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   **Aug. 2024 – Jun.2026 (Expected)**
Ph.D. of Science in Computer Science and Technology   **Beijing, China**
Supervisor: Prof. Jun Zhu 🔗 Homepage, and Prof. Jianfei Chen 🔗 Homepage.

Tsinghua University   **Aug. 2021 – Jun. 2024**
Master of Science in Computer Science and Technology (GPA: 3.92)   **Beijing, China**
Supervisor: Prof. Guoliang Li 🔗 Homepage.

Xidian University   **Sep. 2017 – Jun. 2021**
Bachelor of Science in Computer Science and Technology (GPA: 3.85)   **Shannxi, China**

SELECTED PUBLICATIONS

* indicates equal contributions

[Aixiv] SageAttention2 Technical Report: Accurate 4-Bit Attention for Plug-and-play Inference Acceleration.
Full Paper **Jintao Zhang**, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, Jianfei Chen
Aixiv, 2024. [PDF] | [CODE].

[Aixiv] SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration.
Full Paper **Jintao Zhang**, Jia Wei, Haofeng Huang, Pengle Zhang, Jun Zhu, Jianfei Chen
Aixiv, 2024. [PDF] | [CODE].

[ICDE 2025] SAGE: A Framework of Precise Retrieval for RAG. [PDF]
Full Paper **Jintao Zhang**, Guoliang Li, Jinyang Su
41th IEEE International Conference on Data Engineering, 2025

[SIGMOD 2024] PACE: Poisoning Attacks on Learned Cardinality Estimation. [PDF]
Full Paper **Jintao Zhang**, Guoliang Li, Chao Zhang, Chengliang Chai
ACM SIGMOD International Conference on Management of Data, 2024

[ICDE 2023] AutoCE: An Accurate and Efficient Model Advisor for Learned Cardinality Estimation.[PDF]
Full Paper **Jintao Zhang**, Chao Zhang, Guoliang Li, Chengliang Chai
39th IEEE International Conference on Data Engineering, 2023

[VLDB 2022] Learned Cardinality Estimation: Design Space Exploration and Comparative Evaluation. [PDF]
Full Paper 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.