

Fanjiang Ye

WORK fy27@rice.edu **DAILY** yvanphys@gmail.com
PHONE +1 (812) 322-7150 **HOME** home.fanjiang.net
PROFILES [Google Scholar](#) [Linked In](#)

Education

Ph.D. Student in Computer Science

Rice University

Advisor: [Dr. Yuke Wang](#)

2025–May 2030 (Expected)

Houston, TX, USA

Ph.D. Student in Computer Engineering

Indiana University

Advisor: [Dr. Dingwen Tao](#)

2023–2025

Bloomington, IN, USA

B.S. in Physics

University of Science and Technology of China

Advisor: [Dr. Changling Zou](#)

2019–2023

Hefei, Anhui, China

Research Interests

- Efficient systems for generative AI
- System optimization for multimodal LLMs
- Data compression and communication in HPC/ML training

Publications

- **Preprint** [Fanjiang Ye](#), Zepeng Zhao, Yi Mu, Jucheng Shen, Renjie Li, Kaijian Wang, Desen Sun, Saurabh Agarwal, Myungjin Lee, Triston Cao, Aditya Akella, Arvind Krishnamurthy, T. S. Eugene Ng, Zhengzhong Tu, Yuke Wang. [SUPERGEN: An Efficient Ultra-high-resolution Video Generation System with Sketching and Tiling.](#) [arXiv](#)
Video Generation : Built SUPERGEN, a training-free tile-based framework with region-aware caching and communication minimized multi-GPU tile parallelism for efficient, high-quality ultra-high-resolution video generation.
- **NeurIPS'25** Jinda Jia, Cong Xie, Hanlin Lu, [Fanjiang Ye](#), Hao Feng, Daoce Wang, Haibin Lin, Zhi Zhang, Xin Liu. [DUO: No Compromise to Accuracy Degradation.](#) [To Appear](#)
Communication Compression in Distributed LLM : Introduced DUO: overlaps an extra high-precision gradient sync within compute to hide communication and recover accuracy under aggressive gradient compression.
- **ICML'25 Spotlight** Xiyuan Wei, Ming Lin, [Fanjiang Ye](#), Fengguang Song, Liangliang Cao, My T. Thai, Tianbao Yang. [Model Steering: Learning with a Reference Model Improves Generalization Bounds and Scaling Laws.](#) [arXiv](#)
CLIP Optimization : Formalized model steering (DRRho/DRO) and introduced DRRho-CLIP for reference-guided training with better generalization, data efficiency, and scaling.
- **ICS'25 Best Paper Runner-up** Boyuan Zhang, Bo Fang, [Fanjiang Ye](#), Luanzheng Guo, Fengguang Song, Tallent Nathan, Dingwen Tao. [BMQSim: Overcoming Memory Constraints in Quantum Circuit Simulation with a High-Fidelity Compression Framework.](#) [arXiv](#)
Compression in Quantum Computing : Designed BMQSim, a compression-aware quantum circuit simulator with GPU-based lossy compression, circuit partitioning, pipeline-integrated data movement, and two-level memory management.
- **SC'24** Hao Feng, Boyuan Zhang, [Fanjiang Ye](#), Min Si, Ching-Hsiang Chu, Jiannan Tian, Chunxing Yin, Zhaoxia (Summer) Deng, Yuchen Hao, Pavan Balaji, Tong Geng, Dingwen Tao. [Accelerating Communication in Deep Learning Recommendation Model Training with Dual-Level Adaptive Lossy Compression.](#) [doi](#)
Compression in DLRM : Accelerated DLRM with error-bounded compression for embedding all-to-all, via dual-level adaptive bounds and GPU-optimized tensors.
- **PPoPP'25 Poster** Boyuan Zhang, Luanzheng Guo, Jiannan Tian, Jinyang Liu, Daoce Wang, [Fanjiang Ye](#), Chengming Zhang, Jan Strube, Nathan R. Tallent, Dingwen Tao. [High-performance Visual Semantics Compression for AI-Driven Science.](#) [doi](#)
Compression in AI Science : Developed ViSemZ, a high-performance AI-based scientific-image compressor that preserves visual semantics via sparse encoding with variable-length truncation and optimized lossless coding.
- **Preprint** Xinrui Zhong, Xinze Feng, Jingwei Zuo, [Fanjiang Ye](#), Yi Mu, Junfeng Guo, Heng Huang, Myungjin Lee, Yuke Wang. [An Efficient and Adaptive Watermark Detection System with Tile-based Error Correction.](#) [arXiv](#)
Diffusion Watermarking : Designed QRMark, an adaptive tile-based watermark detector with QR code error correction

and resource-aware GPU scheduling for efficient, robust large-scale detection.

- **Preprint** Xiyuan Wei, **Fanjiang Ye**, Ori Yonay, Xingyu Chen, Dingwen Tao, Tianbao Yang. [FastCLIP: A Suite of Optimization Techniques to Accelerate CLIP Training with Limited Resources](#). [arXiv](#)

CLIP Optimization : Engineered FastCLIP, a distributed CLIP training framework that leverages compositional optimization and comm.-efficient gradient reduction for efficient training on limited resources.

Research Experience

Rice University, Yuke's Laboratory

Graduate Research Assistant

2025–Present
Houston, TX, USA

- Research in designing efficient system techniques for image/video generation.
- Exploring in heterogeneous and high-performance MLLM serving system.

Indiana University, HiPDAC Laboratory

Graduate Research Assistant

2023–2025
Bloomington, IN, USA

- Research in designing accelerator-based lossy compression for HPC/ML applications.
- Developing the efficient distributed CLIP training framework.

Coding Language

- Python
- C/C++
- CUDA

Professional Service

- ASPLOS'26 Artifact Evaluation Committee
- SOSP'25 Artifact Evaluation Committee
- QCE'24 Sub-Reviewer

Honors and Awards

- Indiana University Travel Awards (\$1500), Indiana University Bloomington 2024
- USTC Fellowship (\$2000), University of Science and Technology of China 2023
- Outstanding Student Scholarship (Top 25%), University of Science and Technology of China 2020, 2021, 2022

Teaching Experience

- Teaching Assistant of ENGR-E 516: Cloud Computing, Spring 2025, Indiana University