Zili Zhang

Ph.D. student · Peking University

Yanyuan Building 818, No.5 Yiheyuan Road, Haidian District, Beijing, Republic of China

zzlcs@pku.edu.cn | www.zilizhang.site | Gold-Sea | Zili Zhang

Education

Peking University — *Beijing, China*

Ph.D. in Computer Science and Engineering — *Sep. 2023 - Present*

- Advisor: Prof. Xin Jin

Peking University — *Beijing, China*

B.E. in Computer Science and Engineering — *Sep. 2019 - Jun. 2023*

- Overall GPA: 3.71/4.0 (top 9%)

Service

2025 **Shadow PC,** The 20th edition of EuroSys (EuroSys 2025) — *Rotterdam*

Publications

StreamRL: Scalable, Heterogeneous, and Elastic RL for LLMs with Disaggregated Stream Generation

In Preprint — *Apr. 2025*

- Yinmin Zhong, **Zili Zhang**, Xiaoniu Song, Hanpeng Hu, Chao Jin, Bingyang Wu, Nuo Chen, Yukun Chen, Yu Zhou, Changyi Wan, Hongyu Zhou, Yimin Jiang, Yibo Zhu, Daxin Jiang

RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation

In Preprint — *Apr. 2024*

- Chao Jin, **Zili Zhang**, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, Xin Jin

Fast Distributed Inference Serving for Large Language Models

NSDI'26 — *Apr. 2026*

- Bingyang Wu*, Yinmin Zhong*, **Zili Zhang***, Gang Huang, Xuanzhe Liu, Xin Jin (* indicates equal contribution)

DistTrain: Addressing Model and Data Heterogeneity with Disaggregated Training for Multimodal Large Language Models

SIGCOMM'25 — *Sept. 2025*

- **Zili Zhang**, Yinmin Zhong, Yimin Jiang, Hanpeng Hu, Jianjian Sun, Zheng Ge, Yibo Zhu, Daxin Jiang, Xin Jin

RLHFuse: Efficient RLHF Training for Large Language Models with Inter- and Intra-Stage Fusion — *Philadelphia, PA, U.S.*

NSDI'25 — *Apr. 2025*

- Yinmin Zhong, **Zili Zhang**, Bingyang Wu, Shengyu Liu, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, Xin Jin

Fast Vector Query Processing for Large Datasets Beyond GPU Memory with Reordered Pipelining — *Santa Clara, U.S.*

NSDI'24 — *Apr. 2024*

- **Zili Zhang**, Fangyue Liu, Gang Huang, Xuanzhe Liu, Xin Jin

Jolteon: Unleashing the Promise of Serverless for Serverless Workflows — *Santa Clara, U.S.*

NSDI'24 — *Apr. 2024*

- **Zili Zhang**, Chao Jin, Xin Jin

dLoRA: Dynamically Orchestrating Requests and Adapters for LoRA LLM Serving — *Santa Clara, U.S.*

OSDI'24 — *Jul. 2024*

- Bingyang Wu, Ruidong Zhu, **Zili Zhang**, Peng Sun, Xuanzhe Liu, Xin Jin

Fast, Approximate Vector Queries on Very Large Unstructured Datasets — *Boston, U.S.*

NSDI'23 — *Apr. 2023*

- **Zili Zhang**, Chao Jin, Linpeng Tang, Xuanzhe Liu, Xin Jin

Ditto: Efficient Serverless Analytics with Elastic Parallelism — *New York City, U.S.*

SIGCOMM'23 — *Apr. 2023*

- Chao Jin, **Zili Zhang**, Xingyu Xiang, Songyun Zou, Gang Huang, Xuanzhe Liu, Xin Jin

Transparent GPU Sharing in Container Clouds for Deep Learning Training *Boston, U.S.*

NSDI'23 *Apr. 2023*

- Bingyang Wu, **Zili Zhang**, Zhihao Bai, Xuanzhe Liu, Xin Jin

Rise of Distributed Deep Learning Training in the Big Model Era: From A Software Engineering Perspective

TOSEM'23 *May. 2023*

- Xuanzhe Liu , Diandian Gu, Zhenpeng Chen, Jinfeng Wen, **Zili Zhang**, Yun Ma, Haoyu Wang, Xin Jin

Optimizing Half Precision Winograd Convolution on ARM Many-Core Processors *Online*

APSYS'22 *Aug. 2022*

- Dedong Xie, Zhen Jia, **Zili Zhang**, Xin Jin

Internship

ByteDance Seed. *Beijing, China*

RESEARCH INTERN OF MULTIMODAL LLM TRAINING *Apr. 2025 - Present*

- Optimization for Multimodal and RL Training

StepFun. *Beijing, China*

RESEARCH INTERN OF LLM TRAINING *Apr. 2024 - Apr. 2025*

- Optimization for Multimodal and RL Training

Alibaba. *Beijing, China*

RESEARCH INTERN OF SERVERLESS GPU *Jun. 2023 - Apr. 2024*

- Optimization for Alibaba Serverless Computing Platform

Moqi. *Beijing, China*

RESEARCH INTERN OF VECTOR DATABASE AND RAG *Oct. 2021 - Feb. 2023*

- Optimization for Faiss runtime

ByteDance Inc. *Beijing, China*

RESEARCH INTERN OF DL COMPILER *Jun. 2021 - Sep. 2021*

- Optimization for TVM compiler runtime

Teaching

Operating System (Honor Track) *Peking University*

TEACHING ASSISTANT *Feb. 2024 - July. 2024*

- Correcting students' homework and assisted teachers in the examination paper.
- Organizing the seminar and prepared topics for discussion.

Introduction to Computer System (Honor Track) *Peking University*

TEACHING ASSISTANT *Sep. 2022 - Jan. 2023*

- Correcting students' homework and assisted teachers in the examination paper.
- Organizing the seminar and prepared topics for discussion.
- Coding for the course project.

Introduction to Computer System *Peking University*

TEACHING ASSISTANT *Sep. 2021 - Jan. 2022*

- Correcting students' homework and assisted teachers in the examination paper.
- Organizing the seminar and prepared topics for discussion.
- Coding for the course project.

Honors & Awards

UNIVERSITY AWARDS

2024 **Merit Student of Peking University (1/56)**, PKU Award *Peking University*
2024 **National Scholarship (Top 0.4% nationally)**, PKU Award *Peking University*
2024 **Presidential Scholarship of Peking University (highest honor for Ph.D. students)**, PKU Award *Peking University*
2023 **Top 10 Bachelor Thesis of Peking University, EECS (10/408)**, Graduation Ceremony *Peking University*
2023 **Outstanding Graduation Thesis of Peking University (33/4239)**, Graduation Ceremony *Peking University*
2023 **Representor of Excellent Graduates (14/4239)**, Graduation Ceremony *Peking University*
2023 **Excellent Graduates (613/4239)**, Graduation Ceremony *Peking University*
2022 **Exceptional Award for Academic Innovation (5/408)**, PKU Award *Peking University*
2022 **Award for Scientific Research**, PKU Award *Peking University*
2022 **Lee Wai Wing Scholarship**, PKU Award *Peking University*
2021 **Award for Scientific Research**, PKU Award *Peking University*
2020 **Award for Academic Excellents**, PKU Award *Peking University*
2020 **Third Prize**, PKU Award *Peking University*

Skills & Interests

Programming AWS, Docker, Kubernetes, C++, Python, CUDA, Java, Golang, Node.js, Latex
Languages Chinese (native), English
Interests Running (www.itra.run/RunnerSpace/ZHANG.Zili/4938114), Climbing, Traveling, Video Games