensemble technique Stacking, our framework outperforms traditional baselines. On the test set, it achieves an Accuracy of 0.6991 and F1-score of 0.6890, with strong balance across metrics such as Cohen's Kappa and MCC. The resulting AQI maps highlight spatial–temporal disparities, with higher pollution levels in Hanoi and lower levels in coastal and mountainous areas. This study demonstrates a practical and scalable approach to bridging monitoring gaps and supporting air quality management at the regional scale.

Thai Minh Do and *Cam-Van Thi Nguyen*

Balanced Multimodal Training through Unified Forward-Backward Modulation Strategy

ABSTRACT. Multimodal learning aims to integrate complementary information from diverse modalities such as language, vision, and audio. A persistent challenge in this field is modality imbalance, where dominant modalities suppress weaker ones during training, limiting the benefits of multimodal fusion. Existing approaches often suffer

from inaccurate estimation of modality contributions and fail to jointly consider feed-forward and backpropagation dynamics. We propose Balanced Classifier-Guided Modulation (BCGM), a general training strategy that dynamically balances modality learning across both forward and backward stages. BCGM introduces three key components: (1) Classifier-Guided Dropout, which uses discriminative scores from lightweight unimodal classifiers to guide information flow; (2) Adaptive Gradient Modulation, which scales gradients based on modality-specific learning progress; and (3) Directional Gradient Alignment, which aligns fusion gradients with unimodal signals to preserve modality diversity. BCGM is a plug-and-play module compatible with diverse multimodal architectures. Experiments demonstrate consistent improvements over strong baselines, and ablation studies confirm the importance of jointly optimizing forward and backward modality learning. Code is available at: https://anonymous.4open.science/r/BCGM-7D83.

Tung Giang Le, *Hoang Viet Vu*, *Xuan Tung Nguyen*, *Van Chien Trinh* and *Won-Joo Hwang*

Vehicle routing problems via Quantum Graph Attention Network Deep Reinforcement Learning

ABSTRACT. The Vehicle Routing Problem (VRP) is a fundamental NP-hard task in intelligent transportation systems with broad applications in logistics and distribution. Deep reinforcement learning (DRL) with Graph Neural Networks (GNNs) has shown promise, yet classical models rely on large multi-layer perceptrons (MLPs) that are parameter-heavy and memory-bound. We propose a Quantum Graph Attention Network (Q-GAT) within a DRL framework, where parameterized quantum circuits (PQCs) replace conventional MLPs at critical readout stages. The hybrid model maintains the expressive capacity of graph attention encoders while reducing trainable parameters by more than 50%. Using proximal policy optimization (PPO) with greedy and stochastic decoding, experiments on VRP benchmarks show that Q-GAT achieves faster convergence and reduces routing cost by about 5% compared with classical GAT baselines. These results demonstrate the potential of PQC-enhanced GNNs as compact and effective solvers for large-scale routing and logistics optimization.

Thao Do Thi Phuong, *Cuong Van Duc*, *Tung Doan Duy* and *Hanh Nguyen Thi*

TaP-GA: A Novel Genetic Algorithm for Target-Prioritized, Orientation-Constrained, and Adaptive Coverage Optimization in Wireless Multimedia Sensor Networks

ABSTRACT. Wireless Multimedia Sensor Networks (WMSNs) are now creating a new era in target monitoring systems where each network node can record multimedia data such as video, images, and audio, but also pose greater challenges than typical Wireless Sensor Networks (WSNs) in ensuring quality of service (QoS), satisfying coverage requirements, and maintaining network lifetime. The Heterogeneous Target Coverage (HTC) problem is particularly noteworthy among these difficulties, as sensor orientation is limited to discrete directions and coverage requirements vary for each target according to its relevance. This paper addresses the HTC problem in two crucial scenarios: a fixed number of sensors and a fixed number of targets. Taking advantage of evolutionary algorithms' ability to find near-optimal solutions, we offer an improved version of the traditional Genetic Algorithm (GA), called Target-Prioritized GA (TaP-GA), with Hybrid Population Initialization, Coverage-Biased Crossover, Target-Guided Exploratory Mutation, and an adaptive selection mechanism. These enhancements allow us to extend the search space while preserving some of the superior properties. Multiple simulation scenarios reveal that the suggested strategy is more efficient and effective than previous solutions.

Hue Nguyen, *Tan Tran*, *Giang Nguyen* and *Canh Pham*

Fast Stochastic Greedy Algorithm for k-Submodular Cover Problem

ABSTRACT. We study the k-Submodular Cover (kSC) problem, a natural generalization of the classical Submodular Cover problem that arises in artificial intelligence and combinatorial optimization tasks such as influence maximization, resource allocation, and sensor placement. Existing algorithms for kSC often provide weak approximation guarantees or incur prohibitively high query complexity. To overcome these limitations, we propose a Fast Stochastic Greedy algorithm that achieves strong bicriteria approximation while substantially lowering query complexity compared to state-of-the-art methods. Our approach dramatically reduces the number of function evaluations, making it highly scalable and practical for large-scale real-world AI applications where efficiency is essential.

Van Minh An and *Trung Kien Do*

A Mathematical Model and Exact Column Generation Approach for RMSA Problem in Elastic Optical Networks

ABSTRACT. The Routing, Modulation, and Spectrum Assignment problem extends the RSA by incorporating modulation format selection to balance transmission reach and spectrum efficiency. As RMSA is NP-complete, most prior work has relied on heuristic, hybrid, or machine learning–based methods to improve scalability. This paper proposes a configuration-based Integer Linear Programming model and an exact solution framework using Column Generation approach for RMSA problem. Experimental evaluations on NSFNET and USNET topologies demonstrate that the proposed approach consistently attains near-optimal solutions within feasible runtimes for medium-scale and large-scale networks, underscoring both its effectiveness and practical applicability.

Pham Dinh Thanh

An Improved Initialization-based Evolutionary Algorithm for the Top k 2-Clubs Problem

ABSTRACT. The s-club problem has been widely applied in social network analysis and biological network studies. Among its variants, the Top k 2-Clubs problem has attracted considerable attention from the research community. The objective of this problem is to identify k large 2-clubs that maximize a combined score reflecting both their sizes and pairwise dissimilarities. This study focuses on enhancing the efficiency of the population initialization process. The proposed initialization method integrates both greedy and random strategies to achieve a balance between individual solution quality and population diversity. Experimental evaluations conducted on DIMACS benchmark datasets demonstrate that the proposed algorithm outperforms the original approach in approximately two-thirds of the test cases.

Quan Nguyen Dang, *Thao Nguyen Thi Phuong* and *Diep Hoang Thi*

Evaluating Phylogenetic and Ancestral Recombination Graph Approaches for Analyzing RNA Virus Recombination: A Case Study of SARS-CoV-2 in Vietnam

ABSTRACT. Detecting recombination in rapidly evolving RNA viruses presents a significant computational challenge. This paper presents a case study using SARS-CoV-2 genomes from Vietnam to compare the performance of two complementary computational approaches: (i) RIPPLES, a specialized, tree-based method designed for SARS-CoV-2, and (ii) ARG4WG, a general-purpose tool for reconstructing ancestral recombination graphs (ARGs). RIPPLES demonstrated a clear advantage, identifying a strong, high-confidence recombination signal supported by the highest parsimony gain. The inferred breakpoints were consistent with previous findings and biologically plausible, being enriched in the Spike coding region. In contrast, ARG4WG inferred an unrealistically high number of recombination events that lacked clear functional patterns, suggesting its methodology is ill-suited for the unique characteristics of

short, fast-mutating viral genomes. This case study demonstrates that specialized tools like RIPPLES are currently more reliable for SARS-CoV-2 analysis and highlights the critical need for novel or adapted ARG-based methods to accurately model the evolutionary history of RNA viruses.

Vu Liem Phan, *Tu Minh Ho*, *Duc Dang Bui*, *Tien An Nguyen* and *Phuong Thai Nguyen*

Comprehensive Assessment of SLM Performance on Vietnamese High School History Tasks

ABSTRACT. Due to their low inference costs and ease of implementation on a large scale, small language models (SLMs) are becoming an increasingly viable choice in the field of education. While SLMs have been validated and benchmarked comprehensively on subjects such as Mathematics, Natural Sciences, and Foreign Languages, when it comes to Social Studies, specifically history, there does not exist a publicly available and standardized benchmark to assess the performance of SLMs. This paper introduces a benchmark, as well as a reproducible evaluation framework designed for models with less than or around 9 billion parameters. Our dataset was compiled from official Vietnamese National High School history examination questions in the 5-year period 2020-2024. Attached to each question are the labels of context and difficulty; labelling was performed manually and cross-validated to ensure consistency and objectivity. Additionally, we ran tests of 25 SLMs on our dataset, evaluating their performance on Multiple Choice and Essay Question tasks in Vietnamese history, analyzing a model's overall accuracy, as well as scrutinizing a model's reasoning as to how it arrived at its final answer. This paper contributes a curated dataset, an evaluation protocol, and standardized analyses, laying the foundation for objective comparisons between models and providing valuable insights into selecting the appropriate SLMs in the context of Vietnamese education. The source code and illustrative charts, as well as results and key findings of our experiments can be found at https://github.com/HoTuMinh/Framework-for-evaluating-SLMs-in-History-QA.

Chien Vu Manh, *Bao Anh Tran*, *Phuong Ngo Viet*, *Luan Le Chi*, *Anh Nguyen Quang*, *Long Nguyen* and *Anh Nguyen Duc*

An Empirical Study of Multi-Agent RAG for Real-World University Admissions Counseling

ABSTRACT. Large Language Models (LLMs) are increasingly considered for educational counseling, yet most existing efforts remain limited to prototypes or synthetic benchmarks, leaving little evidence from real-world deployments. We present MARAUS (Multi-Agent and Retrieval- Augmented University Admission System), a domain-specific conversational platform designed for higher-education admissions in Vietnam. MARAUS combines hybrid retrieval, multi-agent orchestration, and LLM based response generation into a lightweight and practical system. We conducted a two-phase study encompassing both development and live deployment. During two weeks of operation, MARAUS processed over 6,000 authentic user queries across six admission-related categories, achieving 87–94% accuracy with mean response times below 4 seconds. The system also proved highly cost-efficient, incurring only USD 11.58 using GPT-4o mini. Our findings provide rare empirical evidence on deploying agentic RAG systems in low-resource educational contexts and offer design insights for building trustworthy, scalable, and domain-adaptive advisory services.

Hai Nguyen, *Binh Mai*, *Hieu Dao*, *Hanh Hoang* and *Cong Tran*

Synthesizing Cultural Heritage: An End-to-End System for Designing Jewelry with Vietnamese Hue Imperial Motifs

ABSTRACT. The preservation and modernization of cultural heritage in the digital age pose multifaceted challenges. Key difficulties include: (1) the risk of cultural dilution

when traditional motifs are adapted into modern forms, (2) the absence of computational frameworks capable of integrating visual and textual modalities for design synthesis, and (3) the lack of standardized benchmarks for evaluating cultural fidelity in AI- generated artifacts. To address these challenges, this paper introduces an end-to-end diffusion-based framework that leverages generative artificial intelligence for the creative reinterpretation of Vietnamese cultural arti- facts. Our system accepts a content image, a style image, and a textual prompt to generate unique jewelry designs that harmonize contempo- rary aesthetics with the rich artistic heritage of Hue imperial motifs. As a foundational contribution, we introduce the HueJewelry-500 dataset, which consists of approximately 500 labeled jewelry images and 200 au- thentic Hue motifs, enabling reproducible evaluation in this emerging domain. Quantitative and qualitative assessments validate the effective- ness of our approach: the framework achieves a CLIP score of 0.28, out- performing contemporary baselines, while a user study with 25 cultural and design experts yielded the highest overall rating of 3.98/5 for quality and cultural authenticity. These results demonstrate that our approach not only supports digital preservation but also facilitates modern rein- terpretation of heritage, bridging historical artistry with contemporary design paradigms.

Thang Ta

Self-training from Self-memory in Data-to-text Generation

ABSTRACT. This paper introduces a novel training model, self-training from self-memory (STSM) in data-to-text generation (DTG), allowing the model to self-train on subsets, including self-memory as outputs inferred directly from the trained models and/or the new data. The quality of self-memory is validated by two models, data-to-text (D2T) and text-to-data (T2D), by two pre-defined conditions: (1) the appearance of all source values in the outputs of the D2T model and (2) the ability to convert back to source data in the outputs in the T2D model. We utilize a greedy algorithm to generate shorter D2T outputs if they contain all source values. Subsequently, we use the T2D model to confirm that these outputs can capture input relationships by demonstrating their capacity to convert text back into data. With 30% of the dataset, we can train the D2T model with a competitive performance compared to full training in the same setup. We experiment with our model on two datasets, E2E NLG and DART. STSM offers the D2T model a generalization capability from its subset memory while reducing training data volume. Ultimately, we anticipate that this paper will contribute to continual learning solutions that adapt to new training data, incorporating it as a form of self-memory in DTG tasks. Our repo is publicly available at https://github.com/hoangthangta/STSM.

Phat Tran and *Long Nguyen*

Vietnamese-guided Post-OCR Processing for Historical Nom Scripts

ABSTRACT. Despite its historical prevalence as a writing system in Vietnam, the majority of Nom documents remain inaccessible due to the obsolescence of the language. While optical character recognition (OCR) offers a pathway to digitize Nom documents, its performance is severely constrained by the limited availability of Nom annotations for training. To circumvent this, this paper proposes a post-OCR processing framework that exploits existing Vietnamese transcriptions, which are often available even when parallel Nom texts are not. Specifically, we propose the Vietnamese-guided Residual Connections (VRC) module, a module that enriches Nom encoder states by attending to contextualized Vietnamese representations, enabling the model to recover from noisy OCR predictions while avoiding overfitting to errors. In the absence of established benchmarks, we also construct a novel corpus through automatic data generation to evaluate post-OCR processing methods. Experimental results demonstrate that even a simple Seq2Seq equipped with VRC

achieves state-of-the-art performance, yielding up to 38% absolute gains in correction F1 and up to 11% in detection F1 over baselines, substantially improving both error detection and correction.

Phuong Nam Nguyen, *Gia Phuc Nguyen*, *Tung Le* and *Huy Tien Nguyen*

SelfCheckHybrid: A Hybrid Framework for Hallucination Detection in Vietnamese Large Language Models

ABSTRACT. Large language models (LLMs) demonstrate strong text generation ability but remain vulnerable to hallucinations, producing responses that are fluent and coherent yet factually incorrect. Although hallucination detection has been studied extensively in English, there is still no systematic approach for Vietnamese. As a low-resource language, Vietnamese lacks annotated datasets and specialized NLP tools, which limits the direct transfer of existing techniques. We address this challenge by adapting SelfCheckGPT, a black-box and zero-resource hallucination detection method, to Vietnamese. We explore multiple variants, including BERTScore, N-gram, Multiple-choice Question Answering and Generation, Natural Language Inference (NLI), and LLM Prompting. Building on these, we propose SelfCheckHybrid, a novel combination of NLI and prompting strategies. SelfCheckHybrid achieves accuracy comparable to SelfCheckPrompt while reducing computational cost by 44 percent, making it more efficient for practical use. Furthermore, we introduce the Vietnamese Hallucination Dataset, which consists of 210 manually annotated sentences with sentence-level hallucination labels. This benchmark represents the first resource for hallucination detection in Vietnamese. Our study provides both methodological advances and a new dataset, offering a foundation for reliable hallucination detection in Vietnamese and other low-resource languages.

Lam Nguyen Thi Khanh, *Tram Nguyen Thi Ngoc*, *Thao Tran Duy*, *Khiet Nguyen Van* and *Huy Nguyen Duc*

R2E - Requirements-to-Execution System

ABSTRACT. Requirements analysis and task planning in software engineering often demand substantial time and effort. Recent advances in Large Language Models (LLMs), combined with Agentic AI, enable these processes to be performed more efficiently. In this work, we develop a system that generates tasks list from requirements. Based on Scrum framework, our system accepts project information as input and breaks down into user stories, and ultimately creates an optimized task schedule visualized by Gantt chart. Experiments conducted on real-world projects across companies demonstrate that the generated outputs align with realistic project plans by up to 80\% in terms of features and task content. These results highlight the effectiveness of the system in supporting task planning and project management, contributing to cost reduction and improved overall performance. This work illustrates the potential of LLMs in improving task scheduling and project management practices.

Mac Thi Quynh Nhu, *Nguyen The Hung*, *Nguyen Le Minh* and *Bui Thu Lam*

P-PQGC: A Proposed Post Quantization Gain Control for Offline and Streaming Whisper under Different Speaker-to-Microphone Distances

ABSTRACT. OpenAI's offline Whisper and its streaming architectures provide robust automatic speech recognition (ASR), which is crucial for real-time communication. However, in single-microphone meeting rooms, increasing the speaker-to-microphone distance (SMD) might lead to a lower-quality speech-to-text model. This paper proposes a Post Quantization Gain Control (PQGC) to address the challenge, focusing on adjusting the signal's PCM amplitude, since factors such as noise, reverberation, and distance-related attenuation, if not properly handled, might distort the waveform and ultimately degrade the ASR performance. Furthermore, the research provides an effective architecture to integrate our PQGC into the streaming

Whisper. By ensuring that every homogeneous segment is normalized separately before being combined, this proposed PQGC streaming approach maintains local consistency and enhances recognition performance under non-uniform SDM situations. Our proposed method consistently outperforms the RMS-based approach, achieving its best performance at a peak value of 0.8 with significant improvements for distant speech. When evaluated on the 100-hour AMI Meeting Corpus benchmark and our self-gathered Vietnamese datasets in both offline and streaming modes, it achieves WER reductions of up to 2.5%, demonstrating stable and consistent gains across conditions.

Ngoc-Tri Nguyen-Dinh, *Truong-Tho Le* and *Khanh-Duy Le*

A No-Code Solution for Creating AR Indoor Navigation Applications

ABSTRACT. Indoor navigation in complex environments such as hospitals, malls, and university campuses is challenging due to complicated layouts and unclear signage. This paper presents an Augmented Reality (AR) navigation system using existing visual features as landmarks, requiring no additional infrastructure. Developed with Unity and Vuforia, it generates AR maps from floor plans and provides desktop and Android apps for map creation and navigation. The system offers prompt response times and reduces implementation costs while preserving interior design, making it suitable for healthcare, education, and retail environments. Future work will address dynamic indoor changes and advanced AR integration.

Thai Nguyen, *Trang Nguyen* and *Thanh Vi*

Fast and Lightweight CNN Model for EEG Person Identification on Constrained Hardware

ABSTRACT. Electroencephalography (EEG)-based person identification has emerged as a secure and spoof-resistant biometric solution, leveraging the uniqueness of individual brain activity. While convolutional neural networks (CNNs) have demonstrated strong performance in this domain, existing models often require large numbers of parameters and high-dimensional inputs due to reliance on many EEG channels, leading to excessive computational demands. Recent works have attempted to introduce lightweight models by reducing the number of EEG channels and network parameters. Yet, these approaches still generate excessive amounts of input data, which significantly slows down training and increases computational cost. This undermines the core objective of lightweight modeling by shifting the computational burden to the data pipeline, making the overall system far from efficient. We introduce a CNN-based EEG identification model that is lightweight in every aspect from architecture to data pipeline. The model uses only three EEG channels and 3-second input segments, keeping data compact and efficient. With approximately 64,000 learnable parameters—making it, to the best of our knowledge, the most lightweight CNN-based EEG identification model reported—it trains rapidly, requiring only ~2 seconds per epoch on a modest CPU. Despite its simplicity, the model achieves a 98.2\% rank-1 test accuracy. This balance of accuracy, speed, and ultra-low complexity makes the model especially suitable for small-scale applications, portable EEG devices, and scenarios where developers lack access to high-end GPUs or CPUs.

Dinh-Thuan Duong-Le, *Duy-Nam Ly*, *My-Le Duong Thi*, *Khanh-Duy Le* and *Minh Phuong Pham*

SolARG: A Collaborative Tangible Augmented Reality Game for Learning Gravity and Solar System Planets

ABSTRACT. Teaching children (aged 6–12) about gravity and the relative sizes of planets in the solar system poses a well-known challenge: these concepts are highly abstract, invisible in daily life, and difficult to grasp through traditional instructional methods. To address this, we designed and implemented SolARG, a collaborative

augmented reality (AR) multiplayer game that transforms these abstract principles into interactive, embodied experiences. In SolARG, two players work together in a competitive mission to stabilize an asteroid by selecting planets on opposite sides that exert equivalent gravitational forces. Because distances between the asteroid and planets differ, players must reason about both mass (represented by planet size) and distance to achieve balance while avoiding planetary collisions. We deployed the game with 60 students aged 6–12 in real-world classroom settings. Results show that SolARG significantly enhanced engagement, supported rapid understanding of gravitational balance, and improved long-term retention of knowledge about gravity and planetary sizes. These findings suggest that collaborative AR games can make abstract STEM concepts more tangible, accessible, and memorable for young learners.

Duy-Nam Ly, *Minh-Triet Tran* and *Khanh-Duy Le*

CAMIronment: Supporting Environmental Design Prototyping With Generative AI and Context-Aware Multimodal Interaction

ABSTRACT. Extended Reality (XR) technologies are increasingly expanding the opportunities of environmental design prototyping (e.g., interior design or MR classroom environments) by enabling users to create and manipulate virtual objects within immersive environments. These capabilities support rapid externalization of ideas, exploration of spatial configurations, and in-situ evaluation of experiential qualities. In contrast, conventional design approaches—such as sketches and CAD-based 3D modeling—are constrained by steep expertise requirements and lengthy iteration cycles. Recent advances in Generative AI present promising alternatives, particularly through text-to-3D pipelines that generate diverse assets from natural language prompts. Yet, systems that rely solely on textual input place significant cognitive and descriptive demands on users and often yield outputs misaligned with intended spatial or contextual requirements. To overcome these limitations, we introduce CAMironment, a context-aware multimodal interaction system that combines voice, gesture, and scene context as inputs for efficient 3D asset creation in environmental design prototyping. We conducted a comparative evaluation of CAMironment against a baseline text-to-3D system to examine usability and user effort. Results demonstrate that CAMironment alleviates users' descriptive burden and enables more effective integration of generated assets into immersive design workflows.

Pham Cong Tu, *Nguyen Quang Vinh*, *Tran Thuan Hoang* and *Pham Thi Phuong Hoa*

Finite-time error control combining neural networks in noisy environments and mobile targets

ABSTRACT. This paper presents an integrated guidance and control approach for a single-channel high-speed aircraft (HSA) based on finite-time error control (FTC) augmented with an adaptive neural network (NN). The FTC scheme is devised to ensure that the tracking error converges to zero within a user-specified finite time, thereby guaranteeing fast response and precise interception performance even against highly maneuvering targets. To enhance robustness, a radial basis function neural network is incorporated into the controller design to estimate and compensate unknown nonlinear dynamics and external disturbances online. The proposed control law consists of an equivalent control component and a switching component, among which the NN serves as a real-time intelligent approximation module embedded in the equivalent part. An adaptive weight update rule is developed to automatically adjust the neural parameters during operation. Lya-punov-based theoretical analysis is conducted, confirming that the closed-loop system is finite-time stable and both tracking and estimation errors remain bounded in the presence of uncertainties. Numerical simulations are carried out to evaluate the effectiveness of the proposed

FTC–NN scheme. Simulation results demonstrate clear superiority over conventional FTC and sliding mode control (SMC) methods in terms of tracking accuracy, convergence time, and disturbance rejection capability.

15:30-16:00 Tea Break

16:00-18:00 Session 5A: SOICT Technical Session IX: AI Applications, AI Foundations and Big Data

CHAIR: *Anna Scius-Bertrand*

LOCATION: Grand Ballroom A, 2F

16:00 *Hai Anh Tran*, *Huy-Hieu Nguyen*, *Quoc-Trung Le*, *Abdelhamid Mellouk* and *Truong X. Tran*

FedEABoost: A Client Entropy Adaptive Boosting Framework for Federated Learning

ABSTRACT. Federated Learning (FL) enables distributed training of machine learning models across decentralized devices without sharing raw data, thus providing privacy-preserving solutions suitable for various applications. However, FL is significantly hindered by data heterogeneity across clients, known as the non-IID problem, leading to degraded model performance, slower convergence, and challenges in achieving fairness across client models. Existing methods typically struggle to adequately handle class imbalance, particularly when certain classes are underrepresented locally. In this paper, we propose FedEABoost, a novel FL framework integrating AdaBoost training principles to address the non-IID issue effectively. FedEABoost generates multiple local models in each client by iteratively emphasizing harder to classify samples, often from minority classes and then employs an entropy-based selection mechanism to choose the most suitable local models for aggregation. Extensive evaluations conducted on Fashion-MNIST and CIFAR-10 datasets demonstrate that FedEABoost outperforms conventional FL baselines, achieving up to 25\% improvement in global model accuracy across various non-IID scenarios. Our results highlight the robustness and effectiveness of FedEABoost in mitigating the impact of data heterogeneity, promoting stable convergence, and enhancing generalization.

16:20 *Teh-Jen Sun* and *Eui-Nam Huh*

Entropy-Based Gradient Weighting and Batch-Size Adaptation for Virtual Data-Parallel Training

ABSTRACT. Distributed deep learning commonly relies on uniform gradient averaging and fixed per-node batch sizes. These choices ignore that model confidence and mini-batch difficulty vary across workers and over training, which can dilute informative updates and produce unstable progress. We provide a simple control signal derived from predictive uncertainty. For each worker we compute entropy- and margin-based measures from its predictions, normalize them to obtain a weight, and use this weight in two ways: (i) to reweight gradients during aggregation, and (ii) to adjust the next-epoch batch size per worker so that computation is allocated where it is most informative. The method is optimizer-agnostic, communication-compatible, and easy to integrate into existing code. Experiments on CIFAR-10 with ResNet-18, VGG-16, MobileNet-V2, and a lightweight CNN show that our approach consistently improves the stability of training, yields higher area under the accuracy curve (AUC), and reaches target accuracy faster or on par with baselines. These results indicate that predictive uncertainty is an effective control signal for both gradient aggregation and resource allocation in distributed training.

16:40 *Quang-Hung Bui*, *Bach Ngoc Pham*, *Anh-Minh Tran*, *Thanh Dat Le*, *The Phong Le*, *Tien Dung Nguyen* and *Anh Son Ta*

AdaFRUGAL: Adaptive Memory-Efficient Training with Dynamic Control

ABSTRACT. Training Large Language Models (LLMs) is highly memory-intensive, largely

due to optimizer state overhead. The FRUGAL framework addresses this with gradient splitting, combining a stateful optimizer on a small subspace and a memoryless one on the remainder. However, FRUGAL relies on two static hyperparameters—the subspace ratio (ρ) and update frequency (T)—which may be suboptimal across training phases. We present AdaFRUGAL, an adaptive extension that introduces (i) a linear decay schedule for ρ to progressively reduce memory usage, and (ii) a loss-aware adaptive schedule for T to lower computational overhead as convergence slows. Experiments on large-scale English (C4), Vietnamese (VietVault), and GLUE fine-tuning show that AdaFRUGAL maintains competitive perplexity and downstream performance compared to AdamW and static FRUGAL, while significantly reducing GPU memory and training time. AdaFRUGAL offers a practical, autonomous solution for efficient LLM training in resource-constrained environments.

17:00 *Wassim Ghommidh* and *Mohamed Farah*

Part-GNN: A partitioning-based graph neural network for efficient memory large scale data classification

ABSTRACT. Graph Neural Networks (GNNs) are a subset of deep learning models that have shown great promise in various forms of classification problems, including image detection and classification as well as predict- ing labels for nodes or edges. They are designed to operate on graphs and rely on graph-structured data, which may include node and/or edge attributes. Several GNN variants have been developed, such as GCN, Cluster-GCN, GraphSAGE, and FastGCN. However, GNNs also have high memory requirements, largely due to the large adjacency matrix, which becomes computationally expensive for large-scale graphs, as well as long training times. In this work, we describe a new technique for graph partitioning in order to reduce the effective size of a graph while still achieving as high a predictive F1-score as possible. Additionally, we combine the graph partitioning method with layer-wise training to achieve greater computational efficiency.

17:20 *Tan Hai Nguyen*, *Do Thanh Huyen Khong* and *Thi Quynh Hoa Nguyen*

Enhancing Survey Efficiency: A Validated Vietnamese Short-Form of the MBTI Developed Through Machine Learning

ABSTRACT. In academic and career counseling, validated personality tools like the MBTI are crucial, but standard versions with over 90 questions are time-consuming and difficult to implement in large items often sacrifice psychometric integrity by providing insufficient coverage of the four personality axes, thereby reducing reliability and classification accuracy. Compounding this, the direct application of international instruments in Vietnam risks cultural and linguistic inaccuracies, necessitating a tool that is not only parsimonious but also culturally resonant. This study addresses this gap by developing and validating an optimized MBTI-based questionnaire for Vietnamese youth. The methodology involved a rigorous language and context conversion process to ensure conceptual and psychometric equivalence, validated through qualitative evaluation by language, psychological, and cultural experts. Subsequently, a feature extraction process utilizing a Random Forest model with Recursive Feature Elimination with Cross-Validation (RFECV) was employed to identify and select items with the highest classification power. The optimized instrument was then administered to a sample of approximately 1800 participants for quantitative validation. The results were positive, with the instrument demonstrating acceptable to excellent psychometric properties and a valid factor structure confirmed by Confirmatory Factor Analysis, indicating a good model fit. The study successfully produced a culturally resonant and psychometrically robust personality assessment tool suitable for the target audience.

17:40 *Mark Jerome Santos*, *Andreas Luy*, *Kenneth Amurao* and *Adriane Brent Castro*

CITADEL: A Web-Based Faculty Performance Evaluation and Decision-Support System for Higher Education Institutions

ABSTRACT. Faculty Performance Evaluation (FPE) in higher education often depends on PDFs and spreadsheets, producing delayed consolidation, non-transparent assessments, and weak decision support. This paper presents a university-wide platform that unifies configurable Performance Indicators (PIs), distributed scoring with endorsements and governed overrides, real-time dashboards and reports, and period archiving with provenance. The architecture adopts an evidence-first, live-consolidation model in which each PI serves as both intake form and data source, enabling permission-scoped views from criterion to PI to evidence. Related work shows gains in automation but persistent fragmentation and limited analytics; the design addresses these gaps via modular services (PI management, scoring, workflow, reporting, audit), role-based access control (RBAC) with attribute-based access control (ABAC), and immutable audit events. An expert appraisal with n = 15 IT professionals used a guided walkthrough and an ISO/IEC 25010 survey. All constructs—Functional Suitability, Performance Efficiency, Usability, Reliability, Security, Maintainability, and Portability—were above neutral (95% CI lower bounds ≥ 5.05) with acceptable to excellent internal consistency. Findings indicate early fitness for purpose and support mid-cycle decision making.

18:00 *Rafael John Castro*, *Earl Gabriel Datu*, *Alex Gaebriel Limio*, *Julian Carlos Torno* and *Jenice Anne Marie Visperas*

AUF iAssist: A Web-Based Helpdesk System for Efficient Support and Concern Resolution

ABSTRACT. This study presents AUF iAssist, a centralized, web-based helpdesk for Angeles University Foundation that consolidates fragmented channels into a unified platform integrating ticket workflows, a knowledge-driven FAQ chatbot, AI-powered solution recommendations, role-based dashboards, and automated notifications. Built on a secure client–server architecture with modular services, the system enhances traceability, scalability, and policy compliance. Using a structured walkthrough with 15 IT professionals (purposive sampling), we evaluated AUF iAssist against ISO/IEC 25010. Results show strong reliability across Functional Suita-bility, Performance Efficiency, Security, and Reliability (Cronbach's α > .90), with Usability highest (M = 5.29). AUF iAssist reduces administrative workload, improves response times, and increases user satisfaction, aligning with SDG 4 and SDG 9. We outline targeted re-finements in usability, reliability, and portability for broader deployment.

16:00-18:00 Session 5 - SOICT Technical Session X: AI Applications

CHAIR: *Cong Tran*

LOCATION: Grand Ballroom B, 2F

16:00 *Tien Dat Phan*, *Vy Anh Tran*, *Bao Long Hoang*, *Thi Minh Ngoc Truong* and *Thi Hau Nguyen*

GRACE: A Knowledge Graph–Enhanced Conversational Recommendation System via Retrieval-Augmented Generation

ABSTRACT. We present GRACE (Graph-Reasoning Augmented Conversational Engine), a multi-stage conversational recommender system that integrates Large Language Models (LLMs) without costly fine-tuning by using a Retrieval-Augmented Generation (RAG) framework grounded in a knowledge graph. Unlike prior CRS approaches that rely on either shallow KG lookups or expensive fine-tuning, GRACE introduces a novel hybrid retriever that fuses three complementary strategies: (i) semantic similarity over plot embeddings, (ii) schema-aligned content filtering with LLM-extracted genres, and (iii) collaborative filtering via graph expansion along creator relationships. GRACE, evaluated on the ReDial and INSPIRED datasets, effectively

addresses the performance–cost trade-off inherent in the existing methods. Without requiring any training, the framework achieves 0.062 Recall@1 and 0.302 Recall@10 on ReDial, as well as 0.116 Recall@1 and 0.307 Recall@10 on INSPIRED, consistently outperforms traditional CRS models, zero-/few-shot LLM baselines, and state-of-the-art knowledge-enhanced approaches. These results demonstrate that tightly coupling LLM reasoning with KG grounding yields a practical, scalable foundation for high-performance conversational recommendation. Our code is publicly available at: https://github.com/DatPhan06/GRACE.

16:20 *Yukinobu Hoshino*, *Namal Rathnayake*, *Tuan Linh Dang* and *Upaka Rathnayake*

Effectiveness of Rolling-Sum Preprocessing in River Mouth Water Depth Prediction Using Machine Learning

ABSTRACT. Accurate prediction of river-mouth water levels is critical for flood preparedness in Japan's steep catchments, where short response times heighten climate-related risks. Conventional hydrologic and statistical approaches demand dense input data and extensive calibration, while machine-learning models trained solely on raw station records often overlook prior flow conditions and deliver weak forecasts. This study tests a physically informed preprocessing method—rolling sums of upstream water depths—within the Niyodo River system. Daily observations from 14 upstream stations were used to compare models fed with raw data against those using rolling-sum features, spanning linear (Lasso, ElasticNet, Ridge) and ensemble (CatBoost, Extra Trees, LightGBM, XGBoost) algorithms under a chronological split. Incorporating rolling sums substantially increased predictive accuracy, raising the coefficient of determination from below 0.25 to around 0.65 and cutting root-mean-square error by more than 10 cm. Feature-importance analyses identified main-stem stations as key predictors, aligning with hydrological expectations. These results show that lightweight, domain-guided preprocessing can significantly enhance water-level forecasting in flood-prone basins.

16:40 *Tuan-Ngoc Nguyen*, *Hai-Dang Kieu* and *Cam-Van Thi Nguyen*

Enhance Sequential Recommendation via Linear Recurrent Units

ABSTRACT. User interactions on online platforms naturally form sequences, making sequential recommendation essential for capturing user's preferences. Self-attention-based models have recently driven progress in this area, but their high computational cost and latency hinder real-time use. In contrast, lightweight models are efficient but often overlook fine-grained item order, which reduces accuracy. To address this gap, we propose PLRec, a sequential recommendation model based on efficiency linear recurrent architectures combined with a dual-task learning framework. PLRec employs linear recurrent units with a recursive parallelization strategy to enable incremental updates. By extending linear recurrent units with a recursive parallelization strategy, PLRec enables incremental updates while preserving sequential order information. The dual-task design incorporates category prediction as an auxiliary objective, which provides complementary supervision and strengthens preference modeling. Extensive experiments on multiple real-world datasets show that PLRec achieves state-of-the-art accuracy with significantly better efficiency on long user sequences, highlighting its potential for real-time and scalable personalized recommendation. Our code is publicly available at this site.

17:00 *Dang Nhat Khuong*, *Nguyen Chi Kien*, *Truong Vinh Linh*, *Cao-Phan Khanh-Duy* and *Pham Minh Tuan*

Aspect-Based Sentiment Analysis for Stock Price Movement Prediction

ABSTRACT. Forecasting short-term stock price movements remains a central challenge in financial prediction, particularly in emerging markets where volatility and data asymmetry complicate traditional approaches. This paper presents an integrated framework that combines Aspect-Based Sentiment Analysis (ABSA), fundamental

indicators, and technical signals to improve market forecasting accuracy and interpretability. Using GPT-4o, we extract fine-grained sentiment on 33 finance-specific aspects from over 16,000 Vietnamese financial news articles and corporate disclosures, then merge these structured signals with technical indicators and firm-level fundamentals for model training. Experiments on 30 VN30 equities (2019–2024) show that gradient boosting models (notably XGBoost) achieve superior performance, with AUC up to 0.67, surpassing technical- and fundamental-only baselines. SHAP analyses highlight that ABSA-derived features enhance predictive stability, especially during neutral market conditions. Backtesting further demonstrates that probability-thresholded trading strategies deliver annualized returns exceeding both buy-and-hold and deep learning baselines. These findings underscore the value of disentangling what the market discusses from how it is discussed, and illustrate how multi-source fusion with modern NLP can provide scalable and explainable decision support for data-driven trading in emerging markets.

17:20 *Abir Linoubli*, *Khairi Abidi* and *Mohamed Farah*

Tokenization in Protein Language Models: Methods, Taxonomy, and Applications

ABSTRACT. The application of natural language processing (NLP) to biological sequences is reshaping computational biology, particularly protein analysis. At the center of this paradigm lies tokenization the process of segmenting protein sequences into discrete units for language models. Unlike human language, protein lacks explicit delimiters, making tokenization both a critical and challenging design choice, with direct consequences for downstream performance. In this survey, we propose a systematic taxonomy of protein sequences tokenization methods and analyze their trade-offs in terms of biological interpretability, computational efficiency, and predictive performance. We also review evaluation metrics, summarize applications across a range of protein analysis tasks, and examine the interpretability of learned protein tokens. Our findings show that no single tokenization strategy dominates across all contexts; rather, the optimal choice depends on the biological objectives and computational constraints at hand.

16:00-18:00 Session 5C: SOICT Technical Session XI:
Applied Operations Research and Optimization
CHAIR: *Chi-Thanh Vi*
LOCATION: Yersion Ballroom A, 2F

16:00 *Quoc-Trung Bui*, *Minh Phan*, *Duy Vu Nguyen*, *Van Son Nguyen* and *Quang Dung Pham*

Non-Parametric Feature Combination For Explainable Credit Scoring

ABSTRACT. Credit scoring has become a cornerstone of modern financial risk assessment, enabling lenders to evaluate borrowers' creditworthiness with unprecedented precision. By transforming complex borrower data into standardized numerical scores, credit scoring predicts the likelihood of loan repayment, thereby reshaping lending practices across diverse industries. Recently, machine learning models have attracted scholars and stakeholders as their accuracy surpasses that of traditional models based on logistic regression and decision trees, which are simple, explainable, and adaptable. However, machine learning models lack interpretability, which leads to a barrier in their real-life implementation. In traditional model development, feature selection is widely utilized to eliminate low-discriminative features, while feature combination, which creates new features by integrating existing ones to capture latent relationships, remains underexplored in credit scoring field. This paper proposes a novel non-parametric approach to feature combination, designed to maximize the use of available features' information. The experimental results on benchmark datasets demonstrate the potential of this approach in enhancing the accuracy of credit scoring models, as the model using the combined features achieves the highest accuracy among the experimental models across all the datasets.

16:20 *Uyen Tran*, *Tan Tran* and *Canh Pham*

Deterministic one-pass streaming algorithm for non-monotone DR-submodular maximization under a size constraint

ABSTRACT. Submodular optimization on the integer lattice has recently attracted significant attention due to its ability to capture problems involving multiple occurrences of elements, with diverse applications in influence maximization, budget allocation, and data summarization. In this paper, we study the problem of maximizing a non-monotone DR-submodular function under a size constraint in a streaming setting, denoted as $\DrSMC$. We propose the first deterministic one-pass streaming algorithms designed for this problem that achieve a theoretical approximation guarantee of $\frac{1}{6}-\epsilon$, with query complexity $O\left(\tfrac{n \log^2 K}{\epsilon}\right)$ and space complexity $O\left(\tfrac{K \log K}{\epsilon}\right)$. Extensive experiments on the Revenue Maximization benchmarks demonstrate that our proposed methods consistently outperform existing baselines in terms of solution quality, query efficiency, and memory usage. These results establish the practicality and robustness of our approach for large-scale streaming applications.

16:40 *Luan Thach*, *Phuong Do*, *Khoa Tan Vo*, *Thu Nguyen*, *Mong-Thy Nguyen-Thi* and *Tu-Anh Nguyen-Hoang*

DESW: Reducing Concentration in Proof-of-Stake with Dynamic Exponential Stake Weighting

ABSTRACT. In permissionless blockchains, Proof-of-Stake (PoS) selects validators in proportion to staked assets, securing the ledger by aligning incentives. However, this proportionality amplifies advantages for large validators through rewards, delegation, and liquid staking, leading to stake concentration and threatening decentralization. To address this issue, we introduce the Dynamic Exponential Stake Weighting (DESW) model, grounded in the principle of weighted probability distribution and equilibrium theory. DESW defines an adaptive validator selection rule that rebalances stake weights according to network-wide inequality, measured by the Gini coefficient. By employing weights in the value domain, DESW reduces the influence of oversized validators as centralization increases, without altering rewards, penalties, or protocol flow. Thus, DESW directly addresses the “rich-get-richer” dynamic in PoS and enhances decentralization. We evaluate DESW across three scenarios that are representative of public PoS networks: (i) stable distributions, (ii) large high-turnover systems, and (iii) adversarial stake injections. Relative to baseline PoS, DESW reduces the Gini coefficient from 0.623 to 0.318 (-48.9%), from 0.876 to 0.501 (-42.8%), and from 0.945 to 0.462 (-52.1%), respectively. Simultaneously, it increases the Nakamoto coefficient from 134 to 303 (+126.1%), from 384 to 2138 (+456.8%), and from 9 to 196 (+2,077.8%), respectively. When applied to Ethereum’s validator set, DESW would raise the Nakamoto coefficient from 3 to 391 (about 130×) and lower Gini from 0.99 to 0.10. These results indicate that DESW strengthens decentralization and fairness while preserving PoS efficiency, offering a practical drop-in improvement for networks.

17:00 *Bui Trong Duc*, *Ho Viet Duc Luong*, *Bui Xuan Son*, *Dang Quang Thang*, *Tran Van Tam*, *Nguyen Hai Dang* and *Vu Van Tan*

Balancing Efficiency and Fairness in the Integrated Truck–Drone Dispatching Problem with Dynamic Endurance via Pareto Front Grid Guided Multi-objective Optimization

ABSTRACT. The coordination of trucks and drones in logistics systems offers new opportunities for cost-effective and sustainable logistics operations while enhancing customer satisfaction, yet significant challenges arise from heterogeneous vehicle capabilities and the payload-dependent endurance of drones. This paper introduces Integrated Truck–Drone Dispatching Problem with Dynamic Endurance (ITDDPDE), a novel multi-objective formulation that jointly schedules a fleet of trucks and drones, capturing two

realistic features of drone operations: (i) dynamic endurance that depends on the payload carried, and (ii) flexible launch and retrieval points along truck routes. The problem simultaneously minimizes three objectives: total operational cost, fairness across customers, and fairness among vehicles. To tackle the ITDDPDE problem, we propose a multi-objective evolutionary algorithm based on Pareto front grid decomposition, incorporating two-level encoding and a heuristic mechanism to determine drone launch and landing points. Numerical experiments on diverse benchmark scenarios demonstrate that the proposed algorithm significantly outperforms state-of-the-art multi-objective algorithms in terms of Hypervolume and Inverted Generational Distance.

17:20 *Dung T.K. Ha*, *Anh T.N. Vu*, *Linh K. Duong* and *Phong T.D. Nguyen*

Budgeted Object Detection via Online Submodular Approximation Algorithm

ABSTRACT. We propose two online algorithms, $\ST$ and $\IST$, for object detection via a reduction to the Submodular Subset Selection problem under a budget constraint. $\ST$ provides a baseline framework, while $\IST$ improves it with an approximation ratio of $1/2$ using only $5n$ queries and $\mathcal{O}(n)$ time. Our methods combine attribution techniques with online approximation to handle streaming inputs efficiently, achieving near-Greedy quality with significant runtime and memory savings. Theoretical guarantees and analysis confirm their practicality for large-scale or real-time scenarios.

16:00-18:00 Session 5D: SOICT Technical Session XII: Networking and Communication Technologies, Software Engineering

CHAIR: *Thien Huynh-The*

LOCATION: Yersin Ballroom B, 2F

16:00 *Thien Huynh-The*, *Minh-Thanh Le*, *Ngoc-Ha Truong*, *Van-Ca Phan* and *Truong-Thinh Le*

A Lightweight and Robust Framework for Waveform Classification Using Dynamic Warping and State-Space Models

ABSTRACT. Accurate and efficient waveform classification is a critical challenge in dense radio frequency environments. While deep learning offers solutions, many models struggle to balance high performance with computational constraints. This paper introduces the Warping State Space Model (WarpSSM), a lightweight and robust framework designed to address this trade-off. Our approach processes spectrograms using two novel components: the Dynamic Adaptive Warp block to mitigate channel-induced geometric distortions, and the Visual State Space Block with Directional Selective Fusion to capture long-range dependencies. Evaluated on a synthetic dataset of 12 waveforms under realistic channel degradation, WarpSSM achieves a state-of-the-art average accuracy of 90.61%. It demonstrates exceptional efficiency, with an inference latency of 0.55 ms from a model of only 42.2K parameters.

16:20 *Tien H. Do*, *Thang V. Nguyen*, *Kien T. Phan*, *Hien T. T. Pham* and *Ngoc T. Dang*

Channel-Aware Power and Rate Control for UOWC with DRL and HARQ Integration

ABSTRACT. Underwater optical wireless communication (UOWC) has difficulties in fulfilling ultra-reliable low-latency communication (URLLC) standards owing to channel distortions, including absorption, scattering, and oceanic turbulence. This paper presents a deep reinforcement learning (DRL) approach utilizing proximal policy optimization (PPO) to concurrently adjust transmit power and coding rate in a point-to-point UOWC system employing hybrid automatic repeat request (HARQ) protocols chase combining (CC-HARQ) and incremental redundancy (IR-HARQ) capitalizing on statistical channel information and signal-to-noise ratio feedback, structured as a Markov decision process (MDP) with rewards that penalize power consumption and delay infractions. By reducing the long-term average power while adhering to stringent delay constraints (e.g., 99.9% dependability at 13 dBm in pristine marine conditions), the approach enables energy-efficient and dependable UOWC for Beyond 5G (B5G)

and 6G applications, such as ocean monitoring.

16:40 *Trung Vu-Thanh*, *Jing He*, *Xuan Quang Truong*, *Nga Phan Thi Thanh*, *Luong Nguyen Thi*, *Ninh Duong-Bao* and *Khanh Nguyen-Huu*

Threshold-based AP Filtering and Distance Measure Analysis for K-means Clustering in WiFi Fingerprinting-based Indoor Localization System

ABSTRACT. WiFi Fingerprinting is widely used in indoor localization due to its cost-effectiveness. However, scalability remains a challenging problem. For large WiFi fingerprint datasets, clustering reduces the computational burden by limiting the search space during matching. K-means is a commonly utilized clustering algorithm valued for its speed and simplicity. This paper examines the effectiveness of the threshold-based access point (AP) filtering, which aims to choose valuable APs, and six different distance measures for k-means clustering, which can optimize the clustering process. To the best of our knowledge, this is the first time various distance measures are considered in k-means clustering in WiFi Fingerprinting-based localization systems to evaluate their effectiveness. We conduct experiments on five public datasets with three AP filtering strategies. The experimental results show that the AP filtering significantly affects the localization performance. Using the Søren-sen distance measure, compared to the original K-Nearest Neighbors meth-od, k-means shows less than a 2% increase in average localization error while achieving over a 98% reduction in average computation time. The analysis shows the scalability potential of the k-means algorithm and its effectiveness in improving the WiFi Fingerprinting technique.

17:00 *Nhat-Hoa Tran* and *Quang-Huy Vuong*

A Bounded Model Checking Approach for Verifying OSEK/VDX Applications

ABSTRACT. We propose a bounded model checking technique for verifying applications running on the OSEK/VDX operating system (OS). Our approach encodes task behaviors into logical formulas by exploring the control-flow graphs of tasks in the application, following the scheduling policy of the OSEK/VDX OS, and the API function calls by unrolling the program with a bound. The formula is then solved using an SAT/SMT solver to detect potential bugs in the applications (within the specified bounds). The method is implemented as an extension of the CBMC model checker. Several experiments were conducted to demonstrate the accuracy and performance of our method.

17:20 *Taejun Choi*, *Boyoon Kim* and *Hichan Moon*

UAV-Based Target Terminal Search System for Emergency Rescue

ABSTRACT. It is very difficult to obtain accurate location information of a target terminal owned by a man-in-distress. Especially, if a target terminal is a legacy mobile phone or located in a GPS (Global Positioning System)-denied area, it is generally possible to obtain the cell information where a target terminal is located. In this case, search time for a target terminal becomes too long to search a man-in-distress within golden time. In this paper, a novel search system is presented to find the accurate location of a target terminal owned by a man-in-distress. In the proposed system, a UAV (Unmanned Aerial Vehicle) is equipped with SME (Signal Measurement Equipment), which measures the signal transmitted from a target terminal. UAV navigates and changes its direction based on the AoA (Angle of Arrival) of the measured signal. To evaluate the performance of the proposed system, a simulator was built. The simulation results show that accurate location information can be obtained within golden time in most cases.

17:40-18:00 Session 6A: SOICT Meetup

17:40-18:00 Session 6B: Competition Session

CHAIR: *Hung Son Nguyen*

LOCATION: Yersin Ballroom A, 2F

DAY 2 - SATURDAY, 13 DECEMBER 2025

08:00-17:40 Registration

08:30-09:10 Session 7: Keynote III:
Josiah Poon (University of Sydney, Australia)
CHAIR: *Cathal Gurrin*
LOCATION: Grand Ballroom, 2F

09:10-09:50 Session 8: Keynote IV:
Tung Kum Hoe Anthony (National University of Singapore, Singapore)
CHAIR: *Wray Buntine*
LOCATION: Grand Ballroom, 2F

09:50-10:20 Tea Break

10:20-12:00 Session 9A: SOICT Technical Session XIII: Lifelog Event Retrieval
CHAIR: *Cathal Gurrin*
LOCATION: Grand Ballroom A, 2F

10:20 *Trong-Le Do*, *Viet-Tham Huynh*, *Hai-Dang Nguyen*, *Thuc Nguyen-Quang*, *Mai-Khiem Tran*, *Trong-Thuan Nguyen*, *Tu V. Ninh*, *Tu-Khiem Le*, *Thanh Duc Ngo*, *Duc-Tien Dang-Nguyen*, *Tu-Trinh Ngo*, *Klaus Schoffmann*, *Cathal Gurrin* and *Minh-Triet Tran*

Toward Abstraction-Level Event Retrieval in Large Video Collections: Leveraging Human Knowledge and LLM-Based Reasoning in the Ho Chi Minh City AI Challenge 2025

ABSTRACT. The Ho Chi Minh City AI Challenge 2025 focused on advancing abstraction-level event retrieval from large, diverse video collections. Key tasks included Known-Item Search (KIS), Question Answering (Q&A), and the demanding Temporal Retrieval and Alignment of Key Events (TRAKE), requiring precise temporal coherence. Solutions integrated robust multimodal architectures (VLM, OCR, ASR) with Large Language Models (LLMs) for semantic reasoning and query refinement. Critical advancements in temporal modeling, often using Dynamic Programming, ensured event sequence coherence. This collection provides 48 papers from participating teams, demonstrating significant progress in scalable and context-aware retrieval systems.

10:40 *Thanh Nhan Vo*, *Minh Doan Ngoc Binh*, *Pi-Dieu Sam*, *Hai Dang Nguyen* and *Khoi-Nguyen Nguyen*

Real-Time Hybrid Multimodal Retrieval System for AI Challenge HCMC 2025

ABSTRACT. This paper presents our team's system for the AI Challenge 2025 (AIC 2025), addressing three video retrieval tasks: Known-Item Search (KIS), Question Answering (Q&A), and Tracking-Based Known Event (TRAKE). Building on our previous work [1], we redesigned the real-time framework to improve semantic precision, temporal consistency, and multi-user scalability. The system integrates a hybrid CLIP-SigLIP2 embedding engine, context-aware reranking, and temporal fusion, deployed on an asynchronous FastAPI-Redis-Milvus stack for low-latency processing. Tested on the official AIC 2025 dataset, it achieved 9% higher Recall@10 over single-model baselines with < 180ms average latency, ranking first in the High-School Division. These results demonstrate that combining hybrid multimodal embeddings with contextual post-processing yields a robust, real-time solution for large-scale video retrieval.

11:00 *Nguyen Mai Vinh*, *Khoi Nguyen Nam Anh*, *Duy Tran Khanh*, *Duy Do Quoc*, *Hung Bach Chan*, *Bao Tran Gia*, *Minh Vo*, *Tien Do* and *Thanh Duc Ngo*

Towards Conversational Video Retrieval with an Intelligent Search Agent

ABSTRACT. Advances in technology have led to an explosive growth of multimedia content, especially videos. This creates an increasing demand for reliable video event

retrieval frameworks capable of efficiently extracting meaningful events from large-scale video databases. Although recent advances have enabled multimodal and temporal matching, most video retrieval systems still rely on rigid query interfaces and lack the ability to engage in conversational refinement. This limitation arises from the absence of an intelligent agent capable of understanding natural language, maintaining context, and dynamically linking user intent to the retrieval modules. To overcome this limitation, we propose a video event retrieval system that incorporates an intelligent search agent, enabling seamless, conversation-guided video retrieval. The agent engages users in dialogue to interpret intent, answer content-related questions, and iteratively refine search results. It leverages existing multimodal and temporal retrieval modules to ground responses in both textual and visual evidence, while a lightweight plug-and-play feedback fusion allows composed retrieval without retraining. Through this interactive loop, users can explore large-scale video collections in a natural and effective manner. Evaluation in the 2025 Ho Chi Minh AI City Challenge demonstrates the effectiveness of our approach, achieving 85 out of 88 successful retrievals and highlighting the potential of conversational video retrieval.

11:20 *Minh-Quan Ho-Le*, *Duy-Khang Ho*, *Huy-Hoang Do-Huu*, *Nhut-Thanh Le Hinh*, *Hoa-Vien Vo-Hoang*, *Tu V. Ninh* and *Minh-Triet Tran*

Applying Large Language Model (LLM) Agents for Automated Lifelog Retrieval

ABSTRACT. Recent advances in LLMs have enabled new levels of autonomy in complex multimedia retrieval systems. This work introduces an LLM-based Planning Agent for fully automated lifelog retrieval, capable of interpreting user queries and orchestrating multimodal retrieval components without human intervention. Given a query and a registry of available retrieval modules (e.g., by text embedding, by OCR, by activity labels, and by object tags), the agent generates and executes complete retrieval plans, selecting appropriate modalities, configuring parameters, and fusing results through normalization and ranking strategies. The system exhibits agentic behavior by reasoning about task structure, adapting execution flow dynamically, and self-evaluating retrieval outcomes to refine its strategy. This fully autonomous mode demonstrates how LLM-driven agents can transform lifelog retrieval from an expert-guided, interactive process into an intelligent, self-directed search paradigm, which bridges the gap between human expertise and machine understanding.

11:40 *An To Vinh*, *Nguyen Nguyen Hoang*, *Nguyen Luong Si*, *Tho Le Doan*, *Minh Vo*, *Tien Do* and *Thanh Duc Ngo*

Leveraging Composed Image Retrieval Principles for Efficient Textual Feedback in Multimodal Retrieval

ABSTRACT. People watch countless videos every day for learning, entertainment, or work, and as video content continues to grow, finding a specific clip or moment among the vast collections has become increasingly difficult. This growing challenge highlights the need for more effective video retrieval systems that can help users quickly locate the videos they are looking for. However, most existing systems produce static results that do not always align with user intent, compelling users to reformulate their queries multiple times while the absence of interactive refinement mechanisms makes it difficult to achieve smooth, conversational search experiences. To address this challenge, we propose a feedback-driven textual search system that enables users to provide direct feedback on retrieved images, iteratively refining retrieval results until the desired outcomes are achieved. Through this approach, our system achieved a score of 84.6 on a maximum scale of 88 in the preliminary round of the HCM AI Challenge 2025, demonstrating strong real-world performance.

12:00 *Xuan Huy Manh*, *Anh Hao Kieu*, *Minh Hung Le*, *Duy Tung Nguyen*, *Thanh Tung Nguyen*, *Viet Hang Dao* and *Hai Vu*

Estimating size of lesions in Endoscopic Images using depth model-based approaches

ABSTRACT. In clinical environments, especially during minimally invasive procedures such as endoscopy, accurate measurement of lesions such as polyps or tumors plays a vital role in determining disease severity and choosing appropriate treatment methods. However, endoscopic cameras typically provide only two-dimensional RGB images that lack depth information, making the size estimation of internal organs or lesions highly challenging. To address this issue, this study utilizes the ZoeDepth model, a deep learning model capable of estimating depth from a single RGB image. However a depth estimation model such as ZoeDepth has not trained for endoscopic images, this study handles this obstacle to solve the problem of estimating the size of the lesion from endoscopic cameras. The proposed method is based on a set-up with chessboards in controlled condition. It combines a series of steps that include geometric correction, depth value optimization, and measurement error assessment. The evaluation of the system was performed on multiple datasets, including synthetic and real endoscopic images, showing that the proposed method could predict object sizes with an error estimation smaller than +\-2.5 mm. These results are stable and relatively accurate measurements with small deviations using 100 real polyps images with different objects' sizes. It opens up a wide range of applications in the medical field, especially diagnostic and surgical endoscopy, with AI-driven clinical support systems

10:20-12:00 Session 9B: SOICT Technical Session XIV: AI Applications

CHAIR: *Van-Duy Nguyen*

LOCATION: Grand Ballroom B, 2F

10:20 *Ma Công Thành*, *Tran Duc Huy*, *Pham Ngoc Loc*, *Hoang Anh Vu* and *Nguyen Trong Khanh*

FA-Net: A Dual-Branch Attention Architecture for Extracting Fine-Grained Anatomical Features of Wood

ABSTRACT. Accurate identification of wood species is a challenging \textit{Fine-Grained Visual Classification (FGVC)}, playing a crucial role in supply chain management and in combating illegal logging. Conventional Convolutional Neural Networks (CNNs) often fail to capture subtle morphological details due to feature compression (global pooling), even though macro-images inherently contain both global structural context and fine-grained cues. To overcome this limitation, we propose \textbf{FA-Net (Fine-Anatomical Network)}, a novel dual-branch architecture that employs a \textit{global branch} to capture global structural context (e.g. porosity types and vessel distribution) and a \textit{local branch} to preserve local morphological details (e.g. parenchyma patterns and vessel/ray sizes) from macro-scale images. Both branches are enhanced with channel–spatial attention mechanisms and are adaptively fused through a pyramid self-attention module, yielding a highly discriminative representation. Comprehensive experiments across five benchmark datasets demonstrate that FA-Net achieves state-of-the-art accuracy, reaching up to 99.32\%—outperforming the DenseNet121 baseline by 4.0\%—while maintaining near-real-time inference speed. Interpretability analysis via EigenCAM further confirms that FA-Net successfully attends to critical anatomical traits (such as porosity types and parenchyma patterns). FA-Net provides an efficient, transparent and deployment-ready solution for practical applications in forestry and customs inspection.

10:40 *Dung Tran*, *Huyen Tran*, *Hong Nguyen*, *Xuan-Vu Phan* and *Nam-Phong Nguyen*

Adaptive Rainfall Forecasting from Multiple Geographical Models Using Matrix Profile and Ensemble Learning

ABSTRACT. Rainfall forecasting in Vietnam is highly challenging due to its diverse climatic conditions and strong geographical variability across river basins, yet accurate and reliable forecasts are vital for flood management, hydropower operation, and disaster preparedness. In this work, we propose a Matrix Profile-based Weighted Ensemble (MPWE), a regime-switching framework that dynamically captures covariant dependencies among multiple geographical model forecasts while incorporating redundancy-aware weighting to balance contributions across models. We evaluate MPWE using rainfall forecasts from eight major basins in Vietnam, spanning five forecasting horizons (1-hour and accumulated rainfall over 12, 24, 48, 72, and 84 hours). Experimental results show that MPWE consistently achieves lower mean and standard deviation of prediction errors compared to geographical models and ensemble baselines, demonstrating both improved accuracy and stability across basins and horizons.

11:00 *Nguyen Tran Minh Nhat*, *Pham Cong Hoang*, *Tran Phong Quan*, *Tran Le Dung*, *Nguyen Duy Long*, *Nguyen Duong Hung* and *Ho Viet Duc Luong*

Toward a Culture-Aware Vietnamese Mental Health Support Chatbot with Large Language Models

ABSTRACT. In recent years, mental health has become a pressing concern in Vietnam, driven by increasing academic, professional, and social pressures. Nonetheless, access to mental health support remains constrained by a shortage of professionals and cultural stigmas that discourage help-seeking. This paper presents an AI-based mental health chatbot powered by Large Language Models (LLMs) to deliver accessible, empathetic, and culturally tailored interactions. The model is pretrained on a 100~MB Vietnamese mental health corpus, comprising 200 documents, forum threads, and books (approximately 1.5~million tokens), capturing local linguistic patterns and psychological themes such as academic stress and familial expectations. It is fine-tuned on an adapted CACTUS dataset, consisting of 36\,577 Vietnamese counseling dialogues (31\,577 translated from English and 5\,000 synthetically generated), incorporating Cognitive Behavioral Therapy (CBT) techniques like decatastrophizing and alternative perspective. The chatbot serves as a virtual companion, providing accurate information, emotional support, and practical strategies for everyday psychological challenges. Its effectiveness is evaluated using 100 client profiles with detailed intake forms and predefined attitudes, assessed via the Cognitive Therapy Rating Scale (CTRS) for general counseling skills (understanding, interpersonal effectiveness, collaboration) and CBT-specific skills (guided discovery, focus, strategy), alongside a custom Change in Attitude Towards Guidance metric to measure shifts in client attitudes, both rated on a 0–6 scale. Results show the model's performance closely mirrors professional counseling standards, highlighting its potential for real-world deployment. With strengths in linguistic and cultural localization, the model offers a scalable solution to bridge mental health care gaps in Vietnam. Future work will focus on dataset expansion, real-world testing, and integration with digital health platforms to enhance scalability and reliability.

11:20 *Cuong Van Duc*, *Thai Tran Quoc*, *Minh Nguyen Dinh Tuan*, *Tam Vu Duc*, *Son Nguyen Van* and *Hanh Nguyen Thi*

MiRAGE: Misconception Detection with Retrieval-Guided Multi-Stage Reasoning and Ensemble Fusion

ABSTRACT. Detecting student misconceptions in open-ended responses is a longstanding challenge, demanding semantic precision and logical reasoning. We

propose MiRAGE - Misconception Detection with Retrieval-Guided Multi-Stage Reasoning and Ensemble Fusion, a novel framework for automated misconception detection in mathematics. MiRAGE operates in three stages: (1) a Retrieval module narrows a large candidate pool to a semantically relevant subset; (2) a Reasoning module employs chain-of-thought generation to expose logical inconsistencies in student solutions; and (3) a Reranking module refines predictions by aligning them with the reasoning. These components are unified through an ensemble-fusion strategy that enhances robustness and interpretability. On mathematics datasets, MiRAGE achieves Mean Average Precision scores of 0.82/0.92/0.93 at levels 1/3/5, consistently outperforming individual modules. By coupling retrieval guidance with multi-stage reasoning, MiRAGE reduces dependence on large-scale language models while delivering a scalable and effective solution for educational assessment.

10:20-12:00 **Session 9C: SOICT Technical Session XV:**
Human Computer Interaction and Intelligent Interactive Systems
CHAIRS: *Dung Le Duy* and *Tuan Linh Dang*
LOCATION: Yersin Ballroom A, 2F

10:20 *Trong Tuan Do*, *Duc Minh Nguyen*, *Van Quyen Bui* and *Dinh Duy Nguyen*

A Co-Simulation Approach for UAV-Network-AI Interaction in Digital Twin Visual Context

ABSTRACT. The growing use of Unmanned Aerial Vehicles (UAVs) in both civilian and military applications demands robust communication and smart real-time decision-making. Existing simulation platforms often lack integration between physical flight dynamics, wireless communication, and AI, resulting in a "reality gap" during real-world deployment. With an emphasis on Wi-Fi networks in contexts with obstacle modeling, this study builds on previous efforts such as CAVIAR by introducing an integrated co-simulation framework as a digital twin for UAVs. ns-3 for realistic network modeling, Microsoft AirSim with Unreal Engine for high-fidelity physical and visual simulations, and a Python-based AI component for object detection through YOLOv11 [6], allowing for synchronized bidirectional interactions under network constraints. Using a PUB/SUB model for scalable and decoupled synchronization, ZeroMQ enables soft real-time data sharing. Key contributions include: (1) a quantitative analysis of how communication constraints impact UAV behavior and AI decision-making; (2) a scalable platform that can model indoor Wi-Fi in high fidelity with obstacle effects; and (3) a reliable testbed for real-time, AI-in-the-loop UAV applications in intricate and safety-critical scenarios. In simulations, experimental results show system stability with an end-to-end latency of less than 80 ms, and real-world tests validate the high fidelity. This adaptable platform promotes research in UAV networks and edge AI while advancing autonomous UAV systems.

10:40 *Viet-Tham Huynh*, *Minh-Khang Nguyen*, *Nhut-Thanh Le-Hinh*, *Duy-Nam Ly*, *Tam V. Nguyen* and *Minh-Triet Tran*

Fairy360VR: Immersive 360° Storytelling with Large Language Models and Generative Diffusion

ABSTRACT. In the growing digital storytelling landscape, most current systems mainly focus on text or static images, which do not effectively exploit the potential of panoramic environments to create immersive experiences. In addition, automatically converting short stories into multimodal storytelling experiences remains a challenge due to the requirements of deep semantic understanding and the ability to synthesize appropriate images according to the story timeline. In this study, we propose a system that automatically generates multimodal storytelling experiences from short stories, combining both text and 360° images. First, we build a dataset of pairs of short stories and scene descriptions, which are preprocessed and mapped into semantically rich text segments to serve as a basis for training and evaluation. Next, we design a system

architecture consisting of two main components: (1) a text processing module that uses a large language model to extract and organize events in a timeline, and (2) a 360° biome module based on the biome model, allowing users to observe the story in panoramic space. The system is deployed on the web, allowing users to directly experience it by dragging and dropping the mouse to rotate and explore the surrounding space. To evaluate, we conducted a user study comparing two storytelling formats: text-based stories generated by Gemini Story Book and a 360° visual experience suggested by the system. Participants were surveyed on criteria such as attractiveness, comprehensibility, immersion, and personal preference. Preliminary results show that the 360° method provides a higher level of immersion and is rated by users as more appealing than the pure text-based stories. This study contributes to opening up a new approach to the field of digital storytelling, where the combination of language models and generative models can bring a more interactive and immersive experience to users.

11:00 *Phi Vu Vo Diep*, *Xuan Uyen Nguyen Vu*, *Chi Thanh Vi* and *Khanh-Duy Le*

Enhancing VR Drink Taste Believability using Olfactory Stimulation

ABSTRACT. Taste perception is inherently multisensory, with olfaction playing a key role alongside vision and other senses. Although virtual reality (VR) has been increasingly applied to simulate multisensory experiences, drinking interactions remain largely underexplored. This paper presents the design and evaluation of a VR system that integrates visual and olfactory cues to enrich drinking experiences and enhance immersion. The system associates predefined liquid colors with specific odors, which are automatically released when a drinking gesture is detected. Guided by a user-centered design approach, a user study was conducted to examine whether combining color and scent could modulate taste perception in VR. Results show that while liquid color alone had limited impact, olfactory cues—particularly strawberry and lemon scents—significantly shaped perceptions of sweetness and sourness. These findings demonstrate the potential of incorporating olfaction into VR to advance Human-Food Interaction, paving the way for future multisensory applications in entertainment, education, and food commerce.

11:20 *Thi Quynh Hoa Nguyen*, *Duy Hai Nguyen*, *Thanh Ha Le* and *Thi Duyen Ngo*

An eye-tracking system for extracting and visualizing visual features of dyscalculia in children

ABSTRACT. Dyscalculia is a learning disability characterised by poor performance on tasks involving spatial and numerical processing. The use of eye tracking to study developmental disorders has been widely used, but its application to dyscalculia has been relatively neglected. Most previous studies lack specific technical methods to identify the visual features characteristic of children with dyscalculia. Our study seeks to bridge this gap by introducing an eye-tracking system designed to elucidate the underlying perceptual problems of Vietnamese dyscalculia children as they perform task demands. It not only explores visual attention strategies but also uncovers fundamental issues such as auditory or spatial order perception when performing calculations. In addition, the visual problems regarding space and time are clarified. The system collects multimodal signal information from the child's eye and mouse-based interactions during the task. Preliminary findings from the comparative analysis between dyscalculic and typically developing children suggest that the system may provide a solid foundation for future research on detection, intervention, and development of assistive applications tailored to the visual processing abilities of these children.

11:40 *Trung-Hau Nguyen-Tran*, *Dung-Minh Nguyen*, *My-Le Duong-Thi*, *Ngoc-Trinh Nguyen-Thi*, *Thi-Hai Vo*, *Hoai-Nam Do* and *Khanh-Duy Le*

MO-PO RM: A Collaborative Mixed Reality Board Game for Engaging Players and Audience in Learning through Playing

ABSTRACT. Collaborative educational activities often rely on physical artifacts such as board games, which encourage manipulation, peer interaction, and active engagement. Yet, traditional board games suffer from several limitations: they require fragile physical resources, manual orchestration by a facilitator, and remain difficult to adapt for formal evaluation. To address these issues, we designed and evaluated MO-PO MR, a Mixed Reality (MR) version of Mono-Poly, a chemistry card game where players create polymers by combining monomers. MO-PO MR serves as an exemplary collaborative MR board games which can eliminate the need for fragile physical components, automate rule enforcement and scorekeeping, and reduce opportunities for error or cheating. The system supports both players and audiences: co-located participants use tablets to interact with an augmented game board, while spectators view a large third-person MR display where virtual objects are spatially fused with real-world participants. This hybrid design ensures accessibility even for users without MR headsets. We conducted a within-subject study with groups of three players and one observer, comparing the MR version against the physical game. Results indicate that MO-PO MR improves clarity, coordination, and immersion, while making the game easier to follow for both players and observers. These findings highlight the potential of MR-based collaborative games to enhance STEM education by offering scalable, traceable, and engaging learning experiences.

10:20-12:00 Session 9D: SOICT Technical Session XVI: Multimedia Processing

CHAIR: *Ichiro Ide*

LOCATION: Yersin Ballroom B, 2F

10:20 *Trong-Thuan Nguyen* and *Minh-Triet Tran*

LOGOS: Language-guided Oriented Object Detection in Aerial Scenes

ABSTRACT. Object detection in geospatial scenes, such as satellite and aerial imagery, presents significant challenges due to the varying orientations and densities of objects, as well as the complex backgrounds inherent in remote sensing images. Traditional methods for oriented object detection have struggled to address issues such as angular discontinuity, fixed query sizes, and inefficiencies in handling sparse or cluttered scenes. In this paper, we propose LOGOS, a novel transformer-based approach that leverages textual prompts to guide the detection of oriented objects in aerial scenes. In particular, our proposed approach incorporates prompt-modulated content queries to dynamically adjust the model's focus based on the given text, ensuring more accurate object detection in complex environments. Empirically, extensive experiments on the DOTA dataset demonstrate that LOGOS outperforms existing state-of-the-art methods, particularly in densely packed and rotated object scenarios. Our approach offers a significant step forward in improving the robustness and scalability of oriented object detection in remote sensing applications.

10:40 *Nguyen Thanh Khoi*

From Text to Thumbnail: A Unified Framework for Automated News Image Generation and Evaluation for Daily Activities

ABSTRACT. Images play a crucial role in online news consumption, attracting attention and driving user engagement. While recent Text-to-Image (T2I) models can generate high-quality images from text, selecting images that accurately reflect news content remains subjective and influenced by user preferences. In this paper, we propose a unified framework for news image generation that systematically addresses this challenge targeting daily activities. Our approach introduces a novel criteria-learning phase to extract salient visual attributes from 6,000 thumbnail images across various

domains and a prompt enrichment pipeline to create Grounded Summaries from article text. Finally, each generated image is assessed using a evaluation protocol that quantifies semantic and visual quality via established human preference models. To validate our approach and facilitate future research, we apply our full pipeline to generate and evaluate over 10,000 images for 1,500 news articles, releasing this collection as the first benchmark dataset for grounded news image generation. Our experiments on this dataset demonstrate the framework's effectiveness and highlight how our prompt enrichment method successfully balances semantic fidelity with aesthetic appeal.

11:00 *Nguyen Duong Hung*, *Hoang Danh Quan*, *Ho Viet Duc Luong*, *Lang Hong Nguyet Anh*, *Nguyen Thi Thuy* and *Dinh Viet Sang*

Self-Supervised ViT for Endoscopy: I-JEPA Pretraining with Label-Free Diffusion Assessment

ABSTRACT. The scarcity of expert-labeled data remains a major barrier for gastrointestinal endoscopy image analysis, as annotation is costly and time-consuming while vast amounts of clinical images remain unlabeled. To address this challenge, we propose a label-efficient pipeline that combines self-supervised pretraining and label-free evaluation. Specifically, we pretrain a Vision Transformer (ViT) encoder on 66,820 unlabeled endoscopy frames using I-JEPA, a joint-embedding predictive approach. To assess the quality of learned representations without relying on annotated data, we introduce a diffusion-based probing mechanism—Reconstruction-Conditioned Diffusion Modeling (RCDM)—which reconstructs images from latent features to provide a qualitative, label-free evaluation. Finally, we transfer the pretrained encoder to downstream tasks including classification, lesion/polyp segmentation, and multitask learning. Experiments on public benchmarks (e.g., Kvasir, CVC series, ETIS) and private datasets demonstrate that I-JEPA pretraining, particularly when combined with sequential MAE to I-JEPA adaptation, yields superior segmentation performance compared to strong baselines such as RaBiT, with larger gains in low-label regimes. Multitask analysis further highlights the role of decoder architecture, where ViT encoders paired with RaBiT-style decoders surpass EndoUnet in most tasks. These results show that our pretrain–probe–transfer framework enables domain-aware, label-efficient representation learning for endoscopic image analysis, providing both practical benefits under label scarcity and actionable insights for multitask model design.

11:20 *Hoang-Phuc Nguyen*, *Phuong-Linh Huynh-Ha* and *Minh-Triet Tran*

Generalizability Evaluation and Anchor-Guided Approach for Category-Agnostic Pose Estimation

ABSTRACT. Category-agnostic pose estimation models may overfit to the limited set of predefined landmarks within the training dataset, resulting in large errors for query points far from these landmarks. We demonstrate this effect by analyzing the error distribution of query points on human faces. We find that query points distant from landmarks exhibit high errors, suggesting existing models struggle to generalize beyond the training data. To handle this problem, we introduce a training-free, anchor-guided geometric mapping approach to improve keypoint prediction. Our method leverages reliably predicted anchor points to construct a pose-consistent geometric basis via Delaunay triangulation. It then uses barycentric coordinate interpolation to map any query point from a support image to a target image, preserving geometric structure across different poses. Quantitative evaluation on human faces and qualitative analysis across diverse categories confirm the overfitting issues and show that our approach significantly improves keypoint accuracy without requiring additional training.

11:40 *Duc-Tuan Luu, Diem Nguyen, Anh-Khoa Nguyen Vu and Vinh-Tiep Nguyen*

RIOT: Robust Incremental Few-Shot Instance Segmentation via Synthetic Feature Generation with Optimal Transport

ABSTRACT. Few-shot instance segmentation (FSIS) extends few-shot detection by requiring both object localization and accurate mask prediction, but remains underexplored in incremental settings where new categories arrive over time with limited annotations. Existing incremental FSIS methods mainly focus on classifier adaptation and knowledge preservation, while neglecting data augmentation or feature generation, which are crucial to mitigate overfitting and distribution mismatch under extreme data scarcity. In this work, we present RIOT, Robust Incremental Few-Shot Instance Segmentation via Synthetic Feature Generation with Optimal Transport. RIOT follows a two-stage pipeline: (1) base training on abundant categories to learn strong segmentation features, and (2) generator training with both Optimal Transport and KL-divergence losses to produce class-conditional synthetic features aligned with real distributions. Unlike prior FSIS approaches, RIOT supports incremental learning without additional fine-tuning stages. Extensive experiments on standard benchmarks demonstrate that RIOT significantly improves recognition of novel classes while maintaining base-class knowledge, establishing a strong baseline for incremental FSIS with synthetic feature generation.

10:20-12:00 Session 9E: Poster Exhibition

CHAIRS: *Van Khu Vu*, *Dung Le Duy* and *Ichiro Ide*

Danh Luu Thanh, *Thien Tran Duc*, *Bao Bui Duy*, *Le Vu Nguyen Dinh*, *Tien Le Nam* and *Tinh Anh Nguyen Nhu*

Integrated Semantic and Temporal Alignment for Interactive Video Retrieval

ABSTRACT. The growing volume of video data and the introduction of complex retrieval challenges, such as the Temporal Retrieval and Alignment of Key Events (TRAKE) task, expose critical limitations in existing systems. Many methodologies lack scalable, holistic architectures and rely on "frozen" embedding models that fail on out-of-knowledge (OOK) or real-world queries. This paper introduces a comprehensive, modular video retrieval framework designed to address these gaps. Our system features a scalable architecture integrating TransNetV2 for scene segmentation, BEiT-3 for visual embeddings in Milvus, and Gemini OCR for metadata in Elasticsearch. We propose two novel components: (1) QUEST (Query Understanding and External Search for Out-of-Knowledge Tasks), a two-branch framework that leverages a Large Language Model (LLM) for query rewriting and an external image search pathway to resolve OOK queries; and (2) DANTE (Dynamic Alignment of Narrative Temporal Events), a novel dynamic programming algorithm that efficiently solves the temporally-incoherent TRAKE task, which has an efficient O(NT) time complexity. These contributions form a robust, scalable, and intelligent system that significantly advances the state-of-the-art in handling complex, real-world video search queries.

Vi Chau The, *Luan Nguyen The*, *Nghi Tran Gia*, *Minh Mai Nguyen Phuc*, *Quyen Nguyen Huu*, *Son Ngo Duc Hoang* and *Duy Phan The*

HelioSearch: A Multimodal Video Retrieval Framework with LLM-Driven Query Expansion and Hybrid Filtering

ABSTRACT. Video event retrieval in large-scale multimedia databases remains a critical challenge due to the inherent complexity of multimodal understanding and semantic alignment across heterogeneous data sources. This paper presents the design and development of a unified multimodal video retrieval system that addresses key limitations of existing approaches in cross-modal representation learning, temporal reasoning, and semantic consistency. The proposed framework leverages BEiT-3 and CLIP as unified transformer encoders to learn shared semantic representations. To overcome single-modality constraints, the system integrates

Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), and YOLOv12-based object detection for fine-grained entity filtering. These modalities are organized within a specialized database architecture that combines vector, document, and search indexes to enable efficient multimodal fusion. The system supports diverse query types, further enhanced through Large Language Model (LLM)-assisted query expansion. Comprehensive experiments conducted on the Ho Chi Minh City AI Challenge 2025 (AIC 2025) dataset demonstrate substantial improvements in retrieval precision and ranking stability, validating the system's effectiveness and generalization capability. Overall, the proposed framework offers a scalable, interpretable, and extensible foundation for real-world multimodal video event retrieval applications.

Van-Thinh Vo, *Minh-Khoi Nguyen*, *Minh-Huy Tran*, *Anh-Quan Nguyen-Tran*, *Duy-Tan Nguyen*, *Khanh-Loi Nguyen* and *Anh-Minh Phan*

Enhanced Multimodal Video Retrieval System: Integrating Query Expansion and Cross-modal Temporal Event Retrieval

ABSTRACT. Multimedia information retrieval from videos remains a challenging problem. While recent systems have advanced multimodal search through semantic, object, and OCR queries - and can retrieve temporally consecutive scenes - they often rely on a single query modality for an entire sequence, limiting robustness in complex temporal contexts. To overcome this, we propose a cross-modal temporal event retrieval framework that enables different query modalities to describe distinct scenes within a sequence. Another key contribution is the Kernel Density Gaussian Mixture Thresholding (KDE-GMM) algorithm, which adaptively determines decision thresholds for scene transition and slide change detection, ensuring optimal keyframe selection. These extracted keyframes act as compact, high-quality visual exemplars that retain each segment's semantic essence, improving retrieval precision and efficiency. Additionally, the system incorporates a large language model (LLM) to refine and expand user queries, enhancing overall retrieval performance. The proposed system's effectiveness and robustness were demonstrated through its strong results in the Ho Chi Minh AI Challenge 2025.

Huu-Tuan Nguyen, *Dien X. Tran*, *Minh-Thinh Vo*, *Thanh-Nhan Nguyen*, *Thanh-Long Nguyen Thai* and *Phuc-Lu Le*

VidAlign: Integrating Multi-Event Alignment and LLM Co-Searching for Video Retrieval

ABSTRACT. The exponential growth of video content demands efficient retrieval systems capable of understanding complex, multi-event scenarios. In this work, we present VidAlign, a fine-grained video retrieval framework that effectively aligns multi-event textual queries with temporally distributed visual content. The framework introduces a novel TDP-Fuse (Temporal Dynamic Programming Fusion) algorithm to dynamically align and fuse partial retrieval results over time. Furthermore, an LLM-guided Co-Searching mechanism is incorporated to assist users in query formulation and refinement, leveraging large language models to enhance semantic understanding and interactive retrieval. VidAlign's architecture combines fast approximate search via FAISS with intelligent reranking, temporal fusion, and adaptive co-searching, ensuring both scalability and retrieval accuracy. The system has been rigorously evaluated in real-world settings and demonstrated outstanding performance at the Ho Chi Minh AI Challenge 2025, ranking among the top solutions. Experiments show that VidAlign effectively enhances semantic alignment and temporal coherence, making it a strong competitor in large-scale event-based video retrieval.

Nam Khanh Tran, *Le Dac Phu Phan*, *Minh Loi Duong*, *Ngoc Thien An Nguyen* and *Truong Thinh Nguyen*

PerceptionBrowswer: Enhancing information retrieval system with spatial-temporal knowledge

ABSTRACT. The ubiquity of internet and media service in the modern world has provided the enormous amounts of data, which is considered the new gold in the informatics era. This emergence has nurtured the development of media retrieval systems, in which the users have the ability to extract the most relevant piece of information from the enormous dataset. Therefore, we propose an effective retrieval system that could run on a portable machine like laptops without the need of internet access, which enhance the system's robustness and usage in privacy-sensitive situations. First, we process the enormous dataset to reduce its size significantly, which enables operation on resource-restrained machines. Second, to facilitate nformation retrieval by text queries, we use CLIP-based features obtained from visual foundation models. This allows us to integrate both spatial (image) and temporal (video) features in our system. Furthermore, we also introduce a temporal combination algorithm to enhance the temporal understanding and retrieval performance. Benchmarking our system on the set of queries provided in the elimination round of Ho Chi Minh AI Challenge 2025 (AIC25), we achieved an impressive score of 84.4/88, equivalent to 95.91\% accuracy, with an average query response time of under 15 seconds. These results underscore our system's robustness in managing diverse and complex queries, demonstrating it as an efficient tool for life-log and media retrieval purposes by significantly enhancing the user experience for both common and advanced usage while maintaining minimal resource requirement. Our code is publicly available at \url{https://github.com/trnKhanh/past-beggars}.

Hoang Vu Do, *Quoc Dat Do*, *Truong Duy Nguyen*, *Le Hai Binh Tran*, *Xuan Tri Pham*, *Ham Duong Tran* and *Cam-Hao Hua*

TARS: Temporal Alignment Retrieval System for Efficient Multi-Segment Video Event Retrieval

ABSTRACT. Temporal video event retrieval requires returning video segments whose frames follow the action order stated by a natural-language query. Existing systems built on global or scene-level similarity often surface visually plausible yet order-inconsistent matches; learning temporal encoders improves ordering but adds training cost and degrades robustness under domain shift. We present TARS, a training-free, order-aware framework that performs temporal reasoning entirely at inference time. A query is decomposed into sub-events, then being embedded by complementary vision–language encoders; and a monotonic dynamic-programming alignment searches the best ordered path on the frame–subevent similarity matrix. A prefix-maximum recurrence yields O(nm) time and O(m) memory per shot and integrates cleanly with candidate retrieval and lightweight re-ranking. On the AI Challenge HCMC 2025 benchmark, TARS attains 93.15% Top-1 accuracy, demonstrating that explicit inference-time temporal alignment over frozen embeddings is a simple, robust, and deployable solution for order-sensitive video retrieval.

Thinh-Phat Vo, *Dang-Khoa Mai*, *Nguyen-Khang Ly*, *Quang-Thang Duong* and *Quoc-Thang Nguyen*

KPT: Enhancing Temporal Event Retrieval in Vietnamese News Videos

ABSTRACT. This paper presents KPT, an enhanced Vietnamese news event retrieval system developed for the 2025 Ho Chi Minh City AI Challenge. The system represents an evolution of our previous solution, focusing on improving temporal event search accuracy and overall system efficiency. First, a key upgrade is the integration of the

Milvus vector database, which replaces the prior in-memory implementation and enables faster retrieval with scalable handling of large video datasets. Second, the web-based user interface and core functionalities have been completely redesigned to provide a more intuitive and efficient user experience, supporting complex multimodal querying and result submission. Third, the temporal search algorithm has been upgraded from a frame-level to a scene-level indexing approach, incorporating a new scoring function that jointly models semantic similarity and temporal order, significantly improving both computational efficiency and the precision of event boundary detection. The enhanced KPT system was successfully deployed in the preliminary round, achieving substantially better performance than the previous version and demonstrating the effectiveness of the proposed architectural and algorithmic enhancements.

Nga Nguyen, *Dat Tien Nguyen* and *Huy M. Le*

TEMPO: A Multimodal Video Retrieval System with Sequential Query Support

ABSTRACT. The Ho Chi Minh AI Challenge 2025, sets the ambitious goal of building a powerful video retrieval system. The competition requires teams to handle a medium-size dataset under a tight timeline, which demands solutions that balance both speed and accuracy. To stay competitive, we design and implement TEMPO, a system that integrates multiple search strategies, including Textual Search, Visual Search, and most importantly, Temporal Search. Starting from the official dataset, we separates and processes both audio and visual streams. These pipelines enable us to build strong features that drive the system: Semantic Search, OCR Search, ASR Search, and Temporal Search. Among them, Temporal Search stands out by supporting sequence-based queries, also known as Sequential Query Video Retrieval. This feature is still rare in commercial systems and represents a novel contribution to the competition. Our system achieves promising results and is currently ranked among the top teams in the preliminary round, based on 50\% of the ground-truth data provided by the organizers. These outcomes highlight the practical potential of TEMPO and its effectiveness in addressing real-world video retrieval tasks.

Nhat-Trung Nguyen-Tran, *Thai Bao Huynh*, *Hai-Dang Pham*, *Hieu Thien Vu*, *Y-Nhi Duong-Thai* and *Thang Cap*

FRED: Unified Multimodal Fusion and Dynamic Temporal Reasoning with Semantic Query Expansion and Exclusionary Search

ABSTRACT. Our proposed system introduces an innovative approach to interactive multimodal video retrieval, developed for the AI Challenge Ho Chi Minh City 2025. The system enhances both retrieval accuracy and user interaction through the integration of Large Language Models (LLMs) for semantic reasoning and query expansion, effectively addressing query ambiguities and improving contextual relevance. The retrieval framework is built upon Vision-Language Models (VLMs) to support text-to-video and image-based search, while incorporating auxiliary components such as Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), and Object Detection to enrich multimodal understanding. These complementary signals enable the system to capture textual, auditory, and visual cues from videos, creating a more comprehensive search foundation. Furthermore, a dynamic temporal search mechanism evaluates frame-level relevance and temporal dependencies, providing adaptive and context-aware retrieval. Overall, our system demonstrates the effectiveness of combining multimodal perception with LLM-driven intelligence to advance the precision, adaptability, and interactivity of modern video retrieval systems.

Kiet Pham Gia, *An Le Kha*, *Vinh Nguyen Quoc*, *Nhi Nguyen Truong Cong*, *Khanh Dinh Quoc*, *Binh Tran Le Hai*, *Tri Pham Xuan*, *Duong Tran Ham* and *Hao Hua Cam*

A Video Retrieval System with Advanced Temporal Algorithm and Vision Language Models Integration

ABSTRACT. Video retrieval is the demanding task of locating the exact moment within an unbounded video collection that corresponds to a user query. This capability is critical as the volume of video data explodes, becoming a "digital haystack" where proficiently finding a specific information is practically impossible. Recent retrieval systems have successfully taken advantages of technology innovations to trivialize many challenges of the task, especially in the era of Large Language Models (LLMs) and Vision Language Models (VLMs). However, limitations remain: (i) most of the current systems only utilize LLMs and VLMs for retrieval operations (query detailing, assisted fusion, etc), undermining their potential; and (ii) - the lack of temporal-capable retrieval methods. Addressing these challenges, this paper presents a pipeline that utilized the strength of VLMs and LLMs in both the pre-processing and retrieval operations, integrated into a system with an advanced dynamic-anchoring temporal algorithm and a refined semantic extraction workflow. Using AIC 2025 - a video retrieval competition - as a benchmarking ground, the proposed systems achieved 97 percent accuracy during preliminary rounds, demonstrating its practicality in terms of performance, interactivity, and scalability.

Lan Tuan Vo, *Le Ngoc Thinh Vo*, *Hong Phuc Nguyen*, *Phuong Duy Do* and *Huu Dang Nguyen Nguyen*

FrameSeeker: Shot-Level Captioning with Multimodal Hints for Efficient Video Retrieval

ABSTRACT. The exponential growth of multimedia data has created an urgent need for video retrieval systems that can deliver fine-grained semantic understanding without excessive computational cost. Existing methods often rely on detailed captioning for every frame, which fails to capture temporal context and results in redundant inference. To address these challenges, a multimodal retrieval framework is proposed that integrates hybrid semantic keyword search, query enhancement, and temporal reasoning. A key feature of the system is the use of multimodal cues and few-shot prompt engineering within a Vision Language Model to generate a single, temporally coherent caption for all frames in a shot. By jointly using information from visual content, object detection, and text recognition, this approach produces rich and semantically grounded embeddings that improve retrieval precision while reducing inference cost. The effectiveness of the proposed method is validated through its strong performance in the 2025 Ho Chi Minh City AI Challenge.

Kim Nguyen, *Quan Nguyen Hung*, *Dat Phan Thanh*, *Tien Huynh Viet*, *Hoang Tran Van*, *Minh Vo*, *Tien Do* and *Thanh Duc Ngo*

Adaptive Agent-Guided Dynamic Programming for Temporal Optimization in Multi-Event Video Retrieval

ABSTRACT. As multimedia content continues to expand rapidly, achieving accurate and efficient video retrieval has become increasingly critical. Among various retrieval tasks, multi-event retrieval poses greater challenges, as it requires understanding both semantic content and temporal dependencies across events. However, existing approaches often rely on locally greedy matching strategies that enforce only pairwise temporal consistency, leading to suboptimal alignments and disrupted event order. To overcome these limitations, we propose a Dynamic Programming (DP)-based video retrieval framework that formulates the task as a global temporal optimization problem. Our method jointly optimizes the entire event sequence to identify the globally optimal keyframe path that best preserves both chronological flow and semantic coherence. This DP formulation effectively reduces the search space while

maintaining temporal consistency and semantic integrity. To enhance automation and adaptivity, the framework is supported by an agent-guided coordination layer powered by a Large Language Model (LLM). This layer interprets user intent, decomposes multi-event queries into structured representations, and autonomously triggers the DP-based retrieval pipeline. Experimental results demonstrate strong performance in the 2025 Ho Chi Minh AI Challenge, highlighting the potential of this agent-guided, DP-centered framework for large-scale, intelligent multimedia retrieval.

Nguyen-Phuc Mac, *Lam-Phuc Nguyen-Le*, *Van-Kinh Nguyen*, *Gia-Bao To* and *Van-Tai Phan*

GalaxyAssistant: An Intelligent Assistant for Multimedia Event Retrieval

ABSTRACT. The exponential growth of large-scale multimedia data necessitates efficient event retrieval systems, a challenge addressed by competitions like LSC, VBS, and the Ho Chi Minh City AI Challenge. To address this, we propose GalaxyAssistant, an intelligent assistant framework designed for in-depth analysis and information retrieval from complex video data. Our system's intelligence is rooted in its tri-modal indexing, which concurrently processes video via three parallel pipelines following shot detection (TransNetV2). First, a visual-language model (SigLIP) generates dense visual embeddings. Second, an image captioning model (InternVL) creates textual descriptions. Third, an ASR model (ChunkFormer) transcribes the audio. Both caption and audio transcripts are encoded using a hybrid dense-and-sparse model (BGE-M3 + BM25) to create robust textual indices. During retrieval, the assistant dually encodes a user's query using both SigLIP and BGE-M3+BM25 text encoders. A concurrent search is then executed via the Hierarchical Navigable Small Worlds (HNSW) algorithm against all three vector databases (visual, caption, and audio). This tri-modal fusion allows our assistant to perform high-precision analysis and retrieval for complex, event-based queries.

Duc-Thang Nguyen and *Minh-Triet Tran*

Text-Guided Filtering to Enhance Open-Vocabulary Object Detection for Sport Event Retrieval

ABSTRACT. Accurately interpreting sports broadcasts requires not only visual perception but also understanding textual cues such as player names, jersey numbers, and scoreboard information. Traditional object detectors, constrained by closed-set vocabularies, struggle to generalize across unseen entities and lack the ability to reason about text embedded in the scene. This paper presents a text-guided filtering framework that integrates YOLO-World, an open-vocabulary object detector, with an Optical Character Recognition module to enhance fine-grained football scene understanding. Experiments on FIFA World Cup 2022 broadcast frames demonstrate that incorporating textual cues improves recognition accuracy and contextual coherence, particularly in event-specific cases such as goals and substitutions. The results highlight the effectiveness of text-guided filtering for multimodal reasoning, offering a scalable direction for open-vocabulary object detection in structured visual domains.

Phu Truong Thien, *Huy Gia Ngo*, *Tan Nhat Nguyen*, *Thuy T. N. Nguyen*, *Dat Tien Nguyen* and *Huy M. Le*

Fusurge: An Accelerated Query-Driven System for Multimodal Information Retrieval

ABSTRACT. With the rapid expansion of multimedia archives, video retrieval systems must strike a balance between scalability, accuracy, and responsiveness. Fusurge addresses this challenge by integrating a compact yet powerful pipeline. Key components include data processing for rapid keyframe sampling, PaddleOCR for multilingual OCR, and faster-whisper for efficient ASR. On the semantic side, multiple CLIP-based encoders are fused to widen coverage, while compact vision–language

models support interactive question answering. Retrieval quality is further refined through reranking with user-guided clarification. An intuitive interface lowers entry barriers and ensures ease of use. Empirically, Fusurge demonstrates strong retrieval quality and robust system performance, thereby achieving consistent top-rank effectiveness up to 86% with responsive, stable runtime — making it suitable for real-world, large-scale use.

Dong Quan Ngo Nguyen, *Nhat Bao Vo*, *Phuc Nguyen Pham*, *Gia Huy Vo Le* and *Thien Bao Ha Van*

ATLAS: Adaptive Temporal Low-rank Alignment System for AI Challenge 2025

ABSTRACT. Text-Video Retrieval (TVR) on large-scale event data requires scalable and semantically rich solutions, particularly within the AI Challenge 2025 - a national competition fostering research in multimodal retrieval. This paper presents ATLAS, the system submitted to the competition, designed to address computational inefficiency and contextual misalignment in cross-modal search. Built upon Milvus and Elasticsearch, and leveraging foundation models such as CLIP and BLIP-2, ATLAS adopts a novel Adaptive Fusion architecture. The system introduces three key innovations: (1) Low-rank Modulation (LoRM), adapted from the RAP architecture, to mitigate temporal redundancy and generate highly representative LoRM-enhanced Activity Vectors; (2) Selective Lite-SGG, which encodes structural context only on salient keyframes, balancing efficiency and expressiveness; and (3) Weighted Reciprocal Rank Fusion (WRRF), a dynamic ranking mechanism that adjusts weights based on query complexity to integrate multi-modal retrieval results effectively. Experimental evaluations demonstrate that ATLAS achieves robust, accurate, and adaptive retrieval performance on large-scale news datasets, setting a strong benchmark for future AI Challenge systems.

Anh Tai Pham Nguyen, *Tung Duong Le Duc*, *Anh Duy Le* and *Trung Hieu Truong Le*

MERVIN: A Unified Framework for Multimodal Event Retrieval in Vietnamese News Videos

ABSTRACT. The rapid growth of online video platforms has created an increasing need for effective and semantically grounded event retrieval systems. To address this, we propose MERVIN, a unified multimodal framework for Vietnamese news video retrieval that integrates visual and textual representations through keyframes, transcripts, and video sum- marizations. The framework enhances textual quality using the Gemini 1.5 Flash model for transcript cleaning and summarization, effectively re- ducing noise caused by accents, background interference, and recognition errors. For visual understanding, features are extracted using the Perception Encoder model, while a Vietnamese-specific language model gener- ates textual embeddings to ensure linguistic relevance. Both visual and textual embeddings are indexed in a Milvus vector database, enabling ef- ficient similarity-based retrieval. On top of this, a web-based interactive interface built with React allows users to iteratively refine queries across modalities, leading to more accurate and semantically aligned search re- sults. Experimental results on Vietnamese news videos demonstrate the effectiveness of the proposed system, with MERVIN achieving 79 out of 88 points in AI Challenge HCMC 2025 qualification phase.

Thuong Phuc Nguyen, *Thanh Hung Nguyen*, *Thanh Dang Phan*, *Thanh Vinh Nguyen* and *Thang Cap*

Aligning Time and Semantics (ATS): A System for Temporal Retrieval and Alignment of Key Events

ABSTRACT. Video retrieval is the process of locating the exact moment in a vast video collection to match the textual description provided by the users. This problem presents a considerable challenge in the era of the digital world, where thousands of media content like videos, images, and audio are everywhere on the Internet. For this

reason, as part of the Ho Chi Minh AI Challenge 2025, we developed an innovative framework called Aligning Time and Semantics (ATS), capable of specifying the video segment that corresponds to the question from the users. This framework is integrated with multimodal models, with the ability not only to combine various types of models but also to effectively manage temporal searches through multi-stage processes.

Nguyễn Văn Minh, *Phan Nhật Tân*, *Nguyễn Văn Hồng Thái*, *Hoàng Đức Dũng* and *Nguyễn Đình Thiên Quang*

Unlocking Arbitrary-Length Querying for Video Retrieval via Advanced Vision-Language Models and Hybrid Temporal Search

ABSTRACT. Video retrieval, a critical task in multimedia retrieval, faces challenges in aligning complex textual queries with video content, especially for long or multimodal queries. This paper introduces a novel multimodal retrieval system designed for the AI Challenge 2025, emphasizing efficient video retrieval. Our approach utilizes LongCLIP as the primary Vision-Language Model (VLM) to address token limitations of previous models, enabling robust processing of extended queries. We integrate OCR, ASR, and object detection, fused via a weighted sum strategy to enhance retrieval accuracy. Additionally, we propose a temporal search algorithm to precisely identify frames corresponding to specific actions described in the query, improving temporal alignment. Experimental results from the AI Challenge 2025 demonstrate the system's superior precision and efficiency in real-world video search scenarios.

An Nguyen Tran Khuong, *Hieu Vu Minh*, *Nguyen Le Binh*, *Hieu Huynh Minh* and *Dai Phan Trong*

Tournament-Inspired Elimination Reranking for Multi-Modal Video Retrieval

ABSTRACT. This paper presents a novel multi-modal video retrieval framework designed to enable efficient and accurate content discovery across large-scale video datasets. The system integrates multiple vision–language models to enhance semantic alignment between textual queries and video content. A key contribution is a double-elimination reranking mechanism inspired by tournament structures, which improves recall and ranking stability through multi-stage candidate evaluation. The framework employs a three-tier storage architecture comprising Milvus for vector similarity search, Elasticsearch for hybrid lexical–semantic retrieval, and MongoDB for metadata management. Reciprocal Rank Fusion is used to combine results from multiple encoders, while post-retrieval reasoning with Gemini is applied to answer user queries and align frames with corresponding events. Demonstrated in the competition setting, the system shows strong potential for large-scale, intelligent multimedia retrieval.

Huy-Giap Bui, *Minh-Huy Trinh*, *Canh-Toan Le*, *Quoc-Lam Vu* and *Duy-Hung Do Hoang*

Multi-modal and Temporally-aware Video Retrieval

ABSTRACT. In this work, we present a multi-modal and temporally-aware video retrieval framework designed for multi-event video search. The proposed method captures diverse semantic information for each keyframe, including image embeddings, optical character recognition text, object detection features, and audio transcripts. Moreover, event dependencies are modelled with a decay-based weighting mechanism to improve long-sequence matching across multiple modalities. In addition, an enhanced user interface with an integrated chatbot assistant is developed to support faster and more streamlined query formulation. Comparative experiments with another video retrieval system demonstrate that the proposed approach enhances contextual understanding and retrieval performance for complex, time-dependent video queries in the HCMC AI Challenge 2025.

Kha Nguyen Hieu, *Dat Le Phat*, *Phat Nguyen Duc*, *Long Pham Hoang*, *Anh Nguyen The*, *Binh Tran Le Hai*, *Tri Pham Xuan*, *Duong Tran Ham* and *Hao Hua Cam*

Cross Segment Coherence Scorer: A Training Free Temporal Framework for Multimodal Video Retrieval

ABSTRACT. Video event retrieval involves identifying and aligning segments that correspond to the events or actions described in a user's query. As large scale video archives continue to grow, managing their multimodal and temporal complexity becomes a critical challenge, requiring accurate cross modal retrieval over visuals, text, and audio while preserving the correct order of events. Many existing systems incorporate temporal checks by verifying frame order or expanding fixed time windows; however, these rigid approaches often fail in real world editing scenarios where interleaved scenes are rejected as out of order and events separated by longer gaps are missed. To overcome these issues, we propose a training free video event retrieval framework that combines late fusion with explicit inference time temporal scoring. Specifically, multimodal fusion based on Reciprocal Rank Fusion (RRF) unifies visual, OCR, and ASR evidence without retraining, and a Cross Segment Coherence Scorer (CSCS) applies soft penalties for order reversals, long temporal gaps, and jumps across shots to handle interleaved and repeated scenes while preserving temporal coherence. This design provides temporal flexibility with minimal inference overhead. The framework was evaluated on the AI Challenge HCMC 2025 benchmark, achieving 95% accuracy and demonstrating competitive performance in event level video retrieval.

Quan Ho Minh, *Duong Tran Anh*, *Hanh Nguyen Thi My*, *Khoa Vo Dang*, *Van Thai Hung*, *Duy-Dinh Le* and *Thanh Duc Ngo*

Poly-Temporal Seach: Bridging Composed and Temporal Queries for Multimodal Video Retrieval

ABSTRACT. Retrieving relevant video content from large-scale datasets requires understanding both what appears in a scene and how events unfold over time. However, existing approaches typically focus on a single aspect of this problem: compositional models capture detailed object–text relationships but fail to model temporal evolution, whereas temporal models track event sequences but overlook fine-grained scene semantics. This separation limits the ability to reason over complex, narrative-style queries that combine spatial composition with chronological order. Moreover, current systems lack robust keyframe selection and unified multimodal representations, frequently relying on static frame-level alignment or local similarity fusion that fails to maintain global temporal coherence. To address this gap, we propose Poly-Temporal Search, a unified framework that integrates compositional and temporal reasoning within a single retrieval process. Our method introduces Adaptive Sampling Keyframe Selection for stable and representative frame selection, Spherical Linear Interpolation for text-grounded compositional retrieval, and a beam-based temporal search to ensure event-level coherence. Together, these components enable joint modeling of intra-scene semantics and inter-event dependencies. Evaluated on the HCMAI 2025 benchmark, Poly-Temporal Search achieved finalist-level performance, demonstrating the effectiveness of unified multimodal reasoning for complex video retrieval tasks and highlighting the importance of bridging compositional and temporal paradigms in multimodal understanding.

Hieu Cao, *Khang Tran*, *Thinh Pham*, *Nghiem Diep* and *Binh Nguyen*

LGCA: Enhancing Semantic Representation via Progressive Expansion

ABSTRACT. Recent advancements in large-scale pretraining in natural language processing have enabled pretrained vision-language models such as CLIP to effectively align images and text, significantly improving performance in zero-shot

image classification tasks. Subsequent studies have further demonstrated that cropping images into smaller regions and using large language models to generate multiple descriptions for each caption can further enhance model performance. However, due to the inherent sensitivity of CLIP, random image crops can introduce misinformation and bias, as many images share similar features at small scales. To address this issue, we propose Localized-Globalized Cross-Alignment (LGCA), a framework that first captures the local features of an image and then repeatedly selects the most salient regions and expands them. The similarity score is designed to incorporate both the original and expanded images, enabling the model to capture both local and global features while minimizing misinformation. Additionally, we provide a theoretical analysis demonstrating that the time complexity of LGCA remains the same as that of the original model prior to the repeated expansion process, highlighting its efficiency and scalability. Extensive experiments demonstrate that our method substantially improves zero-shot performance across diverse datasets, outperforming state-of-the-art baselines.

Quoc-Duy Tran, *Anh-Tuan Vo* and *Trung-Nghia Le*

Visual Retrieval-Augmented Generation for Silhouette-Guided Animal Art

ABSTRACT. Generative AI has advanced the ability to render photorealistic or artistic images, yet it remains limited in a key aspect of human creativity: interpreting ambiguous shapes. This phenomenon, rooted in pareidolia, allows humans to perceive meaningful forms in random patterns such as clouds, stones, or leaves. To computationally replicate this imaginative process, we introduce Visual Retrieval-Augmented Generation (Visual-RAG), a framework that generates animal art directly from natural silhouettes. Our method retrieves structurally similar animal shapes from a curated corpus of 28,586 high-quality silhouettes and uses them as reference exemplars to guide diffusion-based generation with ControlNet and IP-Adapter. Ablation studies confirm that shape Context with RANSAC provides the most accurate alignment, while removing shape standardization reduces the inlier ratio to just 13.4%, underscoring the importance of structural fidelity in Visual-RAG. A user study with 12 participants evaluated the outputs in terms of aesthetics, silhouette fidelity, and overall impression. Results reveal that while Visual-RAG provides plausible interpretations, challenges remain in achieving high perceptual impact. This work lays the foundation for computational pareidolia, showing how machines can contribute to the early stages of imaginative discovery.

Cuong Van Duc, *Tam Ta Dinh*, *Chinh Tran Duc* and *Hanh Nguyen Thi*

FLUID: Flow-Latent Unified Integration via Token Distillation for Expert Specialization in Multimodal Learning

ABSTRACT. Multimodal classification requires robust integration of visual and textual signals, yet common fusion strategies are brittle and vulnerable to modality-specific noise. In this paper, we present FLUID-Flow-Latent Unified Integration via Token Distillation for Expert Specialization, a principled token-level pipeline that improves cross-modal robustness and scalability. FLUID contributes three core elements: (1) Q-transforms, learnable query tokens that distill and retain salient token-level features from modality-specific backbones; (2) a two-stage fusion scheme that enforces cross-modal consistency via contrastive alignment and then performs adaptive, task-aware fusion through a gating mechanism and a Q-bottleneck that selectively compresses information for downstream reasoning; and (3) a lightweight, load-balanced Mixture-of-Experts at prediction time that enables efficient specialization to diverse semantic patterns. Extensive experiments demonstrate that FLUID attains 91% accuracy on the GLAMI-1M benchmark, significantly outperforming prior baselines and exhibiting strong resilience to label noise, long-tail class imbalance, and semantic heterogeneity. Targeted ablation studies corroborate both

the individual and synergistic benefits of the proposed components, positioning FLUID as a scalable, noise-resilient solution for multimodal product classification.

Saikiran Korla, *Sadwik Gummadavelli*, *Trung-Nghia Le* and *Tam V. Nguyen*

Research Paper Quality Recognition Through Textual Feature Analysis

ABSTRACT. Knowledge and innovations are shaped by using the quality and credibility of the scientific research. Yet, distinguishing between impactful, high-quality work and flawed studies remains a challenge. This paper introduces a benchmark for classifying research papers into two categories: good (highly cited) and non-good (retracted), using only textual features from titles and abstracts. We evaluate multiple embedding techniques, including SBERT, Word2Vec, FastText, USE, and TF-IDF, combined with classifiers such as Support Vector Machines (SVM), Random Forests, and Neural Networks. Our contributions include: (1) hyperparameter transparency, (2) feature space visualizations using t-SNE, (3) model interpretability analysis with SHAP, and (4) detailed examination of error cases. Experimental results show that a neural network with SBERT embeddings achieves 87.22% accuracy, while FastText combined with SVM reaches 91.12%. These findings highlight the value of textual information in assessing research quality, with ethical considerations for deployment. This work contributes toward the development of academic integrity tools that promote trustworthy scholarship

Minh Duc Nguyen, *Thi Tam Ngo*, *Duc-Anh Nguyen*, *Oky Dicky Ardiansyah Prima* and *Ngoc Hoa Nguyen*

Efficient Probabilistic Cross-Modal Retrieval via Top-k Selection and Fast Embedding Learning

ABSTRACT. Image-Text Matching (ITM) is a core task in vision-language research, enabling cross-modal retrieval and zero-shot classification. While deterministic embedding methods map inputs to fixed vectors in a shared space, they often struggle to capture the semantic diversity inherent in multimodal data. Probabilistic embeddings offer a more expressive alternative by modeling inputs as distributions, but existing approaches face challenges in feature selection and training efficiency. In this work, we propose FAST-PCME, a novel probabilistic ITM framework that enhances the PCME architecture with two key innovations: (i) a top-k token selection strategy that filters out less informative features before pooling, and (ii) a fast probabilistic embedding learning mechanism that reformulates the matching objective for accelerated convergence. Extensive experiments on ECCV Caption, CxC, and MS COCO benchmarks demonstrate that FAST-PCME achieves state-of-the-art performance while reducing training time and improving semantic expressiveness.

Van-Bang Tran and *Thi Lan Le*

Text-Based Person Search in Low-Resource Scenarios

ABSTRACT. Text-based person search (TBPS) that aims to identify in- dividuals from natural language descriptions has many potential appli- caitons in surveillance, security, and forensics. While impressive results have been achieved with large-scale annotated training data, performance in low-resource scenarios where only a few text–image pairs are available, or only images exist, remains a significant challenge. This paper aims at addressing text-based person search (TBPS) in low-resource scenarios where the number of image-text pair is limited. To address the scarcity of high-quality image–text pairs in TBPS, we integrate a new module: Description generation module into an existing framework TBPS-CLIP [1]. Three scenarios have been proposed for description generation mod- ule: Scenario 1: Direct Multimodal Generation, Scenario 2: Captioning via Fine-tuning, and Scenario 3: VQA-Based Generation. Scenarios 2 and 3 require a small number of image–text pairs to fine-tune the mod- els, whereas Scenario 1 operates without image-text pairs. Furthermore, to better guide the description generation module, we design a new set of questions.

Extensive experiments on the CUHK-PEDES dataset demon- strate that our method achieves strong retrieval performance, reaching 71.60% in R@1, 90.66% in R@5, and 95.25% in R@10, even with only a limited number of image–text pairs for fine-tuning the description gener- ation module. Moreover, in the best case, the proposed question set yields a 16.1% improvement in the R@1 metric over the baseline question list.

Hoàng Bách Nguyễn Phan, *Minh Nghĩa Lê* and *An Trần Đức*

GigaCount: Enhancing Crowd Counting by Integrating a Multi-Scale Feature Fusion Model into CLIP-EBC

ABSTRACT. Crowd counting has emerged as a vital task in computer vision, driving applications ranging from urban planning to public safety. Despite advances, challenges remain in handling diverse crowd scenarios, such as low-light scenes, distorted human figures, and extremely dense crowds. To handle these problems, we propose GigaCount, a multi-scale vision-language model that leverages Contrastive Language-Image Pretraining with Enhanced Blockwise Classification to enhance crowd counting performance. Our approach integrates ConvNeXt and its multi-scale feature fusion capabilities into CLIP, further addressing key challenges in crowd analysis. We evaluate the model's effectiveness by analyzing density maps and conducting ablation studies, which reveal patterns in prediction errors and their underlying causes. These findings guide targeted enhancements, including data augmentation to boost robustness across diverse lighting conditions, loss function adjustments to enhance accuracy in dense scenes, and layer removal to minimize model size and computational cost. Achieving a competitive MAE of 103.3, our model introduces a novel, lightweight architecture that integrates multi-scale feature fusion into CLIP's image encoder. Although it does not outperform state-of-the-art methods, this approach surpasses almost all conventional CNN-based techniques, underscoring the potential of multiscale visionlanguage models in crowd analysis and laying a foundation for further advancements. The implementation is available at https://github.com/AdamHermes/GigaCount

Nguyen Minh Kien, *Ma Thi Chau*, *Vu Thanh Long*, *Nguyen Tien Phuc*, *Hoang Quoc Viet* and *Dinh Quang Trung*

Integrating Motion-based Technique and Deep Learning for Expression Analysis in Vietnamese Traditional Chèo

ABSTRACT. Due to the lack of specific datasets and studies on expression in Vietnam's traditional Chèo theatre, this paper presents a framework that integrates motion-based preprocessing with deep learning to analyze those expressions within this art form. We utilize Eulerian Video Magnification (EVM) and dense optical flow techniques to automatically detect and segment subtle facial movements, resulting in a dataset of 7,166 expression segments. Of these, 3,353 apex frames are manually annotated by experts at the Hanoi Academy of Theatre and Cinema. Using this dataset, we evaluate several popular convolutional and transformer-based architectures under transfer-learning settings. Among them, VGGFace achieves the highest accuracy (81.15%) and Cohen's kappa score (0.703), closely followed by ResNet18 (80.77% accuracy and 0.6993 kappa). These results demonstrate the effectiveness of motion-based extraction in the challenging performing-arts context of Ch`eo and lay a foundation for future cultural-heritage preservation and the development of educational tools.

Thanh-Hai Nguyen, *Thinh-Phuc Nguyen*, *Gia-Huy Dinh*, *Lam-Huy Nguyen* and *Trung-Nghia Le*

VisionGuard: Synergistic Framework for Helmet Violation Detection

ABSTRACT. Enforcing helmet regulations among motorcyclists is essential for enhancing road safety and ensuring the effectiveness of traffic management systems.

However, automatic detection of helmet violations faces significant challenges due to environmental variability, camera angles, and inconsistencies in the data. These factors hinder reliable detection of motorcycles and riders and disrupt consistent object classification. To address these challenges, we propose VisionGuard, a synergistic multi-stage framework designed to overcome the limitations of frame-wise detectors, especially in scenarios with class imbalance and inconsistent annotations. VisionGuard integrates two key components: Adaptive Labeling and Contextual Expander modules. The Adaptive Labeling module is a tracking-based refinement technique that enhances classification consistency by leveraging a tracking algorithm to assign persistent labels across frames and correct misclassifications. The Contextual Expander module improves recall for underrepresented classes by generating virtual bounding boxes with appropriate confidence scores, effectively addressing the impact of data imbalance. Experimental results show that VisionGuard improves overall mAP by 3.1% compared to baseline detectors, demonstrating its effectiveness and potential for real-world deployment in traffic surveillance systems, ultimately promoting safety and regulatory compliance.

Duy-Dat Tran and *Trung-Nghia Le*

Edit3DGS: Unified Framework for Dynamic Head Editing via 2D Instruction-Guided Diffusion and 3D Gaussian Splatting

ABSTRACT. We present Edit3DGS, a unified framework for dynamic 3D head editing that integrates 2D instruction-guided diffusion with 3D Gaussian splatting. Unlike prior approaches that separately address frame-based edits or static 3D reconstruction, our method couples semantic controllability in the image domain with photorealistic, temporally consistent 3D representations. Given an input video, editable facial regions are masked and modified using a text-conditioned diffusion model to support fine-grained operations such as expression transformation, attribute modification, and appearance refinement. The edited frames are then aggregated through 3D Gaussian splatting to produce a coherent, high-fidelity avatar that preserves both identity and motion dynamics. To enforce consistency, Edit3DGS incorporates multi-view batch editing and lightweight inpainting strategies that recover lost expressions across timesteps. Experimental results demonstrate that our framework enables controllable, artifact-free head editing with smooth temporal transitions, offering practical applications in virtual avatars, immersive communication, film production, and interactive media.

Tien Le Thanh, *Huu-Tri Luu*, *Trong-Nguyen Ha*, *An Nguyen-Huu* and *Quang Dung-Cam*

VNProductKIE: A Dataset and Three-Stage Pipeline for Key Product Information Recognition on Vietnamese Packaging Labels

ABSTRACT. Food waste poses serious environmental and economic concerns, often worsened by the lack of accessible product information. Automated extraction from packaging labels offers a promising solution, yet existing datasets fall short in representing the linguistic and visual diversity found in the Vietnamese markets. This paper introduces VNProductKIE, a new dataset of high-resolution images capturing Vietnamese food and beverage packaging. It features both English and Vietnamese text, diacritic-rich scripts, local date formats, and real-world distortions such as blur, curvature, and clutter. To extract structured information, this paper proposes a three-stage pipeline including: (1) a YOLO11x-based detector for locating key regions (e.g., product name, weight, brand, expiration date), (2) a word-level detector for segmenting individual words, and (3) a VietOCR-based recognizer for transcription. The final output is structured into complete product metadata. In experiments on VNProductKIE, the pipeline achieved a word-level recognition accuracy of 98.85%, highlighting the effectiveness of the proposed approach.

Thai-Viet Dang, *Ngoc-Viet-Hoang Tran*, *Nhu-Nghia Bui*, *Xuan Tan Phan* and *Ngoc-Tam Bui*

MMCS: Multimodal Mamba Channel Switching for Object Detection via RGB-IR Fusion

ABSTRACT. Object detection in low-light conditions presents significant challenges due to the presence of noise and reduced contrast in conventional RGB images. Furthermore, optimizing the number of model parameters and computational efficiency remains problematic. The paper proposes MMCS, an efficient object detection framework that harnesses multimodal data derived from paired RGB and Infrared (IR) images. Environmental features are processed via two distinct streams, extracted using the Mamba backbone, and propagated through successive layers. To effectively integrate and prioritize salient information, channel switching spatial attention module blocks are incorporated, employing pooling and attention mechanisms. The refined features are subsequently forwarded to the decoder for final processing. The proposed model was evaluated on two benchmark datasets, LLVIP and FLIR. Experimental results demonstrate that MMCS outperforms existing approaches, achieving substantially higher accuracy, exhibiting a 1.3-fold increase in mean Average Precision (mAP) compared to YOLO-based models and other methods utilizing both RGB and IR modalities. Ultimately, the combination of the robust state-space modeling capabilities of Mamba with an intelligent multimodal information exploitation strategy enhances object recognition performance under varying environmental conditions.

Nhat-Huy Ho and *Minh-Triet Tran*

Balancing Quality, Speed, and Compactness of 3D Gaussian Splatting

ABSTRACT. 3D Gaussian Splatting has transformed the field of novel view synthesis by enabling real-time rendering of high-quality, photorealistic scenes. However, its practical application is often hindered by long training times and the large memory footprint of the resulting models. While methods like DashGaussian accelerate training and GaussianSpa creates compact models, they operate independently; GaussianSpa's sparsification comes at a significant training time cost, and accelerated methods like DashGaussian still produce large final models. The potential for a combined framework that integrates these approaches to efficiently tackle these bottlenecks remains unexplored. To address this, we first introduce a novel three-stage training schedule that integrates the coarse-to-fine acceleration of DashGaussian with the sparsity-enforcing framework of GaussianSpa. Our experiments demonstrate that this method achieves state-of-the-art model compactness, reducing model sizes by up to an order of magnitude with an acceptable trade-off in visual fidelity. Crucially, this compression is achieved efficiently, with training times significantly shorter than standalone sparsification methods. Furthermore, we introduce an experimental variant that replaces gradient-based densification with a direct sampling strategy to explore the limits of training acceleration. Our results show that this second approach achieves the fastest training times by a significant margin. Overall, our work presents two distinct solutions to key 3DGS bottlenecks: one that yields exceptionally compact models, and another that generates models in a fraction of the standard training time.

Duc-Tuan Luu, *Gia-Nghia Tran*, *Khoa Nguyen* and *Vinh-Tiep Nguyen*

OTGen-FSIS: Optimal Transport–Driven Feature Generation for Few-Shot Instance Segmentation

ABSTRACT. Few-shot instance segmentation (FSIS) extends the chal- lenges of few-shot object detection (FSOD) by requiring not only object localization but also precise pixel-level mask prediction for novel cate- gories with only a few labeled samples. This task is particularly difficult because limited supervision makes it hard to capture intra-

class vari- ations, and existing generative approaches often produce synthetic features that misalign with real distributions, resulting in degraded segmen- tation quality. To overcome these limitations, we propose OTGen-FSIS, an Optimal Transport–Driven Feature Generator for FSIS. Our approach introduces a conditional generator trained with an OT-based loss and clustering, enabling the synthesis of diverse and representative features for novel classes by leveraging variations from base classes. Unlike prior work, our method relies solely on an unsupervised OT loss to optimize the generator. This design not only stabilizes generator learning but also offers flexibility for future extensions to tasks without labeled samples. By capturing global geometric relationships between distributions, op- timal transport (OT) loss provides stronger alignment than point-wise losses, reducing distributional mismatch and improving generalization to novel categories. Extensive experiments on standard FSIS benchmarks demonstrate that OTGen-FSIS significantly enhances novel class seg- mentation while maintaining strong performance on base classes, vali- dating the effectiveness of OT-based distribution matching in few-shot instance segmentation

Quynh-Trang Pham Thi

DAKTA: Directional Kolmogorov-Arnold Classifier for Task Arithmetic in Continual Learning

ABSTRACT. Continual learning requires models to acquire new knowledge while preserving previously learned information-a fundamental challenge known as the stability-plasticity tradeoff. While recent advances in model compositionality through task arithmetic show promise, existing approaches primarily rely on linear classifiers that struggle to maintain stable classification spaces during incremental learning. In this paper, we propose Directional Arithmetic Kolmogorov Task Architecture(DAKTA), a strategy that addresses both compositionality and classification stability in continual learning. DAKTA leverages second-order Taylor approximation theory to ensure task vectors remain within the pre-training basin, enabling effective model composition for enhanced stability. Our key innovation lies in replacing conventional linear classifiers with learnable Gaussian Radial Basis Functions (RBF) that provide selective channel activation, enhanced locality properties, and significantly reduced interference between tasks, thus boosting plasticity capabilities. Furthermore, we proposed a novel directional logit fusion mechanism that combines RBF-based magnitude information with feature directional cues, enabling the classifier to capture both local similarity patterns and global directional relationships for more robust class discrimination. Experiments on ImageNet-R, CUB-200, and CIFAR-100 demonstrate significant improvements of DAKTA compared to state-of-the-art methods.

Thi Thu Hien Trinh and *Trung-Nghia Le*

CIAN: Multi-Stage Framework for Event-Enriched Image Captioning via Retrieval-Augmented Generation

ABSTRACT. Event-enriched image Captioning seeks to generate descriptions that convey not only what is visible in an image but also the broader context surrounding the depicted event. However, most existing models remain limited to pixel-level information and fail to integrate non-visual knowledge such as timing, location, or participants. To address this limitation, we propose the Contextual Image-Article Narrator (CIAN), a multi-stage framework that enriches image captions with external contextual narratives. CIAN first employs a Context Retrieval leveraging the SigLIP model to retrieve semantically relevant articles for each query image. The retrieved textual content is then summarized and used to guide a Narrative Generation stage, where a LoRA-fine-tuned Qwen model produces an event-aware caption. Finally, an N-Gram-based Refinement step enhances linguistic fluency and domain-specific coherence. Evaluated on the OpenEvents-V1 benchmark, CIAN achieves strong

retrieval performance (mAP of 0.979) and improves captioning quality, boosting the CIDEr score from 0.030 to 0.094 after refinement. These results demonstrate the effectiveness of combining retrieval-augmented reasoning with progressive linguistic refinement, advancing the development of AI systems capable of generating holistic, human-like visual narratives.

Luong Ho, *Hao Do*, *Duc Chau* and *Minh Nguyen*

Improving Code-Switching Speech Synthesis via Concatenated Tokenizers

ABSTRACT. Code-switching text-to-speech (CS-TTS) remains a challenging task due to phonetic ambiguity, orthographic overlap, and prosodic discontinuity between languages. Existing approaches have not fully addressed the issue of cross-lingual interference, often resulting in degraded pronunciation accuracy and unnatural transitions. In this work, we draw inspiration from the concatenated tokenizer strategy originally introduced for code-switching automatic speech recognition (ASR) and adapt it to the CS-TTS setting. The core idea is to assign each language its own tokenizer and allocate disjoint token identifier spaces by shifting the token IDs of the secondary language by an offset equal to the vocabulary size of the primary language. This design ensures that even visually identical graphemes are mapped to distinct token IDs according to their language origin, thereby encoding both character form and language identity directly at the token level. By eliminating cross-lingual ambiguity in the text encoding stage, the model learns language-specific pronunciation rules while maintaining the ability to generate smooth and natural cross-lingual speech. Experimental evaluations, including subjective mean opinion scores and objective metrics demonstrate that our method significantly improves pronunciation accuracy, naturalness, and the smoothness of cross-lingual transitions. To the best of our knowledge, this is the first application of the concatenated tokenizer paradigm to CS-TTS, providing a simple yet principled solution to cross-lingual interference in multilingual speech synthesis.

Phat Nguyen Cuong and *Khang Nguyen Tan Tran Minh*

Impact of Foggy Weather on Anomaly Detection in Aerial Traffic Surveillance Videos: An In-Depth Analysis

ABSTRACT. Video anomaly detection is vital for intelligent traffic surveillance systems to enhance public safety and security. Despite significant advancements, current methods struggle with rough weather conditions, particularly fog, which degrades image quality and visibility, complicating the identification of abnormal events. This study investigates the impact of fog on anomaly detection performance in aerial traffic surveillance videos. We utilize the DW-GAN algorithm to generate realistic fog scenarios on two benchmark datasets, including UIT-ADrone and Drone-Anomaly. Extensive experiments are conducted with six state-of-the-art anomaly detection methods to evaluate their performance under foggy conditions. Our results reveal significant performance degradation across all methods, highlighting the challenges posed by fog. Additionally, we evaluate the impact of data preprocessing by testing model performance on dehazed datasets by using GridFormer, demonstrating the advantage of dehazing as a preprocessing step for anomaly detection under foggy conditions. Furthermore, we perform in-depth analyses to identify the practical limitations of current approaches and discuss potential directions for future research. Our findings contribute to a better understanding of anomaly detection in adverse weather and provide insights to develop more robust models. The datasets and source code are available online at \url{https://github.com/PhatNC/AnomalyDetection}.

Toan-Thinh Truong, *Hai-Yen Vong*, *Thanh-Ha Ung-Dung*, *Minh-Triet Tran* and *Anh-Duc Duong*

Lightweight digital signature algorithms based on linear public-key

ABSTRACT. Digital signatures represent one of the most widely adopted applications

of public-key cryptography. Over the years, numerous schemes have been introduced, underscoring the importance and continued relevance of research in this field. As a cornerstone of modern security infrastructures, digital signatures support a wide range of applications, including secure economic systems, financial transactions, and national security frameworks. Consequently, a comprehensive understanding and effective implementation of these schemes are essential not only for cryptography specialists but also for the broader user community. Classical signature algorithms such as RSA and ElGamal have long served as foundational cryptographic solutions. Over time, advanced models such as blind signatures and ring signatures have emerged, extending their applicability across diverse domains. Moreover, the adoption of digital signatures in cloud-based environments has grown significantly in recent years. Among current approaches, schemes that utilize public-keys of the form k + r × p have attracted considerable attention due to their efficiency and structural simplicity. Building on this trend, the present work introduces a lightweight digital signature algorithm grounded in this principle, aimed at addressing the constraints of resource-limited environments. In addition to detailing the proposed construction, we provide a thorough security analysis and performance evaluation, comparing it with existing solutions to demonstrate its practicality and effectiveness.

Nguyen Hoang Cao, *Hoang Bui Le*, *Nam Vo Hoang* and *Trung-Nghia Le*

Exploring Multi-Modal Large Language Models and Two-Stage Fine-Tuning for Fashion Image Retrieval

ABSTRACT. Composed Image Retrieval (CIR) retrieves a target image using a composed query of a reference image and a modified text description. In the fashion domain, this task requires understanding subtle attribute variations such as color, pattern, and texture. However, existing approaches face limitations due to scarce annotated data and simplistic negative sampling. We propose a novel framework that integrates a multi-modal large language model (LLaVA) to generate attribute-aware triplets and introduces a two-stage fine-tuning strategy to enhance contrastive learning. We leverage pretrained vision-language models, such as CLIP-ViT/B32, to generate and concatenate sentence-level prompts with the relative caption and to scale the number of negatives using static representations. Due to resource constraints, experiments were conducted on a representative subset of the dataset. Preliminary findings demonstrate improved compositional reasoning and fine-grained retrieval behavior, indicating the feasibility and potential of the proposed framework for future full-scale evaluation.

Thi-Huong Nguyen and *Van-Toi Nguyen*

A RGB-D Dataset of Isolated Vietnamese Sign Language

ABSTRACT. In this paper, we introduce ViSL120, the first large-scale multimodal dataset for Vietnamese Sign Language (ViSL). In contrast to previous tiny, RGB-only, or multi-view datasets, ViSL120 provides more than 50,000 videos in 120 glosses that were taken with a single Intel RealSense D435 camera in both RGB and depth. This dataset ensures easy data collection and deployment while offering rich indications from hands, bodies, and facial expressions. To demonstrate its utility, we establish benchmark results with state-of-the-art sign language recognition models, revealing both the challenges of ViSL and the potential for robust model development. ViSL120 enhances Vietnamese sign language resources, supports assistive technology for the deaf community, and advances the research community.

12:00-13:30	Lunch LOCATION: Feast Restaurant, 1F
13:30-15:30	**Session 10A: SOICT Technical Session XVII: Lifelog Event Retrieval** CHAIR: *Cathal Gurrin* LOCATION: Grand Ballroom A, 2F

13:30 *Duc-Nhuan Le*, *Hoang-Phuc Nguyen*, *Thanh-Duy Lam*, *Minh-Nhut Dang* and *Minh-Hoang Le*

U-CESE: Unified Clip-based Event Search Engine for AI Challenge HCMC 2025

ABSTRACT. Retrieving events from large-scale video datasets is challenging due to complex temporal, spatial, and multimodal information. This paper presents U-CESE, our solution for the AI Challenge HCMC 2025, a Unified Clip-based Event Search Engine for multimodal event retrieval across diverse video sources. Building on CESE, U-CESE integrates its three modules into a single cohesive framework, ensuring consistent processing and retrieval across query types. A core component is the Unified Clipping Algorithm, which merges separate clipping algorithms into one efficient pipeline. To handle large-scale data, we propose DAKE, a lightweight, training-free keyframe extraction method using JPEG file size variations to identify significant scene changes. Finally, we introduce ReCap, a temporally consistent captioning framework inspired by Recurrent Neural Network, generating detailed and context-aware textual descriptions. Experiments show that U-CESE delivers robust, consistent, and efficient performance in large-scale multimodal event retrieval.

13:50 *Van-Loc Nguyen*, *Gia-Huy Vuong*, *Ngoc-Do Tran*, *Tien-Thanh Nguyen-Dang*, *Van-Son Ho*, *Van-Tu Ninh* and *Minh-Triet Tran*

Visionary: Optimized Temporal Video Retrieval via Large Language Model-Enhanced Query Processing

ABSTRACT. The rapid growth of video content necessitates efficient, real-time event retrieval systems. Addressing the Ho Chi Minh City AI Challenge 2025, we present Visionary, a new generation of the NewsInsight systems. Our system introduces four key contributions: (1) a novel adaptive keyframe extraction algorithm; (2) an enhanced pre-processing pipeline using the Qwen3-VL model for metadata generation and integrated optical character recognition; (3) a flexible architecture supporting multiple embedding models; and (4) the use of Reciprocal Rank Fusion to synthesize retrieval results. These enhancements aim to substantially improve retrieval accuracy and overall performance for complex, large-scale video retrieval tasks.

14:10 *Khoa Dinh Duc Anh*, *Duc-Tai Dinh*, *Trung Nguyen Le Hoang* and *Nhan Nguyen Thanh*

KPTER: K-Pointer for Temporal Event Retrieval

ABSTRACT. The explosive proliferation of online video content necessi- tates advanced retrieval systems, as promoted by the Ho Chi Minh AI Challenge (AIC) 2025. This competition comprises three tasks: Known- item Search (KIS), Visual Question Answering (VQA), and the newly introduced Temporal Retrieval and Alignment of Key Events (TRAKE). To address these challenges, we propose a comprehensive multimodal retrieval framework. Our system integrates heterogeneous data sources, including semantic embeddings from CLIP and BEIT-3 , Optical Char- acter Recognition (OCR), Automatic Speech Recognition (ASR), and open-vocabulary object detection. The architecture features a dual-layer search mechanism. For discrete queries (KIS, VQA), a Multi-modal Search layer retrieves and ranks results from parallel data streams using the Weighted Reciprocal Rank Fusion (WRRF) algorithm. To address the sequential nature of the TRAKE task, we introduce a novel Temporal Search module built upon an efficient K-pointer sequential re-ranking algorithm. This algorithm effectively validates and ranks video segments containing events in a specified temporal order. Finally, we present a competition-oriented user interface designed for real-time interaction, supporting multi-stage query construction and precise temporal refine- ment tools.

14:30 *Huu An Vu*, *Van Khanh Mai*, *Trong Tam Nguyen*, *Quang Duc Dam*, *Tien Huy Nguyen* and *Thanh Huong Le*

MADTempo: An Interactive System for Multi-Event Temporal Video Retrieval with Query Augmentation

ABSTRACT. The rapid expansion of video content across online platforms has intensified the need for retrieval systems capable of understanding not only isolated visual moments but also the temporal structure of com- plex events. Existing approaches often fall short in modeling temporal dependencies across multiple events and in handling queries that ref- erence unseen or rare visual concepts. To address these challenges, we introduce MADTempo, a video retrieval framework that unifies tempo- ral search with web-scale visual grounding. Our temporal search mech- anism captures event-level continuity by aggregating similarity scores across sequential video segments, enabling coherent retrieval of multi- event queries. Complementarily, a Google Image Search–based fallback module expands query representations with external web imagery, ef- fectively bridging gaps in pretrained visual embeddings and improving robustness against out-of-distribution (OOD) queries. Together, these components advance the temporal reasoning and generalization capabil- ities of modern video retrieval systems, paving the way for more seman- tically aware and adaptive retrieval across large-scale video corpora.

14:50 *Duong Nghia*, *Nguyen Tu*, *Pham Nhan*, *Le Truong* and *Le Khoi*

Althena-Vision: Adaptive Temporal Multimodal Event Retrieval with LLM-generated Multiperspective Fusion

ABSTRACT. Event retrieval in large-scale video collections remains a chal- lenging task due to the complexity of multimodal content and the seman- tic gap between user queries and visual data. In this paper, we present AIthena-Vision, our entry for the AI Challenge HCMC 2025, an adaptive system to retrieve large-scale video events designed to effectively address these challenges. The core of our system combines a state-of-the-art Per- ception Encoder (a CLIP variant) for superior text–visual alignment with an LLM-driven multiperspective query expansion to improve retrieval ac- curacy. Furthermore, the system incorporates adaptive temporal search and integrates multiple sources of multimodal evidence, including OCR, ASR, and object detection. With an interactive interface that supports collaborative searching, AIthena-Vision offers a competitive and effective retrieval solution. In our latest evaluation with the competition ground truth, the system achieved 92% precision, outperforming our previous model - which obtained the highest score among all teams in the third preliminary round last year - with 81% precision, demonstrating signifi- cant improvement and robustness.

15:10 *Duy Ho Khanh*, *Diep Tran Van*, *Sơn Nguyễn Hồng*, *Duy Nguyen*, *Hữu Trần Chí* and *Binh Nguyen*

Lucifer-TRACE: Dynamic Programming and LVLM-Aided Verification for Event-Based Video Retrieval

ABSTRACT. Retrieving complex real-world events from large-scale videos recorded has been one of the most fundamental yet underexplored chal- lenges. Most existing systems still treat videos as collections of indepen- dent frames, overlooking the temporal continuity and semantic coherence that define real events. As a result, they often fail to align fragmented ev- idence across time or verify whether the retrieved segments truly match the user's intent. We propose Lucifer-TRACE, a multi-modal event re- trieval framework that unifies structured temporal reasoning with early semantic validation. At its core, the Soft Temporal Search mechanism employs dynamic programming to link semantically related yet tempo- rally scattered frames into coherent event chains. In order to complement this, we present an Early Semantic Verification (ESV) module, leverag- ing a large vision–language model (LVLM) to assess the semantic cor- rectness of candidate segments, providing interpretable feedback without altering retrieval scores. By integrating visual and textual cues in a uni- fied pipeline, Lucifer-TRACE achieves both precision and transparency. While evaluated in the Ho Chi Minh City AI Challenge 2025, our system can obtain a score of 86/88,

ranking among the top-performing teams. These results highlight the potential of combining dynamic program- ming–based temporal alignment with lightweight VLM-aided verification for robust, interactive event retrieval.

13:30-15:30 **Session 10B: SOICT Technical Session XVIII: AI Applications**
CHAIRS: *Van Khu Vu* and *Khoat Than*
LOCATION: Grand Ballroom B, 2F

13:30 *Dam Thai Ninh*, *Nguyen Duc Kien*, *Trinh Ngoc Huynh*, *Dinh Tran Hiep*, *Nguyen Hai Anh*, *Bui Duc Manh*, *Miguel D'Haeseleer*, *Jeroen Van Schependom*, *Stijn Denissen*, *Tran Quoc Long*, *Nguyen Linh Trung* and *Guy Nagels*

The Privacy–Utility Trade-off in Brain MRI Synthesis: A Comparative Framework for Generative Models

ABSTRACT. Generative Adversarial Networks (GANs) showed significant potential for synthesizing realistic brain MRI scans. However, this capability introduced a critical risk of sensitive patient data leakage due to sample memorization. This study evaluated data leakage mitigation techniques for GANs applied to brain MRI synthesis. We assessed a standard Deep Convolutional GAN (DCGAN) model and its variants incorporating four distinct mitigation strategies: Maximum Entropy GAN (MEGAN), Spectral Normalization (SN), Wasserstein GAN (WGAN) and Differential Privacy (DP). The models' efficacy was assessed based on their ability to reduce memorization and maintain image quality. We revealed that both MEGAN-SN and WGAN-SN-D provide an optimal balance, significantly reducing privacy risks while maintaining acceptable image quality. Conversely, DPGANs substantially compromised image quality to achieve their strong theoretical privacy guarantees.

13:50 *Bao Bui-Quoc*, *Khang Nguyen-Vi*, *Hoa Nguyen-Phuong* and *Nidal Kamel*

Task-Aware Harmonization of Sentinel-2 for Canopy Height Mapping: A Deep Learning Application in the Ngoc Linh Mountains, Vietnam

ABSTRACT. Accurate mapping of forest canopy height is essential for biomass estimation, carbon accounting, and long-term forest monitoring. However, the heterogeneous spatial resolutions of Sentinel-2 imagery ($10 \mathrm{~m}, 20 \mathrm{~m}$, and 60 m bands) present significant challenges for reliable canopy height estimation. Conventional approaches typically decouple the task into two stages-superresolution followed by regression-which often introduces error propagation and reduces accuracy. In this study, we propose an end-to-end deep learning framework that jointly performs resolution harmonization and canopy height regression. The model incorporates frequency-enhanced residual blocks and channel attention mechanisms to align all Sentinel-2 bands to a uniform 10 m resolution while extracting task-specific spectral-spatial features. Experiments conducted over the Ngoc Linh Mountains in Central Vietnam demonstrate that the proposed method achieves a mean absolute error of $7.061 \pm 0.110 \mathrm{~m}$ and a root mean square error of $9.175 \pm 0.155 \mathrm{~m}$, outperforming existing baselines. Qualitative analyses further confirm robust canopy structure reconstruction with low prediction uncertainty (STD < 15). These results highlight that integrating harmonization and regression into a unified architecture leads to more accurate and stable canopy height predictions in complex tropical forest landscapes.

14:10 *Thanh Tam Tran*, *Ba Hung Ngo*, *Thu Thuan Pham* and *Tae Jong Choi*

Adaptive Multi-Level Attention for Effective Cross-Domain Brain Tumor Detection

ABSTRACT. We present an innovative approach for unsupervised domain adaptation (UDA) in the classification of brain tumors, Adaptive Multi-Level Attention (AMLA) to confront the UDA challenges of domain shift in medical imaging datasets. AMLA solves the brain tumor classification problem across domains without target domain labeled data through the use of Efficient Channel Attention (ECA) and Dual Self-Attention (DSA) mechanisms. While ECA operates at the early stages of the network,

DSA which incorporates Spatial Attention Module (SAM) and Channel Attention Module (CAM) applies to the final network stage. \texttt{ECA} applies to the early network stage since it captures low-level features in an efficient manner. On the other hand, DSA applies to the final stage captures long-range spatial and channel interactions and is important for the separation of complicated tumor types. This stage-specific adaptation strikes a balance between cost-effectiveness and feature expressiveness, which is important in alleviating overfitting when performing UDA. Experimental results over three brain tumor datasets demonstrate that the target test accuracy of our AMLA outperforms previous UDA methods. Our backbone-agnostic approach ensures robustness and scalability for medical imaging applications.

14:30 *Leonhard Bürkner* and *Markus Westner*

Critical Success Factors for AI Adoption: A Multivocal Literature Review and a Top Management Perspective

ABSTRACT. The strategic adoption of Artificial Intelligence (AI) is a critical determinant of competitive advantage, yet organizations face high failure rates. This paper aims to identify and categorize the Critical Success Factors (CSFs) that influence suc-cessful AI adoption by synthesizing academic and practitioner knowledge from a top management perspective. We conducted a multivocal literature review of 57 academic and 20 practitioner sources, screened per PRISMA. From vote-count synthesis we identified 16 CSFs grouped via the Technology-Organization-Environment (TOE) framework, with organizational factors dominating such as leadership support, AI literacy, and cultural readiness. Practitioner evidence corroborates most academic CSFs and adds implementation-centric aspects (e.g., AI scalability, use-case–value alignment, partnerships). Implications for top management/Chief Information Officers include prioritizing data governance, change capability, and portfolio-level value realization. Scholarly, the paper validates the TOE framework's continued relevance for AI while highlighting the amplified importance of its organizational dimension.

14:50 *Thi Quynh Hoa Nguyen*, *Duy Hai Nguyen*, *Tuan Long Tran*, *Thi Trang Tran* and *Van Khoa Le*

A Computational Framework for the Personalized Remediation of Reading Difficulties Using Dynamic Bayesian Networks

ABSTRACT. Reading disorders in children pose a significant clinical challenge, with current interventions often limited by their reliance on static assessments. These approaches fail to capture the dynamic nature of skill acquisition and the hierarchical dependencies between cognitive domains that underlie reading ability. This study develops an intervention approach based on the Dynamic Bayesian Network (DBN), a computational model that represents skill acquisition over time, as a diagnostic tool. By mapping the relationships between underlying language abilities, DBNs aim to identify core deficits that lead to reading failure. We conducted an experimental study with two children with word decoding difficulties. The DBN model was used not only to assess performance but also to create a diagnostic profile of each child's specific underlying weaknesses. This profile is then used to develop interventions tailored to the nature of the child's intrinsic difficulties. The positive results demonstrate the effectiveness of this targeted intervention, opening up promising new opportunities for individualized remediation of reading disorders.

15:10 *Quang Nguyen* and *Khang Nguyen*

Towards Reliable Oriented Surgical Instrument Detection: Benchmark and Evaluation

ABSTRACT. Accurate detection of surgical instruments is crucial for computer- and robotic-assisted minimally invasive surgery (MIS). Segmentation-based methods provide pixel-level localization but are computationally demanding for real-time use,

while axis-aligned detection cannot capture the orientation of elongated and articulated tools such as threads and suturing needles. To address these limitations, we adapt the SAR-RARP50 dataset by converting segmentation masks into oriented bounding box annotations, establishing the first benchmark for oriented detection in robotic surgery. Using this benchmark, we evaluate ten state-of-the-art detectors. Experimental results show that YOLO-based single-stage models achieve the highest mean Average Precision, with YOLOv9 reaching 81.2%, while rotation-aware architectures such as Rotated RetinaNet and SASM achieve the most accurate orientation predictions, with 9.9° error and 92.1% accuracy, and 14.0° error and 86.3% accuracy respectively. These findings highlight a trade-off between real-time detection robustness and orientation precision, providing a foundation for hybrid designs, class-balanced strategies, and the development of reliable perception systems for surgical robotics.

13:30-15:30 **Session 10C: SOICT Technical Session XIX: Recent Advances in Cyber Security**

CHAIR: *Hai Anh Tran*

LOCATION: Yersin Ballroom A, 2F

13:30 *Minh Tran Dang Quang*, *Tung Bui*, *Tran Dinh Kien Giang*, *Tran Quang Duc* and *Tuyen Ngoc Le*

Robust Intrusion Detection and Classification in EVSE Using Ensemble Methods

ABSTRACT. This paper presents a novel machine learning-based approach for intrusion detection and classification in Electric Vehicle Supply Equipment (EVSE). Focusing exclusively on network traffic data, we explore how ensemble learning methods can enhance threat detection in both binary and multi-class classification tasks. By leveraging a combination of classifiers such as K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost), our approach demonstrates improved accuracy and robustness over individual models. The study evaluates multiple ensemble strategies, including majority voting and soft voting, to identify the most effective techniques for real-time threat classification in EVSE environments. Evaluated on the CICEVSE2024 dataset, our approach achieves exceptional performance, with 99.93% accuracy in binary classification and 99.57% accuracy in multi-class classification. Our findings contribute to the growing field of intelligent cybersecurity solutions in electric mobility systems, highlighting the critical role of network-level machine learning analysis in protecting EVSE against evolving cyber threats.

13:50 *Hung Nguyen Tuan*, *Kiet Le Tuan*, *Nguyen Nguyen Trung*, *Luong Ho Nguyen*, *Duy Vu Ba* and *Hoa Nguyen Ngoc*

FOAMI: Enhancing ICS Threat Detection via Feature Optimization, Realistic Augmentation, and Mutual Inference

ABSTRACT. Industrial Control Systems (ICS) are now encountering a multitude of advanced cyber dangers, particularly those associated with the Internet. However, the present study continues to delineate several research gaps, including the restricted sample size, the subpar quality of the training data, and the inability to capitalize on the advantages of various AI models. This study introduces a novel ICS threat detection system, named FOAMI, which incorporates feature optimization, realistic data augmentation techniques, and mutual inference. Our FOAMI facilitates superior feature performance, increased layer separation by targeted aggregation, and overall learning enhancement via mutual inference. Extensive experimental findings on the standard industrial dataset IEC 60870-5-104 indicate that our approach significantly enhances detection accuracy, attaining a detection rate of 89.00% and decreasing the false alarm rate to only 0.99%, outperforming state-of-the-art sophisticated approaches.

14:10 *Vu Minh Manh*, *Nguyen Thanh Chung* and *Cho Do Xuan*

A Novel Framework for Android Malware Detection Based on Function Call Graph Pruning and Contrastive Learning

ABSTRACT. The widespread popularity of the Android operating system, along with its open-source nature, has made it a prime target for malware attacks. Given the rapid evolution and increasing sophistication of Android malware, developing effective detection techniques remains a critical challenge. Recently, many studies have explored the potential of function call graphs (FCGs) combined with graph neural networks for Android malware detection. However, existing approaches still face two main limitations: (1) the FCGs of Android applications are often extremely large, leading to high computational costs and reduced effectiveness in representation learning; and (2) the embedding spaces generated by graph representation learning models often lack clear separability between benign and malicious applications, which negatively impacts classification performance. To address these limitations, we propose PruCLDroid, a novel Android malware detection framework with two key contributions. First, we introduce a graph pruning technique based on structural similarity between node pairs to remove less significant edges, thereby simplifying the graph while preserving essential call relationships. Second, we employ a contrastive learning strategy to optimize the embedding space by pulling together representations of samples from the same class and pushing apart those from different classes. Experimental results demonstrate that PruCLDroid consistently outperforms existing methods across all evaluation metrics.

14:30 *Huynh Minh Hien*, *Ngo Trung Hieu*, *Nguyen Huu Quyen*, *Pham Van-Hau*, *Do Thi Thu Hien* and *Phan The Duy*

MPPO-GEM: Reinforcement Learning Approach for Generating Evasive Malware against Static and Dynamic Malware Detectors

ABSTRACT. Machine learning(ML)–based malware detectors are widely adopted in both research and practice. However, they remain vulnerable to adversarial attacks in both static and dynamic settings. In the static case, simple semantic-preserving edits (e.g., byte padding, section manipulation, control-flow redividing) can alter model features without breaking executability. In dynamic settings, attackers may evade sandboxing by delaying payload execution, simulating benign API interactions, or employing anti-debug/user-interaction techniques. To analyze these challenges in depth, we propose MPPO-GEM, a MaskablePPO-based framework that jointly applies problem-space primitives for generating adversarial malware, including call-based redividing, section-level modifications, semantic NOP insertion, user-interaction simulation, and anti-debugging bypasses within a constrained, action-masked Reinforcement Learning(RL) environment. The generated adversarial samples are evaluated across multiple platforms (VirusTotal, Kaspersky static/dynamic, MalConv, LightGBM). Empirical results show our method nearly doubled the evasion rate compared with the original malware while preserving executable integrity.

14:50 *Xuan Hung Truong*, *The Dung Luong* and *Anh Tu Tran*

Pri-WeDec: A Private Deep Learning Approach for Weapon Detection in Digital Forensics

ABSTRACT. Modern digital forensic investigations increasingly rely on Artificial Intelligence (AI) tools to screen and analyze vast volumes of image-based evidence. However, uploading this sensitive data to cloud systems or third-party servers for analysis poses significant challenges to privacy, data security, and the integrity of the chain of custody. To address this issue, we propose Pri-WeDec, a novel framework that enables the detection of weapon imagery directly on encrypted data, ensuring that the original content of the evidence is never exposed to untrusted environments. Our solution integrates Fully Homomorphic Encryption (FHE) with a specially customized

Convolutional Neural Network (CNN). Specifically, we utilize the CKKS encryption scheme, which supports real-number arithmetic, to encrypt the images. Subsequently, a custom-designed "FHE-friendly" CNN—which employs polynomial activation functions in place of ReLU and utilizes Average Pooling layers—performs inference directly on these ciphertexts. Experimental results show that our model achieves high accuracy on encrypted data, demonstrating the feasibility of performing complex forensic image analysis securely and with full privacy. This work opens a new avenue for the development of next-generation digital forensic tools, where the efficiency of AI can be leveraged without sacrificing core security principles.

15:10 *Nhat Duy Dang*, *Linh Giang Nguyen*, *Dong Le Van* and *Tuyen Ngoc Le*

Few-Shot Intrusion Detection using Model-Agnostic Meta-Learning with Deep Neural Networks

ABSTRACT. Nowadays, Intrusion Detection Systems (IDS) have to deal with dynamic and evolving cyberattacks. Albeit, traditional methods often fail to generalize to new threats because of the need for a large amount of labeled data. This study applies Model-Agnostic Meta-Learning (MAML) to a few-shot intrusion detection scenario. We developed a MAML-based multilayer perceptron (MAML-MLP) and experimented with it using the NSL-KDD dataset and CICIDS2017 dataset, focusing on fast adaptation to new attack classes and a few labeled samples. In this paper, comparative experiments of MAML-MLP and baseline deep learning models, including supervised MLP, LSTM, GRU, 1D-CNN, and CNN-LSTM, are presented to demonstrate MAML has superior and stable performance in the few-shot setting. We conclude that in a few-shot scenario, MAML is an optimal solution for intrusion detection to quickly respond to new and emerging cyber threats with less supervision. This model archive an average result of 90.62% and 95.89% accuracy on completely unseen attack type in CICIDS2017 and NSL-KDD dataset, respectively.

13:30-15:30 Session 10D: SOICT Technical Session XX: Multimedia Processing

CHAIR: *Tho Quan*

LOCATION: Yersin Balroom B, 2F

13:30 *Quynh Vo*, *Dung Phan* and *Tho Quan*

Scene Graph for Vietnamese Video Understanding: An Agentic Approach with Reasoning

ABSTRACT. Vietnamese Video Understanding (VVU) requires not only recognizing objects and actions but also reasoning about their interactions over time. A central approach is Video Scene Graph Generation (VSGG), yet traditional methods are prohibitively costly, demanding dense frame-level annotations of object–relation pairs and language-specific supervision—constraints infeasible for low-resource Vietnamese. Even modern VSGG pipelines and Vision–Language Models (VLMs) often rely on benchmark-specific fine-tuning or adapters, resulting in high computational/memory demands and brittleness under context shifts. These challenges motivate zero-shot, training-free approaches as a natural alternative. A key insight we emphasize is masking/segmentation before feeding frames into VLMs. Isolating objects and relevant background regions helps preserve scene information, reduce noise from irrelevant context, and avoid spurious correlations. While frequently overlooked, scene context is crucial for accurate relational reasoning. We present VISTA (Video Intelligence with Scene-graph Team-based Agents), a zero-shot multi-agent framework that decomposes video reasoning to reduce computational and memory load. VISTA orchestrates pre-trained components—frame selection, segmentation/masking, vision–language grounding, entity linking/refinement, and graph construction—via role-specialized agents connected through a lightweight shared memory. This design produces dynamic temporal scene graphs and answers questions without further training. By activating only query-relevant tools, pruning

candidates early, and reusing cached evidence, VISTA achieves cost-efficient compositional inference. To evaluate, we adapt part of the NEXT-QA dataset into Vietnamese and conduct real-world YouTube case studies. Results across graph quality, agent efficiency, and QA performance show that coordinated multi-agent decomposition can rival large-scale training, underscoring modularity and collaboration as practical pathways to advance video understanding in underrepresented languages.

13:50 *Quang-Linh Tran*, *Hoang-Bao Le*, *Tuong-Nghiem Diep*, *Binh Nguyen*, *Gareth J. F. Jones* and *Cathal Gurrin*

OpenLifelogQA: An Open-Ended Multimodal Lifelog Question-Answering Dataset

ABSTRACT. We introduce OpenLifelogQA, a large-scale open-ended lifelog QA dataset constructed from 18 months of multimodal lifelog data. Lifelogging is the passive collection and analysis of personal daily activities using wearable devices, producing rich multimodal data such as images, locations, and biometrics. Question answering (QA) over lifelog data enables users to interactively query their own experiences, supporting applications in memory support, lifestyle analysis, and personal assistance. OpenLifelogQA contains 14,187 Q&A pairs spanning multiple question types and difficulty levels, designed to support robust evaluation in realistic settings. Compared with prior resources, OpenLifelogQA offers greater diversity and practicality for real-world applications. To establish baselines, we evaluate the LLaVA-NeXT-Interleave 7B model, achieving 89.7% BERTScore, 25.87% ROUGE-L, and an average LLM Score of 3.97. By releasing OpenLifelogQA, we aim to promote future research on lifelog technologies, paving the way for personal lifelog assistants capable of memory augmentation, healthcare support, and lifestyle coaching.

14:10 *Thanh-Son Nguyen*, *Van-Loc Nguyen*, *Cong-Luan Le* and *Viet-Tham Huynh*

EnAug: ENT Endoscopy Images Classification Using Ensemble and Augmentation Methods

ABSTRACT. Ear, nose, and throat (ENT) endoscopy is a key diagnostic tool for detecting a variety of head and neck conditions. However, automated analysis of endoscopic images remains challenging due to inconsistent image quality, limited labeled data, and the inherent variability in human interpretation. In this work, we present a robust classification framework based on an ensemble of deep learning models, designed specifically for ENT endoscopy images. To address class imbalance and improve generalization, we employ a novel data augmentation strategy that combines symmetry-based label flipping with Mixup, Mosaic, and other augmentation techniques. Our approach is evaluated on a curated ENT dataset covering seven anatomical categories, achieving an accuracy of 95.82%, which surpasses several competitive baselines. In addition to strong overall performance, our method demonstrates improved robustness on underrepresented classes, showing its potential for real-world deployment in clinical settings. This work highlights the effectiveness of deep learning ensembles and thoughtful augmentation in building scalable AI tools for medical imaging. The full implementation and models are available at our repository at:
https://github.com/thanhson28/acmmm2025_entrep_challenge_was.

14:30 *Minh-Khoa Le-Phan*, *Minh-Hoang Le*, *Minh-Triet Tran* and *Trong-Le Do*

EDGER: EDge-Guided with HEatmap Refinement for Generalizable Image Forgery Localization

ABSTRACT. Text-guided inpainting has made image forgery increasingly realistic, challenging both SID and IFL. However, existing methods often struggle to point out suspicious signals across domains. To address this problem, we propose EDGER, a patch-based, dual-branch framework that localizes manipulated regions in arbitrary resolution images without sacrificing native resolution. The first branch, Edge-Guided

Segmentation, introduces a Frequency-based Edge Detector to emphasize high-frequency inconsistencies at manipulation boundaries, and fine-tunes a SegFormer to fuse RGB and edge features for pixel-level masks. Since edge evidence is most informative only when patches contain both authentic and manipulated pixels, we complement Edge-Guided Segmentation with a Synthetic Heatmapping branch, a classification-based localizer that fine-tunes a CLIP-ViT image encoder with LoRA to flag fully synthetic patches. Together, Synthetic Heatmapping provides coarse, patch-level synthetic priors, while Edge-Guided Segmentation sharpens boundaries within partially manipulated patches, yielding comprehensive localization. Evaluated in the MediaEval 2025, SynthIM challenge, Manipulated Region Localization Task's setting, our approach scales to multi-megapixel imagery, remains robust to resizing, compression, and cropping, and exhibits strong cross-domain generalization. Extensive ablations highlight the complementary roles of frequency-based edge cues and patch-level synthetic priors in driving accurate, resolution-agnostic localization.

14:50 *Minh-Loi Nguyen*, *Xuan-Vu Le*, *Long-Bao Nguyen*, *Hoang-Bach Ngo* and *Trung-Nghia Le*

Hierarchical Multi-Modal Retrieval for Knowledge-Grounded News Image Captioning

ABSTRACT. Traditional image captioning methods often struggle to generate comprehensive, context-rich descriptions, especially for details not directly observable from visual cues. To overcome this, we propose a novel retrieval-augmented image captioning framework that generates captions with deeper insights, such as object attributes, event context, and underlying significance, by leveraging external knowledge. Our approach features a hierarchical multi-modal article retrieval mechanism that moves beyond monolithic text entities. This retrieval considers article structure-aware features, including weighted textual components (e.g., headlines, body sections) and visual placement patterns, alongside multi-faceted similarity computations (content--visual, visual--visual, and discourse positioning). A subsequent contextual relevance refinement stage further enhances the retrieved information. The retrieved articles then serve as the knowledge base for caption generation: first, a VLM generates a concise image description; second, we segment relevant information from the retrieved articles based on this description; and finally, an LLM utilizes both the description and extracted knowledge to generate a comprehensive, contextually detailed caption. We participated in the ACM Multimedia EVENTA 2025 Challenge and achieved 5th place with an overall score of 0.2824 on the private test set of the OpenEvent-V1 dataset. Source code is publicly released at https://github.com/mf0212/EVENTA-Challange.

15:10 *Ngoc Nguyen Tran*, *Duc-Duy Nguyen* and *Nhat-Duc Le*

From Relative to Absolute: Monocular Depth Estimation in Aerial Imagery

ABSTRACT. Monocular depth estimation from single aerial images is fundamental to autonomous navigation, 3D reconstruction, and terrain analysis, yet practical deployment requires absolute (metric) depth instead of the purely relative outputs produced by most deep models. This work benchmarks three state-of-the-art methods - Marigold, NDDepth, and Unidepth - on three aerial datasets (ESPADA, ENRICH-Aerial, Skyscene) with a comprehensive metric suite, then investigates two routes to metric recovery: a least-squares scale-shift alignment and a learned scale-shift predictor based on ResNet50. On ESPADA, Marigold outperforms Unidepth with lower error and higher accuracy (absErrRel 0.1699 vs 0.1972, LinearRmse 5.1449 vs 6.0058, delta1Acc 0.8467 vs 0.8255). On ENRICH-Aerial, the two models are comparable (Marigold absErrRel 0.0673, delta1Acc 0.9733; Unidepth 0.0694, 0.9751). Prompt conditioning analysis on ESPADA shows that adding aerial-scene prompts improves over no-prompt (e.g., absErrRel drops from 0.1699 to 0.1402 with true

prompts and to 0.1365 – 0.1373 with partly-correct prompts), while ENRICH-Aerial exhibits near-neutral sensitivity to prompts. The classical least-squares alignment yields only marginal numerical gains across datasets. On ESPADA, the absErrRel improves trivially from 0.1699 to 0.169815, while on ENRICH-Aerial it changes from 0.0673 to 0.06726. These negligible differences indicate that once predictions are stable, the benefit of least-squares calibration is limited. Finally, our Res-Net50-based regressor directly predicts scale and shift to convert relative depth to absolute depth, achieving strong results that match the best calibrated outputs (ESPADA: absErrRel 0.1694, delta1Acc 0.8471; ENRICH-Aerial: 0.0674, 0.9734). Collectively, these findings demonstrate that absolute depth from monocular aerial imagery is attainable with high fidelity via lightweight calibration or learned scale–shift prediction, enabling reliable downstream drone and GIS applications.

13:30-17:20 **Session 10E: Poster Exhibition**

CHAIRS: *Van-Duy Nguyen*, *Dung Le Duy* and *Ho Bao*

Bach Nguyen Hoang, *Tung Pham Quang*, *Huy Nguyen Sinh*, *Thanh Nguyen Chi* and *Hai Vu*

Anatomy-based Brain Hemorrhage Segmentation and Application in Assessment of Traumatic Brain Injury Severity

ABSTRACT. Traumatic brain injury (TBI) requires an objective and timely severity assessment to support clinical decision making and integration into digital health workflows. In this study, we present a multimodal pipeline based on features extracted from CT images, integrated with structured clinical variables. The proposed method utilizes a dataset named 103_TBI, which comprises 504 records. To extract features from brain CT images, a U-Net-based neural network is used to quantify lesion volumes (e.g., epidural, subdural, and intraparenchymal hematomas), midline shift, and subarachnoid characteristics. In addition, the unified workflow performs clinically informed binning, KNN-based imputation, Z-score normalization, and class rebalancing using SMOTE. Tree-based ensemble models (Random Forest and XGBoost) trained on the 103_TBI dataset achieve accuracies of up to 94.20\%. The results highlight the added value of segmentation features, particularly midline shift and hematoma burden, when combined with key clinical indicators such as the Glasgow Coma Scale. The proposed framework demonstrates a practical approach to integrating CT imaging features with clinical data to assess the severity of TBI. The experimental results confirm the feasibility of deploying this solution to support the evaluation of the severity of TBI in clinical diagnostics.

Amira Yousif, *Kunal Agrawal*, *Ba-Thinh Tran-Le*, *Vatsa Patel* and *Tam Nguyen*

House Price Prediction via Attribute, Visual, and Economic Features

ABSTRACT. In this paper, we investigate the impact of economic factors such as interest rates inflation, employment, and local market conditions on the house pricing along with the attribute and visual features. To this end, we collect the House Attribute, Visual, and Economic Network Dataset (HAVEN-3000), which includes 3,000 houses from 22 states and 74 cities. Then, we propose a multimodal approach that combines house photos with property attributes and economic factors. Our proposed work achieves a Mean Absolute Error (MAE) of 10.9178 and an R2 of 0.3863. We further benchmark a direct prediction XGBoost model, which delivers improved results with MAE values of 9.245805 (2023), 10.978892 (2024), and 6.896334 (2025), alongside R2 scores of 0.6528, 0.6106, and 0.8219. Results high- light interest rate fluctuations as the strongest driver of housing price dynamics, highlighting the importance of incorporating multimodal data into predictive models for real estate.

Kasturi Jamale, *Kunal Agrawal*, *Ba-Thinh Tran-Le*, *Jayanth Merakanapalli*, *Soham Chousalkar*, *Vatsa Patel*, *Trung-Nghia Le* and *Tam V. Nguyen*

AEye: Avian Monitoring from Streaming Videos

ABSTRACT. The conservation of bird species, especially those that are endangered or at risk of extinction such as eagles and hawks, has become an urgent ecological priority. In this paper, we address the bird detection and classification problem, aimed at supporting large-scale avian monitoring and conservation efforts. Our work focuses on recognizing both adult birds and their chicks, with chicks observed from the time of hatching until fledging. We meticulously collect a realistic and diverse AEye dataset using YouTube streaming videos, covering 20 bird species, each classified into two categories, namely, parent and chick. The AEye dataset includes both daytime and night vision footage to evaluate model performance under different lighting conditions. We leverage the YOLO object detection model on the newly collected dataset, which demonstrates strong performance in detecting and classifying birds in different environments. We also evaluate performance in both daytime and nighttime settings.

Thi Quynh Mai Banh, *Trong Dung Bui*, *Van Chien Trinh* and *Thi Hanh Nguyen*

Optimizing UAV Swarm Routing with Optical Communication Systems

ABSTRACT. Fifth generation (5G) and beyond 5G (5G/B5G) networks deliver high-speed, widespread data transfer with reduced latency and enhanced connectivity compared to earlier systems. Among them, unmanned aerial vehicles (UAVs) and optical communication technologies have been increasingly applied in 5G/B5G networks, addressing the growing need for rapid data transfer and extensive connectivity in dynamic environments. This study explores routing algorithms for multi-hop UAV swarm networks, with a focus on minimizing latency during data transmission. We develop both path-delay optimal and heuristic algorithms to address routing challenges in large-scale UAV swarm networks, ensuring efficient data delivery across complex topologies. We also add must-pass intermediate UAVs that ensure critical information delivery, which are additional constraints to the routing design. Numerical results demonstrate that the proposed path-delay optimal algorithm consistently achieves lower latency compared to the heuristic approach for data transmission compared to the heuristic algorithm, though it requires higher computational complexity, highlighting a trade-off between performance and resource demands.

Khanh-Linh Dinh, *Long-Giang Nguyen*, *Thi-Bac Do*, *A. A. Kostina*, *A. A. Moldovyan* and *D. N. Moldovyan*

Practical multivariate algebraic signature scheme with one hidden group

ABSTRACT. In the field of post-quantum two-key cryptography, a significant interest is the development of practical algebraic schemes of electronic digital signature (EDS) with a secret (hidden) group, the security of which is based on the computational complexity of solving systems of power equations with many unknowns. Associative non-commutative finite algebras (ANFA) are used as an algebraic carrier in such crypto schemes. A critical concern in developing these types of EDS schemes lies in guaranteeing a sufficient level of randomness in the fitting element of the digital signature, which is a vector S, repeatedly included in the verification equation as a multiplier. A known solution to this problem is based on calculating S depending on two vectors randomly selected from two commutative secret groups such that the elements of one of them are non-commutative with the elements of the other. This mechanism requires the use of auxiliary fitting signature elements, which are calculated independently of the vector S, which creates potential prerequisites for calculating the secret key in parts. The paper proposes a new method for enhancing signature randomization, which is distinguished by calculating the value of S depending on two vectors randomly selected from one commutative secret group. Based on the proposed method, a practical post-quantum algebraic digital signature algorithm is developed, in which the calculation of auxiliary fitting elements of the

signature and the vector S is carried out in conjunction and simultaneously. Due to this, the specified potential vulnerability of algebraic digital signature algorithms with a hidden group is eliminated.

Duc Do Minh, *Anh Tran Le Duc*, *Dung Hoang Viet*, *Ngoc Tuyen Le* and *Tran Quang Duc*

QuantaMind: A Robust and Efficient Framework for Quantum Machine Learning Applications

ABSTRACT. Quantum Neural Networks (QNNs) possess the capability to significantly reduce the complexity of training neural networks. This research examines the potential advantages of QNNs compared to classical neural networks. We present QuantaMind, a novel framework for rapidly building hybrid classical-quantum models. We investigate hybrid classical-quantum neural networks combining quantum circuits with classical layers, specifically examining their performance on classification and regression tasks across many datasets by constructing Hybrid Quantum Feedforward Neural Networks (HQFNNs) and Hybrid Quantum Convolutional Neural Networks (HQCNNs). In most circumstances, these models show competitive performance with their classical counterparts in terms of accuracy and loss. QuantaMind demonstrates its ability to be suitable for many application domains, indicating an important step forward in Quantum Machine Learning.

Hai Nam Nguyen, *Truong Son Pham*, *Viet Anh Phan*, *Huu Noi Nguyen* and *Van Loi Cao*

GENLog: Enhance Generalization to Log-based Anomaly Detection

ABSTRACT. Log-based anomaly detection is crucial for maintaining the reliability and security of modern software systems. While deep learning models, particularly those based on Transformers like NeuralLog, have shown significant promise, they often suffer from limited generalization. This limitation arises because purely discriminative training, driven by cross-entropy loss, can lead to "shallow" representations over-specialized to the training data. To address this, we propose GENLog, a novel architecture designed to enhance generalization by integrating generative and discriminative learning. GENLog augments the standard Transformer encoder with a decoder module, creating a multi-task framework. In addition to the primary classification task, the model is simultaneously trained to reconstruct the original input sequence. This reconstruction objective acts as a powerful regularizer, compelling the encoder to learn a rich, comprehensive latent representation that preserves essential semantic and structural information. We conduct extensive experiments on two benchmark datasets, HDFS and BGL, demonstrating that GENLog significantly outperforms state-of-the-art methods, especially in challenging low-data regimes. Our analysis further includes an ablation study on the trade-off hyperparameter, providing insights into the synergistic relationship between the two learning objectives.

Guillaume Guérard and *Soufian Ben Amor*

Architecting Trustworthy AI: The Cyber-Resilient AI (CRAI) Framework

ABSTRACT. The inductive biases and optimization objectives integral to modern artificial intelligence, particularly in large-scale generative models, create inherent and exploitable security vulnerabilities. The differentiability and high-dimensional nature of deep neural networks, for instance, give rise to adversarial attack surfaces that are not mere implementation flaws but fundamental properties of the models themselves. Current research often addresses these vulnerabilities in isolation, resulting in a fragmented landscape of point defenses that lack a unifying structure. This paper addresses this methodological gap by conducting a structured critical analysis of the AI-driven threat landscape. A novel, unified framework for proactive defense—the Cyber-Resilient AI (CRAI) architecture—is introduced. The CRAI framework is built on three synergistic pillars: (1) Hardening collaborative learning through cryptographic integrity proofs, (2) Enhancing model auditability via causal and

counterfactual explainability (XAI), and (3) Implementing adaptive governance informed by real-time model state analysis. This work's primary contribution is a formal synthesis that connects specific AI model properties to emergent threat vectors and maps them to a coherent, multi-layered defense strategy. It provides a new research roadmap for developing verifiably robust and secure AI systems, moving beyond reactive patching toward a paradigm of security-by-design.

Thang Phung Duc and *Dai Tho Nguyen*

Adaptive Federated Learning for Software Vulnerability Detection

ABSTRACT. Federated learning has recently been adopted in cybersecurity to enable collaborative model training on distributed threat data without disclosing sensitive information. However, existing approaches for software vulnerability detection rely on static aggregation and require shared test data to monitor convergence. To address these challenges, we introduce Adaptive Federated Learning (Adaptive FL), which dynamically allocates computation to clients with harder-to-learn vulnerability profiles and uses client-side validation to orchestrate the learning process without any centralized test data. We evaluate Adaptive FL on multiple vulnerability datasets and demonstrate that it achieves an F1-score of approximately 70%, outperforming VDBFL baselines by 5–10 percentage points, while maintaining stable runtimes of 100–150 s per round—over 20 times faster than competing methods. These results establish Adaptive FL as a practical, efficient, and privacy-preserving solution for federated vulnerability detection in heterogeneous software development environments.

Ha Nguyen, *Phuc Le*, *Dang Do*, *Cuong Nguyen* and *Chung Mai*

A Method for Building QA Corpora for Low-Resource Languages

ABSTRACT. Building high-quality question–answering (QA) datasets for low-resource languages is challenging due to the lack of annotated corpora. We propose a fully automated and scalable pipeline for constructing large-scale QA corpora from authoritative online sources. The pipeline includes five stages: (i) selecting authoritative QA websites; (ii) automated crawling; (iii) extracting question–context–answer (QCA) triples via site-specific templates that leverage semantic HTML; (iv) applying an AI-assisted fact-checking filter that uses Sentence-BERT retrieval with a high-similarity threshold followed by LLM verification, both against curated references; and (v) final canonicalization and deduplication to remove redundant items and maintain corpus diversity. Unlike conventional QA pairs, the QCA structure preserves contextual grounding, enhancing corpus utility and model robustness. Applied to Vietnamese, our method produced 30,000 QCA triples from four reputable sources. To demonstrate usability, we fine-tuned vit5-base, a Vietnamese sequence-to-sequence model, achieving strong results on a 1,000-triple test set (BLEU 89.1; semantic similarity >=0.8: 91.5%) and in a human evaluation of 500 samples (grammaticality 4.58/5, usefulness 4.29/5). Compared with existing baselines (question-generation-vietnamese-v2, Ollama, GPT-5), our model yields substantially higher performance, underscoring both the effectiveness of the pipeline and the quality of the corpus. The released dataset and tools are publicly available as open-source resources, providing a valuable benchmark for future research on Vietnamese QA and question generation in low-resource settings.

Marcus Jeremy Cariño, *Iverson David*, *Brylle Renzy Diminsil*, *Kate Louise Naguit* and *Adriane Brent Castro*

JuanQueue: A Digital Appointment and Queuing System for a Government Organization

ABSTRACT. This paper presents JuanQueue, a web-based digital appointment and queuing system designed to modernize document-request services in local government offices. The study addressed inefficiencies in manual workflows that cause long waiting times and administrative bottlenecks. It followed a quantitative

research design and Rapid Application Development (RAD) for system creation and iterative refinement. Evaluation was performed across four objectives: (1) correlation between manual waiting time and citizen satisfaction, (2) expert validation of the developed system under ISO/IEC 25010 software-quality standards, (3) user-experience and adoption analysis using the Unified Theory of Acceptance and Use of Technology (UTAUT), and (4) assessment of data-driven decision support via the Technology-to-Performance Chain (TPC) framework. Results showed very weak negative correlations between waiting time and perceived efficiency, usability, and satisfaction, confirming that citizens tolerate manual processes for familiarity rather than effectiveness. IT experts rated all ISO/IEC 25010 criteria "Very Acceptable," while users reported mean ratings of ≈5.6 / 6 across UTAUT constructs. Barangay staff rated information quality and decision support highly. Findings indicate that JuanQueue is technically ready, widely accepted, and capable of supporting data-informed local governance. The study demonstrates the feasibility of digitizing barangay services and highlights opportunities for scalable Software-as-a-Service deployment.

Nhan Vu

Smart Mobility through Hybrid Offline-Online Scheduling for Ridesharing

ABSTRACT. In recent years, ridesharing has emerged as one of the most cost-effective and ef-ficient transportation solutions, allowing more than one people to share a single vehicle. Nonetheless, scheduling remains a critical challenge that must be ad-dressed to enhance user adoption. This paper addresses this issue through a two-fold objective: firstly, by providing frequent riders with reliable, pre-arranged routes; and secondly, by enabling the dynamic insertion of new riders into active shared trips. To this end, we develop an algorithm named aVC, which clusters regular users into ridesharing groups based on the similarity of their frequent travel routes. This approach eliminates the need for users to repeatedly search for rides or experience extended waiting times, as trip details and drivers assignments are communicated to users in advance. Additionally, when a regular ridesharing trip commences and vacant seats remain, drivers can accommodate real-time re-quests from new users without disrupting the planned itinerary. To efficiently handle such scenarios, we introduce a method named biSearchIns designed to quickly process shared trip queries. The proposed method are assessed against existing approaches using simulated datasets. Experimental results demonstrate that our methods surpass current solutions in terms of computational efficiency, the number of riders served, and the overall reduction in the required number of vehicles

15:30-16:00 Tea Break

16:00-17:20 Session 11A: SOICT Technical Session XXI: Lifelog Event Retrieval

CHAIR: *Minh-Triet Tran*

LOCATION: Grand Ballroom A, 2F

16:00 *Minh Nguyen*, *Nga N.T. Nguyen*, *Cuong Dinh*, *Dang Nguyen*, *Dat Tien Nguyen* and *Huy M. Le*

CLIPAR: Multimodal and Temporal-Aware Video Retrieval System

ABSTRACT. The Ho Chi Minh AI Challenge 2025 sets an ambitious goal of building a powerful video retrieval system that can compete with other teams. To address this challenge, CLIPAR is designed and implemented, integrating multiple search strategies, including Semantic Search, OCR Search, ASR Search, Object Detection, and Image Matching. CLIPAR processes audio and visual streams separately, extracting keyframes, transcripts, and embeddings to support diverse retrieval tasks. One of the system's key strengths is its flexibility: users can search using text, objects, or images, and the system quickly returns the most relevant video segments. CLIPAR's system demonstrates strong performance in the preliminary stage of the competition, achieving top ranking among participating teams. Its ability to handle multiple types

of queries allows it to return relevant video segments quickly and accurately. These results highlight CLIPAR's practical potential for real-world video retrieval tasks and show that a system designed with flexible, multimodal search capabilities can outperform more specialized approaches.

16:20 *Duc-Tho Nguyen*, *Hieu-Hoc Tran-Minh*, *Khanh-Hoa Lam*, *Hoang-Nhut Ly*, *Huu-Phuc Huynh*, *Thanh-Tien Tran* and *Trung-Nghia Le*

Vortex: A Multi-Modal Fusion System for Intelligent Video Retrieval

ABSTRACT. This paper presents Vortex, a multimodal video retrieval system developed for the Ho Chi Minh City AI Challenge 2025, designed to advance intelligent multimedia search and temporal reasoning. The system integrates adaptive keyframe extraction, multimodal metadata generation from vision-language and speech models, and a hybrid retrieval strategy that fuses CLIP and SigLIP2 embeddings through Reciprocal Rank Fusion to balance global and fine-grained semantics. To enhance interactivity, Vortex incorporates Rocchio-based relevance feedback and a multi-stage temporal search mechanism for sequential event alignment. Built on Milvus and Elasticsearch, the architecture enables scalable indexing and efficient retrieval. Evaluated in the official competition, Vortex achieved a final score of 79.6/88 (90.5\%), demonstrating the complementary strengths of CLIP and SigLIP2 and confirming the effectiveness of the hybrid retrieval approach. The system establishes a robust foundation for future research in intelligent, context-aware, and interactive video retrieval.

16:40 *Danh Nguyen Duc*, *Duy Dang Phu*, *Lac Tran Nguyen Khai*, *Bao Luong Huynh Gia*, *Minh Quang Le*, *Van Thai Hung*, *Duy-Dinh Le* and *Thanh Duc Ngo*

Efficient Video Retrieval for Less-Resourced Languages via Multi-Modal Semantic Search

ABSTRACT. The surge of multimedia data has increased the demand for intelligent video retrieval systems capable of understanding complex natural-language queries. Current multimodal approaches often underperform in less-resourced languages, where semantic ambiguity, linguistic diversity, and cultural context hinder retrieval accuracy. To address this, we propose a multilingual video retrieval framework that integrates multimodal embeddings and a caption-based module to enhance performance. The caption-based module provides fine-grained, language-adaptive captions for each keyframe, segmented into overlapping chunks for embedding and re-ranking. This design enables precise semantic alignment between video content and user queries, improving both contextual relevance and retrieval precision. We evaluate the framework on Vietnamese, an under-resourced but linguistically rich language, demonstrating its adaptability and effectiveness. Experiments on large-scale Vietnamese video datasets, including the 2025 Ho Chi Minh AI Challenge, show that our approach significantly improves cross-modal understanding and retrieval performance, highlighting its potential as a robust foundation for multilingual multimedia search systems.

16:00-17:20 Session 11B: SOICT Technical Session XXII: AI Applications

CHAIR: *Hoang Ta*

LOCATION: Grand Ballroom B, 2F

16:00 *Huyen Nguyen*, *Hieu Dam*, *Thuy Nguyen Thi Thu* and *Viet Nguyen Kim*

AuMoM: A Framework for Learning Discriminative Speaker Embeddings using a Mamba-based Mixture of Experts and Contrastive Loss

ABSTRACT. This paper presents Audio MoE-Mamba (AuMoM), an innovative speaker verification system optimized for scalability and efficiency. AuMoM begins by transforming audio waveforms into spectrogram patches, which are then embedded as vectors and processed by a Mamba Encoder enhanced with a Mixture of Experts (MoE) architecture. Unlike conventional attention-based models, the bidirectional

Mamba Encoder leverages state-space modeling and convolutional operations to achieve faster processing speeds. The MoE layer dynamically routes inputs to specialized expert sub-models, increasing model capacity and efficiency. A classification token placed within the sequence facilitates learning bidirectional context. Finally, a Siamese Network compares the learned audio features using a contrastive loss function to distinguish between different speakers. This architecture enables AuMoM to produce discriminative speaker embeddings effectively, combining attention-free processing with MoE for enhanced computational efficiency.

16:20 *Anh Nguyen-Thi-Mai*, *Anh Nguyen-Thi-Van*, *Bao Doan-Quoc*, *Minh Tran-Duc*, *Tran Hung*, *Miroslav Voznak*, *Tu Dac Ho*, *Van Vo Nhan*, *Symeon Chatzinotas* and *Tran Dinh-Hieu*

A Survey on Challenges and Emerging Frontiers of Multi-Agent Systems

ABSTRACT. Multi-Agent Systems (MAS) have emerged as a fundamental approach for solving dynamic and distributed problems across domains such as robotics, communication networks, and intelligent decision-making. MAS agents are characterized by autonomy, sociality, and flexibility [1]. Recent advances such as deep reinforcement learning (DRL), large lan- guage model (LLM)-based agents, and context awareness have broadened MAS capabilities, but existing surveys are often limited to specific subdo- mains or focus on outdated platforms, ignoring the convergence between learning-based and language-based systems. This review provides a com- prehensive, technically rigorous view that connects classical MAS theory with emerging paradigms. We analyze cross-cutting system challenges, including scalability, cybersecurity, and privacy, and synthesize recent model families across five functional research areas: (1) MAS for Com- plex Problem Solving and Planning; (2) Embodied MAS for Physical Environments; (3) MAS for Emergent Communication and Decentral- ized Coordination; (4) MAS for Human-AI Teaming and Social Intelli- gence; (5) Toward Generalist and Multi-tasking MAS. This work aims to consolidate the fragmented literature, highlight common challenges, and outline future research opportunities for developing scalable, general, and interoperable MAS systems.

16:40 *Nhân Nguyễn Tiến*, *Sơn Lê Mai Thanh*, *Anh Vũ Đức*, *Bảo Bùi Quốc* and *Nidal Kamel*

Improving Plant Species Distribution Models with Hydrologic and Topographic Features

ABSTRACT. Species distribution models (SDMs) for plants typically prioritize climate and broad remote sensing while under-using hydrologic and topographic information. Using European presence–absence surveys from GeoPlant (~94k plots) and an XGBoost baseline built on Location + Climate + Land-cover (LCL), we quantify the added value of hydrologic and terrain context and test sensitivity to elevation source. We derive river and lake descriptors from HydroRIVERS/HydroLAKES (e.g., network position, connectivity, discharge, and lentic morphology), compute the Topographic Position Index (TPI) at multiple radii (150–3000 m), and compare predictors derived from ASTER GDEM versus Copernicus GLO-30 (COP). Models are trained and evaluated over five seeds using standard discrimination and retrieval metrics. Three consistent findings emerge. First, DEM choice matters: across like-for-like configurations, COP provides more informative elevation for SDM predictors than ASTER. Second, fine-scale topography helps when derived from high-quality elevation: among TPI scales, the 150 m radius yields the clearest improvement when paired with COP. Third, hydrologic context is complementary: river features improve the LCL baseline, and combining rivers with lake descriptors yields further gains. The best configuration—COP elevation + TPI (150 m) + river + lake features—achieves the highest overall accuracy, raising the F1 score to 29.97. Feature importances

corroborate these trends: latitude dominates among location variables, several BIO-climatic predictors remain near the top, river-network distance appears among the top five features, and TPI adds a measurable (though smaller) contribution alongside land cover and elevation. Together, these results provide a practical recipe for integrating hydro-topography into scalable plant SDMs and highlight the benefits of high-quality elevation data and fine-scale terrain context for continent-scale prediction.

16:00-17:20 **Session 11C: SOICT Technical Session XXIII: Recent Advances in Cyber Security**
CHAIR: *Tuan Dam*
LOCATION: Yersin Ballroom A, 2F

16:00 *Nghi Hoang Khoa*, *Vo Duc Chinh*, *Tu Chi Kien*, *Thai Hung Van* and *Phan The Duy*

Password Generation Based on GenAI for Evaluating the Security of Password-Based Control Systems

ABSTRACT. The rapid growth of Generative AI (GenAI) brings new challenges for password security. Traditional rules based only on length or character complexity are insufficient to measure real strength. Currently, many AI models for password guessing show severe limits. They often repeat guesses, generate in random order, and fail to follow the real attack patterns. This study presents a new framework for pattern-aware, search-based, ordered password generation. It combines a pattern-conditioned Generative Pretrained Transformer (PagPassGPT) with a tree-search algorithm (SOPG). This design creates password guesses in order, without repetition, and with correct patterns. The model was trained with the RockYou dataset and compared with advanced password-guessing models. Results show substantial improvement: our model reached 20.8% success in 1,000 guesses, while PassGPT reached only 2.2%. It also avoided repetition (0%), unlike others that repeated up to 99.9%. In conclusion, the framework is reliable for testing password policies and simulating AI-based attacks.

16:20 *Giang Tran Dinh Kien*, *Duy Anh Hoang*, *Van Tong*, *Tung Bui*, *Tran Quang Duc* and *Tuyen Le Ngoc*

FusionMalNet: A Hybrid Ensemble Architecture for Windows Malware Detection

ABSTRACT. Malware detection remains a critical yet challenging task, as traditional static analysis techniques struggle with increasingly sophisticated obfuscation methods. In this paper, we propose FusionMalNet, an ensemble deep-learning framework leveraging multimodal representations for robust Windows malware classification. FusionMalNet combines visual patterns extracted from Gramian Angular Field images via a RegNetY convolutional architecture and structural information captured from static file features using XGBoost. These complementary representations are integrated through a compact neural fusion module, enabling accurate and confident predictions. To facilitate evaluation, we also introduce a dataset of Portable Executable (PE) files from which aligned visual and tabular representations are derived. Extensive experiments demonstrate that FusionMalNet achieves state-of-the-art performance among the compared methods, attaining 99.67% accuracy and 99.97% ROC-AUC, surpassing multiple existing approaches across diverse evaluation metrics.

16:40 *Nghi Hoang Khoa*, *Dang Quang Huy*, *Cao Minh Duc*, *Ngo Duc Hoang Son* and *Phan The Duy*

PowerGAN: Enhancing PowerShell Attack Detection through GAN-Driven Data Generation

ABSTRACT. As defensive solutions advance, Living Off the Land (LoTL) attacks have emerged as a powerful evasion technique by exploiting native system tools. Among them, PowerShell—deeply integrated into Windows—has become a prime vector for stealthy LoTL attacks that frequently bypass traditional detection methods. While

machine learning (ML) and deep learning (DL) approaches have been widely applied, limitations in data quantity, and class balance for PowerShell scripts hinder their effectiveness. To address this challenge, we propose PowerGAN, a generative deep learning approach for producing additional training data to enhance ML- and DL-based PowerShell detection. Experimental results demonstrate that PowerGAN significantly improves detection performance, and we further compare different GAN variants to identify the most suitable model for this problem.

16:00-17:20 **Session 11D: SOICT Technical Session XXIV: Multimedia Processing**

CHAIR: *Duc-Hau Le*

LOCATION: Yersin Ballroom B, 2F

16:00 *Thanh-Nhan Vo*, *Trong-Thuan Nguyen*, *Tam V. Nguyen* and *Minh-Triet Tran*

SimGraph: A Unified Framework for Scene Graph-Based Image Generation and Editing

ABSTRACT. Recent advancements in Generative Artificial Intelligence (GenAI) have significantly enhanced the capabilities of both image generation and editing. However, current approaches often treat these tasks separately, leading to inefficiencies and challenges in maintaining spatial consistency and semantic coherence between generated content and edits. Moreover, a major obstacle is the lack of structured control over object relationships and spatial arrangements. Scene graph-based methods, which represent objects and their interrelationships in a structured format, offer a solution by providing greater control over composition and interactions in both image generation and editing. To address this, we introduce SimGraph, which is a unified framework that simultaneously integrates scene graph-based image generation and editing, enabling precise control over object interactions, layouts, and spatial coherence. In particular, our framework integrates token-based generation and diffusion-based editing within a single scene graph-driven model, ensuring high-quality and consistent results. Through extensive experiments, we empirically demonstrate that our approach outperforms existing state-of-the-art methods.

16:20 *Duc-Manh Phan*, *Quoc-Duy Tran*, *Duy-Khang Do*, *Anh-Tuan Vo*, *Hai-Dang Nguyen*, *Trong Le Do*, *Mai-Khiem Tran*, *Vinh-Tiep Nguyen*, *Tam V. Nguyen*, *Isao Echizen* and *Trung-Nghia Le*

Forged Calamity: Benchmark for Cross-Domain Synthetic Disaster Detection in the Age of Diffusion

ABSTRACT. The rapid advancement of text-to-image diffusion models has enabled the creation of highly photorealistic synthetic images that closely resemble real photographs, making it increasingly difficult to distinguish authentic content from AI-generated fabrications. This poses challenges for cybersecurity, digital forensics, and disaster response, where fake imagery of floods, fires, or earthquakes can spread misinformation or disrupt emergency operations. To address this, we introduce Forged Calamity, a benchmark dataset for synthetic disaster detection containing 30,000 images, including 6,000 real and 24,000 synthetic samples generated by four diffusion models. Comprehensive experiments across fine-tuned and zero-shot settings reveal consistent weaknesses in current forensic approaches. Fine-tuned detectors perform well in-distribution but lose up to 50% accuracy on unseen generators or disaster types, showing overfitting to model-specific artifacts. Zero-shot generalized detectors also struggle to maintain stable accuracy, with only limited resilience in a few representation-robust models. These findings highlight persistent generalization gaps and the urgent need for domain- and model-agnostic detection methods to ensure visual authenticity in the diffusion era.

16:40 *Thi-Minh-Thu Vu*, *Quoc-An Nguyen*, *Bich-Dat Nguyen*, *Dinh-Quang-Minh Tran* and *Hoang-Quynh Le*

HERF: Hybrid Evidence Retrieval Framework for Entity-Centric Question Answering

ABSTRACT. Question answering (QA) systems are typically built on either knowledge bases or the open web, each with distinct advantages and limitations. Structured knowledge bases provide high-precision answers but are often incomplete, while open web data offer broader coverage at the cost of factual reliability. To address this trade-off, this paper presents HERF (Hybrid Evidence Retrieval Framework), a hybrid architecture designed for entity-centric question answering—a task focused on queries about specific entities. HERF's parallel architecture retrieves complementary evidence: one stream queries a knowledge base, while the other extracts contextual evidence from open web text. An aggregator then fuses evidences from both streams into a unified set of candidate answers, ensuring both precision and coverage. On the EntityQuestions benchmark, HERF achieves superior performance compared to existing comparative models. The results demonstrate that a parallel fusion strategy is a highly effective approach, highlighting the potential of integrating hybrid evidence to build more robust and accurate QA systems.

18:30-21:30 Gala Dinner

LOCATION: Lakshmi Hall 1 - Champa Island Nha Trang - Resort Hotel & Spa

SOICT HISTORY

2010

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Ta Quang Buu Library, HUST, Hanoi, Vietnam**

2011

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Ta Quang Buu Library, HUST, Hanoi, Vietnam**

2012

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Halong Plaza Hotel, Ha Long Bay, Quang Ninh City, Vietnam**

2013

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Pullman Danang Beach Resort, Da Nang City, Vietnam**

2014

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Ta Quang Buu Library, HUST, Hanoi, Vietnam**
- SoICT'14 was in conjunction with iiWAS2014

2015

- Organizer: **Hanoi University of Science and Technology, Hue University**
- Venue: **Imperial Hue Hotel, Hue City, Vietnam**

2016

- Organizer: **Hanoi University of Science and Technology, Nguyen Tat Thanh University**
- Venue: **Rex Saigon Hotel, Ho Chi Minh City, Vietnam**

SOICT HISTORY

2017

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Sheraton Nha Trang Hotel, Nha Trang City, Vietnam**

2018

- Organizer: **Hanoi University of Science and Technology, Danang University**
- Venue: **Pullman Danang Beach Resort, Da Nang City, Vietnam**

2019

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Wyndham Hotel, Ha Long Bay, Quang Ninh Province, Vietnam**

2022

- Organizer: **Hanoi University of Science and Technology**
- Venue: **Wyndham Hotel, Ha Long Bay, Quang Ninh Province, Vietnam**

2023

- Organizer: **Hanoi University of Science and Technology, VNU-HCM University of Science, Laboratory Informatics, Modelling and Optimisation System (LIMOS), The French National Centre for Scientific Research (CNRS), Vietnam Institute for Advanced Study in Mathematics**
- Venue: **Rex Saigon Hotel, Ho Chi Minh City, Vietnam**

2024

- Organizer: **Hanoi University of Science and Technology, VNU-HCM University of Science, The University of Danang - University of Science and Technology**
- Venue: **Furama Resort Danang, Da Nang City, Vietnam**

2025

- Organizer: **Hanoi University of Science and Technology - VNU-HCM University of Science**
- Venue: **Sheraton Nha Trang Hotel, Nha Trang City, Vietnam**

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATIONS TECHNOLOGY

Hanoi University of Science and Technology (HUST) is a leading public university in Vietnam. Established in 1956, HUST is committed to research, knowledge development, and intellectual training. Renowned for its influential alumni, HUST remains dedicated to quality education while embracing innovation to meet the evolving needs of students and society, particularly in the digital age.

The **School of Information and Communications Technology (SoICT)** is a subsidiary of HUST, founded in March 1995. Towards the 30th anniversary of the establishment of the School, SoICT has achieved significant milestones, solidifying its position as a leader in information and communications technology education, research, and technology transfer in Vietnam. This is evident in its strong performance in global university rankings as well as social recognition, maintaining its position as the top public university in Vietnam.

SoICT adheres to the educational philosophy ***"Towards excellence in digital age".*** The school aims to cultivate learners with a strong academic foundation, a keen ability to acquire new knowledge and advanced technologies, an innovative mindset, and a passion for applying scientific and technological advancements to improve lives. SoICT provides a comprehensive learning experience that includes rigorous academic lectures, practical exercises, experiments, and impactful internships.

(+84) 24 3869 2463
vp@soict.hust.edu.vn

Room 505, B1 Building
Hanoi University of
Science and Technology

Dai Co Viet Str. Hanoi, Vietnam

soict.hust.edu.vn/en

HCMUS
Viet Nam National University
Ho Chi Minh City
University of Science

UNIVERSITY OF SCIENCE VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY

The University of Science, Viet Nam National University Ho Chi Minh City has a history of over 80 years of establishment and development. The University's mission is to become the leading university in Viet Nam and Southeast Asia in education and research on science, knowledge-based technology and the digital economy.

The University's core values are science, creativity, integration, empathy, nurture, community, and empowerment.

The University has 10 faculties: Faculty of Mathematics and Computer Science, Faculty of Information Technology, Faculty of Physics and Engineering Physics, Faculty of Chemistry, Faculty of Biology and Biotechnology, Faculty of Environment, Faculty of Geology, Faculty of Materials Science and Technology, Faculty of Electronics and Telecommunications, Faculty of Interdisciplinary Science.

The University also has 02 research institutes, 11 specialized research laboratories, with 04 VNUHCM key laboratories, and 16 research and education centres of technologies, natural sciences, languages and other fields.
The University currently has 22 undergraduate courses, 34 master's degree courses and 30 doctorate degree courses in 07 fields: natural sciences, life sciences, mathematics and statistics, computers and information technology, environment, technology and engineering.

The University has cooperated with more than 60 scientific institutions and over 50 prestigious universities and research institutes in 30 countries and territories, in both undergraduate and postgraduate programs. Annually, the University and Faculties organize exchange activities, sign cooperation memorandums, and exchange lecturers and students; significantly contributing to the research expertise and qualified improvement of lecturers, staff and students.

The University has two main campuses:

Campus 1: 227 Nguyen Van Cu, Ward 4, District 5, HCMC | Campus 2: Urban area of VNU-HCM, Di An - Thu Duc, HCMC

FINANCIAL SPONSORS

VinIF
Powered by VinBigdata

TECHNICAL SPONSORS

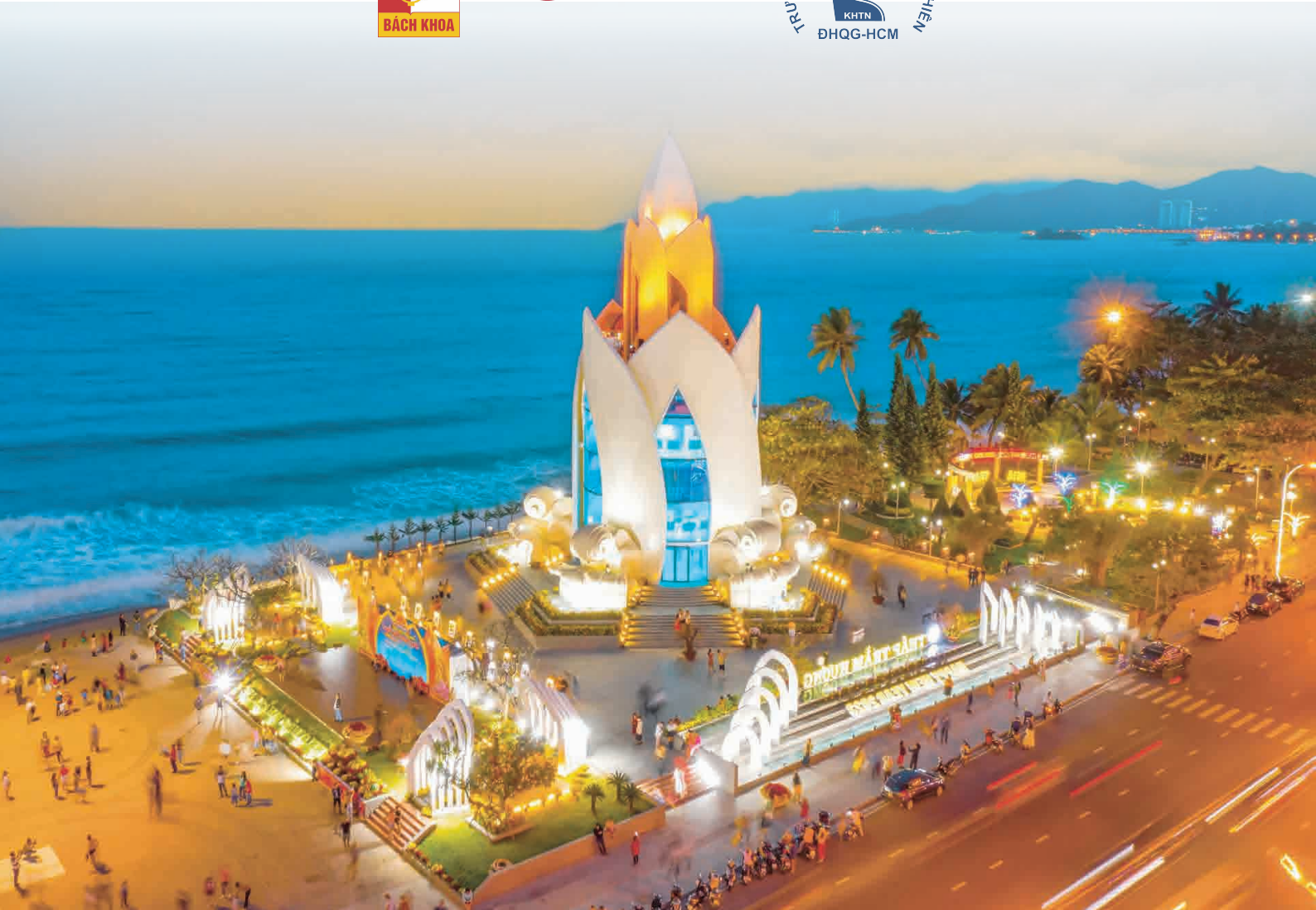
Springer

CCIS

ORGANIZERS

ĐẠI HỌC BÁCH KHOA SOICT

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN KHTN ĐHQG-HCM



SOICT 2025