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

* indicates equal contributions

[ICML 2025] Full Paper	SpargeAttn: Accurate Sparse Attention Accelerating Any Model Inference. **Jintao Zhang**, Chendong Xiang, Haofeng Huang, Haocheng Xi, Jia Wei, Jun Zhu, Jianfei Chen Aixiv, 2025. [PDF] ∣ [CODE (0.6K Stars)].
[ICML 2025] Full Paper	SageAttention2: Efficient Attention with Thorough Outlier Smoothing and Per-thread INT4 Quantization. **Jintao Zhang**, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, Jianfei Chen Aixiv, 2024. [PDF] ∣ [CODE (2K Stars)].
[ICLR 2025] Full Paper	SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration. **Jintao Zhang**, Jia Wei, Pengle Zhang, Jun Zhu, Jianfei Chen International Conference on Learning Representations, 2025. [PDF] ∣ [CODE (2K Stars)]
[ICDE 2025] Full Paper	SAGE: A Framework of Precise Retrieval for RAG. **Jintao Zhang**, Guoliang Li, Jinyang Su 41th IEEE International Conference on Data Engineering, 2025. [PDF]
[SIGMOD 2024] Full Paper	PACE: Poisoning Attacks on Learned Cardinality Estimation. **Jintao Zhang**, Guoliang Li, Chao Zhang, Chengliang Chai ACM SIGMOD International Conference on Management of Data, 2024. [PDF]
[ICDE 2023] Full Paper	AutoCE: An Accurate and Efficient Model Advisor for Learned Cardinality Estimation. **Jintao Zhang**, Chao Zhang, Guoliang Li, Chengliang Chai 39th IEEE International Conference on Data Engineering, 2023. [PDF]
[VLDB 2022] Full Paper	Learned Cardinality Estimation: Design Space Exploration and Comparative Evaluation. Ji Sun*, **Jintao Zhang***, Zhaoyan Sun, Guoliang Li, Nan Tang. 48th International Conference on Very Large Databases, 2022. [PDF]

RESEARCH STATEMENT

- My research interests focus on Efficient Machine Learning System, specifically on accelerating training and inference of large models, including language, image, and video generation models.

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 for three years.
- **Oral Reports.** ICLR 2025, ICDE 2025, SIGMOD 2024, ICDE 2023, and VLDB 2022.