

JINTAO ZHANG

ZiQiang Technology Building, Tsinghua University, Haidian District, Beijing, China 100084
✉ zhang-jt24@mails.tsinghua.edu.cn | 📞 +86 185-1911-5711 (WeChat ID) | 🌐 <https://jt-zhang.github.io>

EDUCATION

Tsinghua University Ph.D. of Science in Computer Science and Technology Supervisor: Prof. Jun Zhu 🌐 Homepage , and Prof. Jianfei Chen 🌐 Homepage .	Aug. 2024 – Jun.2026 (Expected) Beijing, China
Tsinghua University Master of Science in Computer Science and Technology (GPA: 3.92) Supervisor: Prof. Guoliang Li 🌐 Homepage .	Aug. 2021 – Jun. 2024 Beijing, China
Xidian University Bachelor of Science in Computer Science and Technology (GPA: 3.85)	Sep. 2017 – Jun. 2021 Shannxi, China

SELECTED PUBLICATIONS

* indicates equal contributions

[Aixiv] Full Paper	SageAttention2: Efficient Attention with Thorough Outlier Smoothing and Per-thread INT4 Quantization. Jintao Zhang , Haofeng Huang, Penge Zhang, Jia Wei, Jun Zhu, Jianfei Chen Aixiv, 2024. [PDF] [CODE] .
[ICLR 2025] Full Paper	SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration. [PDF] [CODE] Jintao Zhang , Jia Wei, Penge Zhang, Jun Zhu, Jianfei Chen International Conference on Learning Representations, 2025
[ICDE 2025] Full Paper	SAGE: A Framework of Precise Retrieval for RAG. [PDF] Jintao Zhang , Guoliang Li, Jinyang Su 41th IEEE International Conference on Data Engineering, 2025
[SIGMOD 2024] Full Paper	PACE: Poisoning Attacks on Learned Cardinality Estimation. [PDF] Jintao Zhang , Guoliang Li, Chao Zhang, Chengliang Chai ACM SIGMOD International Conference on Management of Data, 2024
[ICDE 2023] Full Paper	AutoCE: An Accurate and Efficient Model Advisor for Learned Cardinality Estimation. [PDF] Jintao Zhang , Chao Zhang, Guoliang Li, Chengliang Chai 39th IEEE International Conference on Data Engineering, 2023
[VLDB 2022] Full Paper	Learned Cardinality Estimation: Design Space Exploration and Comparative Evaluation. [PDF] Ji Sun*, Jintao Zhang* , Zhaoyan Sun, Guoliang Li, Nan Tang. 48th International Conference on Very Large Databases, 2022

RESEARCH STATEMENT

- My research interests focus on Efficient Machine Learning System, specifically on accelerating large models' training and inference.

SELECTED AWARDS

Siebel Scholars Scholarship (35,000 USD, Less than 100 globally/year)	May. 2023
President's Award of Xidian University (Top 0.02%)	Dec. 2020
China National Scholarship	2019 & 2020

PROFESSIONAL EXPERIENCES

- **Teacher Assistant.** "Programming Fundamentals" in Tsinghua University. Feb. 2022 – Present
- **Oral Reports.** SIGMOD 2024, ICDE 2023, and VLDB 2022.