Fanjiang Ye

WORK fy27@rice.edu **DAILY** yvanphys@gmail.com
PHONE +1 (812) 322-7150 **HOME** home.fanjiang.net
PROFILES Google Scholar ▪ LinkedIn

Education

Ph.D. Student in Computer Science *2025–May 2030 (Expected)*
Rice Univeristy *Houston, TX, USA*
Advisor: Dr. Yuke Wang

M.S. in Computer Science *2023–2025*
Indiana University *Bloomington, IN, USA*
Advisor: Dr. Dingwen Tao

B.S. in Physics *2019–2023*
University of Science and Technology of China *Hefei, Anhui, China*
Advisor: Dr. Changling Zou

Research Interests

- **Generative AI**: super-resolution image/video generation, efficient diffusion inference (cache, sparsity, pruning, parallelism), long real-time video generation.
- **AI Infra Systems**: cost-friendly LLM/diffusion/unified model serving and scalable distributed training and inference.
- **Efficient LLM**: quantization and sparse attention kernel design.

Publications

- **[A1]** **Fanjiang Ye**, Zepeng Zhao, Yi Mu, Jucheng Shen, Renjie Li, Kaijian Wang, 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**
- **[C1]** Jinda Jia, Cong Xie, **Fanjiang Ye**, Hao Feng, Hanlin Lu, Daoce Wang, Haibin Lin, Zhi Zhang, Xin Liu. DUO: No Compromise to Accuracy Degradation. **NeurIPS'25**
- **[C2]** 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. **ICML'25 Spotlight**
- **[C3]** 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. **ICS'25 Best Paper Runner-up**
- **[C4]** 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. **SC'24**
- **[C5]** 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. **PPoPP'25 Poster**
- **[A2]** Yihua Liu, **Fanjiang Ye**, Bowen Lin, Rongyu Fang, Chengming Zhang. TIDE: Text-Informed Dynamic Extrapolation with Step-Aware Temperature Control for Diffusion Transformers. **arXiv**
- **[A3]** Xinheng Ding, Tianyi Zhang, Hao Li, **Fanjiang Ye**, Wenya Xie, Qingquan Song, Yuke Wang, Zirui Liu. Teach to Fish, Not to Feed Internalizing Retrieval for External Memory in Large Language Models. **arXiv**
- **[A4]** 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**
- **[A5]** Bowen Lin, **Fanjiang Ye**, Yihua Liu, Zhenghui Guo, Boyuan Zhang, Weijian Zheng, Yufan Xu, Tiancheng Xing, Yuke Wang, Chengming Zhang. SDiT: Semantic Region-Adaptive for Diffusion Transformers. **arXiv**
- **[A6]** 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**

Experience

Rice University, **Yuke's Laboratory** *2025–Present*
Graduate Research Assistant *Houston, TX, USA*

- *Efficient Video Generation*: built 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 [A1].
- *Efficient Image Generation*: proposed a semantic region-adaptive diffusion transformer to allocate computation according to regional complexity [A5].
- *Diffusion Watermarking*: designed an adaptive tile-based watermark detector with QR code error correction and resource-aware GPU scheduling for efficient, robust large-scale detection [A4].

Indiana University, **HiPDAC Laboratory** *2023–2025*
Graduate Research Assistant *Bloomington, IN, USA*

- *Efficient Distributed LLM*: introduced an extra high-precision gradient sync within compute to hide communication and recover accuracy under aggressive gradient compression [C1].
- *Efficient Distributed CLIP training*: introduced 1) reference-guided training with better generalization, data efficiency, and scaling [C2]. 2) a distributed CLIP training framework that leverages compositional optimization and communication-efficient gradient reduction for efficient training on limited resources [A6].
- *Compression in AI applications and Quantum Computing*: [C3] [C4] [C5]

Coding Language

- Python
- C/C++
- CUDA, Triton

Professional Service

- Artifact Evaluation Committee: EuroSys'2026 Fall, ASPLOS'26 Summer, PPoPP'26, ASPLOS'26 Spring, SOSP'25
- Program Committee: ECCV'26, CVPR'26, QCE'24