

## Education

### Ph.D. Student in Computer Science

Rice University

Advisor: Dr. Yuke Wang

2025–May 2030 (Expected)  
Houston, TX, USA

### M.S. in Computer Science

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

- **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, Kajian 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] 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](#)
- [A3] 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](#)
- [A4] 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](#)

## Research Experience

### Rice University, Yuke's Laboratory

Graduate Research Assistant

2025–Present  
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 [A3].
- *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 [A2].

**Indiana University, HiPDAC Laboratory**

Graduate Research Assistant

2023–2025

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 [A4].
- *Compression in AI applications and Quantum Computing*: [C3] [C4] [C5]

## Coding Language

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

## Professional Service

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