Jiahao Fang
Zhejiang, China—(217) 418-1139—jiahaof3@illinois.edu—*Homepage*

EDUCATION

University of Illinois Urbana-Champaign **Illinois, USA**
Graduate, The Master of Science (M.S.) in *Computer Science*, GPA 3.89/4.00 **August 2024 - May 2026 (Expected)**

University of Illinois Urbana-Champaign **Illinois, USA**
Undergraduate, Major in *Electrical Engineering*, Minor in *Computer Science*, GPA 3.89/4.00 **August 2020 - May 2024**

Zhejiang University **Zhejiang, China**
Undergraduate, Major in *Electrical Engineering and Automation*, GPA 3.90/4.00 **September 2020 - June 2024**

RESEARCH EXPERIENCE

KV Cache Management for LLM Serving **August 2024 - Present**
Master Research and Thesis Supervised by **Professor Fan Lai**
Contribution to Conference Publication, University of Illinois Urbana-Champaign

- Minimize Time to First Token (TTFT) for both Large Language Model (LLM) and Multimodal Large Language Model (MLLM)
- Overlap cache retrieval from multiple storage backends with the GPU-bound recomputing prefill

The Application of Machine Learning to Datebase Query **May 2023 - May 2024**
Undergraduate Research Assistant in Computer Science Department Supervised by **Professor Daniel Kang**
Contribution to Conference Publication, University of Illinois Urbana-Champaign

- Implemented and reproduced functioning code of specify approximate selection with guarantees (SUPG) to prepare for Approximate Selection with Guarantees using Proxies in *GitHub*
- Generated experiment datasets, modify codes, and did experiments about query optimization for Applied AI for Database Systems and Applications **(AIDB)**

Self-study for Machine Learning and Computational Imaging **January - May 2023**
ECE 397, Individual Study in ECE Problems Supervised by **Professor Jane Zhao**
Independent Study Course about image processing, University of Illinois Urbana-Champaign

- Developed object detection Machine Learning models in different types of unstructured datasets, such as images, audios, and videos
- Summaried advantages of various Machine Learning algorithms based on math proof in *Course Report*

CONFERENCE PUBLICATION

Jiahao Fang, Sean Nian, Qilong Feng, Zhiyu Wu, **Fan Lai***.
Cross-layer KV Cache Parallelism for LLM Serving at Scale. (Expected) submitted to 20th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Seattle, USA, July 13–15, 2026.

Tengjun Jin, Akash Mittal, Chenghao Mo, **Jiahao Fang**, Chengsong Zhang, Timothy Dai, **Daniel Kang***.
AIDB: a Sparsely Materialized Database for Queries using Machine Learning. accepted to 8th Data Management for End-to-End Machine Learning (DEEM) workshop at SIGMOD, Santiago, Chile, June 9th, 2024.

HONORS AND AWARDS

The Third Prize of Zhejiang University Academic Excellence **August 2022 - May 2023**

The Third Prize of Zhejiang University Academic Excellence **September 2020 - June 2021**

Meritorious Winner (Top 7 percent) **January 2021**
The Mathematical Contest in Modeling in Problem A: Fungi

COMPUTER SKILLS

Basic Languages:	Python, C++
Database:	MySQL, MongoDB, Neo4j (SQL and NoSQL)
Web Design:	HTML, CSS, JavaScript