Neural Networks and Model Compression

Date: August 3, 2026 (Mon)

Time: From 9:00 AM (Full-day Event, HKT)

Venue: Rooms A & B, 8/F, Bio-Informatics Centre,
2 Science Park West Avenue,
Hong Kong Science Park, Pak Shek Kok, New Territories, Hong Kong

Organizers: Fenglei Fan, Shuren Qi, and Jingxiao Liao

Hosts: Huawei Leibniz Research Center,
Frontier of Artificial Network Group, CityU,
Specialized Committee on Mathematical AI, HKSIAM

August 3, 2026 (Monday, Morning)	**Speaker/Activity**
9:00-9:10	*Opening*
9:10-9:50	Junchi Yan
9:50-10:30	Xiangyu Zhao
10:30-10:50	*Tea Break*
10:50-11:30	Kai Zhao
11:30-12:10	Gen Li
August 3, 2026 (Monday, Afternoon)	**Speaker/Activity**
14:30-14:40	*Opening*
14:40-15:20	Yongduan Song
15:20-16:00	Fenglei Fan
16:00-16:20	*Tea Break*
16:20-16:40	Ran Guan
16:40-17:20	Bohan Zhuang
17:20-18:00	Liang Chen

Junchi Yan
Shanghai Jiao Tong University
yanjunchi@sjtu.edu.cn

Machine Learning For Constrained Problems: From Combinatorics To Differential Equations

Various constraints exist across discrete graph theory and combinatorics, continuous differential equations, quantum domains, and beyond. Integrating machine learning with generalized constraint solving, as well as exploring cutting-edge applications thereof, constitutes a valuable endeavor for artificial intelligence to evolve from probabilistic intelligence toward constraint-centric intelligence. This report primarily presents the speaker's representative research achievements in the aforementioned fields over recent years, spanning fundamental theories, methodologies, and practical applications. It will also share insights and experience gained throughout the corresponding research journey.

Bio: Junchi Yan is a Professor with School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China. Before that, he was a Senior Research Staff Member with IBM Research and later an affiliated consultant Researcher with AWS AI Lab. His research interests include machine learning and AI4Science and was once featured on IEEE Explorer. He serves the Associate Editor for IEEE TPAMI/TNNLS/TEVC/CIM, JMLR, TMLR, Pattern Recognition, and (Senior) Area Chair for CVPR, ECCV, ICML, NeurIPS, ICLR, AAAI etc. He is a Fellow of IAPR and IET. He received the IEEE CS AI'10 to Watch 2024, IEEE CIS Early Career Award 2026, and CVPR 2024 and IROS 2025 Best Paper Candidate, ACL 2025 Outstanding paper. He serves on the Board of ICML, and as the program co-chair for ACM Multimedia 2026.

Xiangyu Zhao
City University of Hong Kong
xy.zhao@cityu.edu.hk

Model Compression For Large Models And On-Device Intelligence

This talk introduces recent advances in model compression for large-scale machine learning systems, with a focus on multimodal large models and ultra-compact language models for real-world deployment. The presentation first

discusses the motivation for compressing large multimodal models in industrial scenarios, where high latency and energy consumption limit practical adoption. To address these challenges, an iterative framework combining multi-stage pruning and pair-wise distillation is presented, enabling substantial reduction in model size and inference cost while preserving task performance. The compressed model has been successfully deployed in an advertisement auditing system, achieving significant latency reduction and energy savings with only marginal performance loss. The talk further presents Cross-Distillation for ultra-compact model compression, a framework designed for highly resource-constrained mobile environments. By integrating student pretraining, embedding pruning, bidirectional teacher-student distillation, and module-wise quantization, the method produces a 1.91MB BERT-based model that maintains strong performance on both public benchmarks and proprietary industrial datasets. This model has been deployed in Alipay's edge recommendation system at scale, demonstrating the effectiveness of model compression in enabling practical, privacy-preserving, and low-latency on-device intelligence.

Bio: Xiangyu Zhao is a tenured Associate Professor of Data Science at City University of Hong Kong (CityU) and a CityU Thetos Young Scholar. He joined CityU as an Assistant Professor in 2021 and received early promotion in July 2025. He earned his Ph.D. at Michigan State University (advised by Prof. Jiliang Tang), his M.S. at USTC, and his B.Eng. at UESTC. His research spans data mining and machine learning, with a focus on recommender systems, information retrieval, large language models and agents, trustworthy AI, AutoML, and AI applications in healthcare, smart cities, science, and education. His honors include the KDD'25 Best Paper Award Runner-Up, ICDM'21/'22 Best-ranked Papers, Global Chinese AI Rising Stars, and multiple industry research funds. Notably, he was the only (then) Assistant Professor in Hong Kong to lead both an HKRGC CRF and RIF key project. His students have won the NeurIPS Competition 2024 Championship and KDD Cup 2024 Runner-Up, among other awards. He is listed among Stanford's World's Top 2% Scientists (2024–2025). He has served as conference chair/organizer for venues including WSDM'27 and WWW'25, and as associate editor for journals such as Neural Networks and Nature NPJ Artificial Intelligence. Twenty of his algorithms have been deployed at leading companies including Alibaba, Tencent, Google, and Amazon, powering systems that serve over 100 million daily active users and contribute over 10 billion RMB in annual revenue.

Kai Zhao
East China Normal University
kaiz@cs.aau.dk

Enhancing Diversity For Data-Free Quantization

Model quantization is an effective way to compress deep neural networks and accelerate the inference on edge devices. Existing quantization methods usually require original data for calibration during the compressing process, which may be inaccessible due to privacy issues. A common way is to generate calibration data. However, the generators in these methods have the mode collapse problem, making them unable to synthesize diverse data. To solve this problem, we leverage the information from the full-precision model and enhance both inter-class and intra-class diversity for generating better calibration data. Besides, regulation losses are proposed to make the generator produce diverse data with more patterns from the perspective of activated feature values and for the quantized model to learn better clip ranges adaptive to diverse calibration data. Extensive experiments show that our method achieves state-of-the-art quantization results for both Transformer and CNN architectures.

Bio: Kai Zhao, professor at East China Normal University and researcher at MBZUAI. He completed PhD at Aalborg University, Denmark, supervised by Prof. Christian S. Jensen, and BSc at Peking University, China. His research interests lie in time series, graph learning, model compression, manifold learning, knowledge graph and reasoning. Related works have been published, including Nature Communications, VLDB, CVPR, KDD, WWW, ICML, ICLR, TKDE etc. He also serves as PC, AC and SPC, including ICCV, NeurIPS, ACM MM and AAAI etc.

Gen Li
Chinese University of Hong Kong
genli@cuhk.edu.hk

Provable Sampling Acceleration In Diffusion Language Models

Diffusion language models (DLMs) have emerged as a compelling alternative to autoregressive (AR) models by enabling parallel, non-sequential token generation. Despite their strong empirical performance, the theoretical understanding of how decoding strategies affect sampling efficiency remains

limited. In this talk, we present recent theoretical advances showing that DLM sampling can be substantially accelerated through decoding strategies. We establish convergence guarantees for both uniform and confidence-based decoding strategies. In particular, the iteration complexity depends on information-theoretic quantities that capture the intrinsic complexity of the target distribution. These results provide a theoretical foundation for efficient diffusion-based language generation and offer principled insights into decoding strategy design.

Bio: Gen Li is currently an assistant professor in the Department of Statistics and Data Science at the Chinese University of Hong Kong. His research interests include diffusion model, generative AI and reinforcement learning.

Yongduan Song
Lingnan University / Chongqing University
yongduansong@ln.edu.hk

Several Key Issues In Neural Network-Driven Control Systems

This report focuses on two core scientific issues concerning neural-network-driven control systems: reliability and interpretability. First, it systematically elucidates the fundamental theories and core principles of neural-network-based learning control, and examines its prominent technical advantages in adaptive learning, global generalization, and real-time dynamic processing. Second, it summarizes the key bottlenecks that currently constrain the engineering deployment and practical application of neural-network-driven control systems. Building on this foundation, the report further investigates the significant effects of network architecture selection and training-strategy optimization on controller performance, with particular emphasis on approaches to enhancing the robustness of neural-network-driven controllers and mechanisms for ensuring their operational effectiveness.

Bio: Professor Yongduan Song is an International Member of the Chinese Academy of Engineering, an IEEE Fellow, and a Fellow of both the Chinese Association of Automation (CAA) and the Chinese Association for Artificial Intelligence (CAAI). He serves as Chair of the CAA Technical Committee on Trustworthy Control Systems, Editor-in-Chief of IEEE Transactions on Neural Networks and Learning Systems (TNNLS), a leading international journal in artificial intelligence, and Founding Editor-in-Chief of the Journal of Automation and Intelligence, an English-language journal selected for

the High-Starting-Point New Journal Program under the China Science and Technology Journal Excellence Action Plan. He previously served as one of six Langley Distinguished Professors at the National Institute of Aerospace, where he founded and directed the Center for Cooperative Systems. He holds over 100 granted patents in China, the United States, and Japan, and has authored 4 textbooks and 8 monographs. He has also led more than 10 key research projects at the provincial, ministerial, and national levels.

Fenglei Fan
City University of Hong Kong
fenglfan@cityu.edu.hk

Model Compression Via Dynamic Systems

Recently, the escalating demand for memory and computational resources by large models has presented formidable challenges for their deployment in resource-constrained environments. One prevalent way to serve large models is to develop effective model compression approaches that crop the size of large models while maintaining acceptable performance levels. Currently, the landscape of model compression methodologies predominantly revolves around four fundamental algorithms: pruning, quantization, knowledge distillation, and low-rank approximation whose basic frameworks have been established years ago. However, the compression efficacy of these methods is often capped or challenging to scale based on theoretical analysis. In this talk, we introduce a novel and general-purpose approach, referred to as hyper-compression, that redefines model compression as a problem of parameter representation. Specifically, we extend the concept of hypernets into what we term a 'hyperfunction'. Then, the hyperfunction is designed based on ergodic theory (ET). This advanced formulation leads to a performant algorithm that offers several distinct advantages, succinctly summarized as PNAS: 1) Preferable compression ratio; 2) No post-hoc retraining; 3) Affordable inference time; and 4) Short compression time. Lastly, in light of the observed stagnation in hardware Moore's Law, we conjecture "Moore's Law of Model Compression", i.e., the efficiency of model compression could double annually in the near future to meet the needs of large model era. We believe that model compression based on hyperfunction can play an important role in "Moore's Law of Model Compression".

Bio: Fenglei Fan is currently an Assistant Professor with Department of Data Science, City University of Hong Kong. He is the Chief Scientist of the

Huawei Key Project. His primary research interests lie in NeuroAI and its applications in model compression and medical imaging. He was the recipients of the IBM AI Horizon Scholarship, the 2021 International Neural Network Society Doctoral Dissertation Award, and He won OlympusMons Pioneering Award, a prestigious award in the field of storage. He has one paper selected as one of few 2024 CVPR Best Paper Award Candidates, one won the IEEE Nuclear and Plasma Society IEEE TRPMS Best Paper Award, and one ESI highly cited paper. He organized special issues in journals like IEEE TRPMS, presented three tutorials in AAAI2023, IJCNN25, and WWW2025, and served as (senior) program committee members in AAAI and IJCAI.

Ran Guan
Huawei

Introduction To openJiuwen And JiuwenSwarm

This talk introduces openJiuwen, an open-source platform for developing and running AI agents, and JiuwenSwarm, an intelligent AI agent built on openJiuwen. It will provide a brief overview of their capabilities and applications.

Bio: Ran Guan is a Principal Engineer at Huawei. His work focuses on agentic systems, wireless AI, and intelligent sensing, and he is involved in the development of Jiuwen.

Bohan Zhuang
Zhejiang University
bohan.zhuang@zju.edu.cn

Efficient Long-Context Inference For Multimodal Understanding And Generation

As multimodal foundation models process increasingly long contexts, their memory footprint and generation cost have become major barriers to deployment at scale. This talk presents our latest advances in efficient multimodal inference through algorithm–system co-design. We will discuss efficient attention mechanisms, including sparse and linear attention for long-context modeling; memory-efficient inference through KV-cache compression and management; and recent parallel decoding methods for accelerating generation. We will then connect these techniques to practical applications in

multimodal understanding and generation. Beyond conventional inference, the talk will also cover efficient diffusion-based reinforcement learning for aligning video generation models, as well as our exploration of evaluation benchmarks for interactive world models. Finally, we will share our perspective on the future of efficient foundation models, particularly the agentic loops.

Bio: Bohan Zhuang is a Research Professor at Zhejiang University. His research focuses on algorithm–system co-design for efficient foundation models, with the goal of maximizing value per token. He also conducts frontier research on agentic loops, world models, and embodied AI. Previously, he was a tenured faculty at Monash University, where he founded the ZIP Lab, and he maintains close research collaborations with leading global technology companies. The lab's first seven PhD graduates have all joined top-tier technology companies or universities, including DeepSeek and ByteDance Seed, with many selected for highly competitive talent programs. Many undergraduate alumni have pursued PhDs at world-leading universities, with one receiving a doctoral fellowship from UC Berkeley.

Liang Chen
Tongji University
liangchen527@gmail.com

Efficient VLM Tuning Via Rational Adaptation

Adapting large VLMs to downstream tasks is powerful but expensive, and almost all existing methods pay that cost by modifying dense intermediate features inside the text or vision encoders. In this talk, I argue for a different place to intervene: the final decision-making stage, where the two modalities actually fuse to produce a prediction. Starting from our earlier work on domain generalization, I will first introduce the rational matrix, a fine-grained view of how a model reaches its decisions, and show how it carries over naturally to VLMs. I will then present Rational Adaptation, which simply attaches one lightweight layer at the end to recalibrate this decision stream, leaving all intermediate features untouched. This shifts adaptation to where it matters most while keeping the footprint small, and I hope it opens up new possibilities for lightweight foundation-model tuning and fresh thinking on model efficiency.

Bio: Liang Chen is an Assistant Professor in the College of Computer Science and Technology at Tongji University. He received his Ph.D. from the

University of Adelaide, where he was awarded the Dean's Commendation for Doctoral Thesis Excellence, and subsequently conducted postdoctoral research at MBZUAI and the City University of Hong Kong. He is a recipient of the Shanghai Overseas High-Level Talent and Huawei Young Tech Talent programs. His research has focused on robust generalization in deep learning and, more recently, on 3D reconstruction. Dr. Chen has published in premier computer vision and machine learning venues, including CVPR, ICCV, ECCV, NeurIPS, and IJCV, and serves as a reviewer for these venues.