Kaijian Wang

kw118@rice.edu · (+1) 858-250-9160 · albedowang.github.io

Education

Rice University, Texas, United States 2025 – Present

Ph.D Student in Computer Science, *Advisor:* Prof. Yuke Wang

University of Science and Technology of China, Anhui, China 2021 – 2025

B.E. in Information Security

Courses: Data Structure and Algorithm, Operating System, Compiler Principle, Machine Learning, Network Security Protocol, Computer Network etc.

Publications

- **RecDM:** Efficient Training System for Large-Scale Recommendation Models on Disaggregated Memory (In submission)
  *Zheng Wang, Zhongkai Yu, **Kaijian Wang**, Yichen Lin, Yikai Li, Liu Liu, Xulong Tang, Yuke Wang, Yangwook Kang, Yufei Ding*
- **SuperGen:** An Efficient Ultra-high-resolution Video Generation System with Sketching and Tiling [arXiv:2508.17756]
  *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*

Experience

DiT Serving on TPU (On-going, Collaborating with Google TPU team) Sep. 2025 – Now

Research Assistant at Rice Advisors: Prof. Yuke Wang, Hongmin Fan, Yarong Mu

- Deploying Diffusion Transformers(DiTs) on TPU by designing parallelism mapping that exploits high inter-device bandwidth and addresses limited on-chip memory for efficient distributed inference.

Efficient Ultra-high-Resolution Diffusion Video Generation May 2025 – Aug. 2025

Research Assistant at Rice Advisor: Prof. Yuke Wang

- Conducted preliminary experiments to evaluate the feasibility of diffusion-based video generation at ultra-high resolutions.
- Assisted in the design of a distributed multi-GPU execution strategy, including workload partitioning and communication scheduling, to improve efficiency in large-scale diffusion video generation.

Efficient Tool-Augmented LLM Serving System Feb. 2024 – Dec. 2024

Research Intern at UC San Diego Advisors: Prof. Yufei Ding, Zheng Wang

- Developed an inference method combining speculative models with tool engines to optimize memory allocation and request scheduling, significantly improving the efficiency of CoT and tool usage in LLM inference serving.
- Achieved **30-40%** speed improvement over using only large models by implementing collaborative reasoning between large and small models.
- Played a key role in implementing nearly the entire code framework over vLLM, profiling system data and performing evaluations.
- Optimized inefficient memcpy kernels in vLLM, achieving a **2x** speed improvement.

Efficient DLRM Training System over CXL Aug. 2024 – Nov. 2024

Research Intern at UC San Diego Advisors: Prof. Yufei Ding, Zheng Wang

- Proposed a novel solution optimizing DLRM training on disaggregated memory, achieving a **5.34x** end-to-end training step latency improvement over traditional frameworks.
- Leveraged CXL technology for flexible memory expansion and introduced an access-aware strategy for embedding placement.
- I formulated embedding table placement for an ILP problem with multiple constraints e.g. memory capacity and bandwidth and achieve **1.37x** improvement over arbitrary place embedding table. And I completed the evaluation part of the project.

Relocation-Robust Deep Learning Watermark Aug. 2023 – Feb. 2024

Research Assistant at USTC Advisor: Prof. Kejiang Chen

- Proposed a watermark recovery module based on image hiding, designed to restore information-embedded images subjected to cropping and spatial manipulation.
- Completed experiments on related research and developed the module framework code.
- The final watermarked image achieved a PSNR greater than 38, and the hidden information extraction accuracy greater than 99%.

CTF Team (USTC-NEBULA) Mar. 2023 – Jun. 2025

Team Member, Main direction: Misc

- Researched software reverse engineering, binary analysis, cryptography, and image processing.
- Participated in several competitions; team ranked 16/1655 (top 1%) in the China CTFTime rating for 2023-2024.

Tiny-C compiler (Course project) Sep. 2023 – Oct. 2023

- Developed a simplified C compiler in C, implementing lexical analysis, syntax analysis, and Intermediate Representation (IR) generation for the Tiny-C language.

Scholarships and Awards

USTC-Fellowship *(~7.5k USD)*, USTC	May. 2025
Outstanding Student Scholarship, USTC	Oct. 2024
DAS Scholarship, USTC & DAS-Security Inc.	Nov. 2023
Outstanding Student Scholarship, USTC	Oct. 2023
Zhang Zongzhi Science and Technology Scholarship, USTC	Oct. 2022
2nd Prize (top 5 /389), Hackergame2024, USTC	Nov. 2024
Top 3rd(1%), TPCTF, Tsinghua University & Peking University	Nov. 2023
3rd Prize (top 10%), Hackergame2023, USTC	Nov. 2023
3rd Prize in competition area (top 20%),	
- National College Student Information Security Contest, China	June. 2023
3rd Prize (top 20%), D^3CTF, Alibaba Inc.	May. 2023

Skills

- Programming Language: Python, C/C++, SQL, HTML, LaTeX
- ML related: vLLM, PyTorch, DLRM, torchax
- Research Interest: Efficient LLM Serving/Training, Distributed Systems, Heterogeneous Devices