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 – Dec.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

[Arxiv 2025]
Full Paper
SageAttention3: Microscaling FP4 Attention for Inference and An Exploration of 8-Bit Training.
Jintao Zhang, Jia Wei, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Haoxu Wang, Kai Jiang, Jun Zhu, Jianfei Chen

[ICML 2025]
Full Paper, CCF-A
SpargeAttention: Accurate Sparse Attention Accelerating Any Model Inference. [code (0.7K Stars)]
Jintao Zhang, Chendong Xiang, Haofeng Huang, Haocheng Xi, Jia Wei, Jun Zhu, Jianfei Chen.

[ICML 2025]
Full Paper, CCF-A
SageAttention2: Efficient Attention with Thorough Outlier Smoothing and Per-thread INT4 Quantization
Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, Jianfei Chen.

[ICLR 2025]
Full Paper, CCF-A
SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration.[code (2K+ Stars)]
Jintao Zhang, Jia Wei, Pengle Zhang, Jun Zhu, Jianfei Chen.

[ICDE 2025]
Full Paper, CCF-A
SAGE: A Framework of Precise Retrieval for RAG.
Jintao Zhang, Guoliang Li, Jinyang Su.

[SIGMOD 2024]
Full Paper, CCF-A
PACE: Poisoning Attacks on Learned Cardinality Estimation.
Jintao Zhang, Guoliang Li, Chao Zhang, Chengliang Chai.

[ICDE 2023]
Full Paper, CCF-A
AutoCE: An Accurate and Efficient Model Advisor for Learned Cardinality Estimation.
Jintao Zhang, Chao Zhang, Guoliang Li, Chengliang Chai.

[TKDE 2025]
Full Paper, CCF-A
A Lightweight Learned Cardinality Estimation Model.
Yaoyu Zhu, **Jintao Zhang#**, (*corresponding author*) Guoliang Li#, Jianhua Feng.

[VLDB 2022]
Full Paper, CCF-A
Learned Cardinality Estimation: Design Space Exploration and Comparative Evaluation.
Ji Sun*, **Jintao Zhang*** (*equal contribution*), Zhaoyan Sun, Guoliang Li, Nan Tang.

[ES-FoMo III]
Short Paper
SageAttention2++: A More Efficient Implementation of SageAttention2.
Jintao Zhang, Xiaoming Xu, Jia Wei, Haofeng Huang, Pengle Zhang, Chendong Xiang, Jun Zhu, Jianfei Chen.

RESEARCH STATEMENT

- My research focuses on (1) Efficient Machine Learning System, specializing in accelerating model training and inference; (2) Data Management, specializing in query optimization and high-quality data acquisition.

SELECTED AWARDS

Siebel Scholars Scholarship (35,000 USD, Less than 100 globally/year)	May. 2023
China National Scholarship	2019 & 2020

PROFESSIONAL EXPERIENCES AND OTHERS

- **Teacher Assistant.** "Programming Fundamentals" in Tsinghua University for three years.
- **Oral Reports.** ICLR 2025, ICDE 2025, SIGMOD 2024, ICDE 2023, and VLDB 2022.   TOEFL Score: 99.