

# 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

<b>Tsinghua University</b> <b>Ph.D. of Science in Computer Science and Technology</b> <b>Supervisor:</b> Prof. Jun Zhu <a href="#">🔗 Homepage</a> , and Prof. Jianfei Chen <a href="#">🔗 Homepage</a> .	<b>Aug. 2024 – Dec.2026 (Expected)</b> <b>Beijing, China</b>
<b>Tsinghua University</b> <b>Master of Science in Computer Science and Technology (GPA: 3.92)</b> <b>Supervisor:</b> Prof. Guoliang Li <a href="#">🔗 Homepage</a> .	<b>Aug. 2021 – Jun. 2024</b> <b>Beijing, China</b>
<b>Xidian University</b> <b>Bachelor of Science in Computer Science and Technology (GPA: 3.85)</b>	<b>Sep. 2017 – Jun. 2021</b> <b>Shannxi, China</b>

## SELECTED PUBLICATIONS

\* indicates equal contributions

<b>[Arxiv 2025]</b> Full Paper	SageAttention3: Microscaling FP4 Attention for Inference and An Exploration of 8-Bit Training. <b>Jintao Zhang</b> , Jia Wei, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Haoxu Wang, Kai Jiang, Jun Zhu, Jianfei Chen Aixiv, 2025. <a href="#">[PDF]</a>   <a href="#">[CODE (2K Stars)]</a> .
<b>[ICML 2025]</b> Full Paper	SparseAttn: Accurate Sparse Attention Accelerating Any Model Inference. <b>Jintao Zhang</b> , Chendong Xiang, Haofeng Huang, Haocheng Xi, Jia Wei, Jun Zhu, Jianfei Chen Aixiv, 2025. <a href="#">[PDF]</a>   <a href="#">[CODE (0.6K Stars)]</a> .
<b>[ICML 2025]</b> Full Paper	SageAttention2: Efficient Attention with Thorough Outlier Smoothing and Per-thread INT4 Quantization. <b>Jintao Zhang</b> , Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, Jianfei Chen Aixiv, 2024. <a href="#">[PDF]</a>   <a href="#">[CODE (2K Stars)]</a> .
<b>[ICLR 2025]</b> Full Paper	SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration. <b>Jintao Zhang</b> , Jia Wei, Pengle Zhang, Jun Zhu, Jianfei Chen International Conference on Learning Representations, 2025. <a href="#">[PDF]</a>   <a href="#">[CODE (2K Stars)]</a>
<b>[ICDE 2025]</b> Full Paper	SAGE: A Framework of Precise Retrieval for RAG. <b>Jintao Zhang</b> , Guoliang Li, Jinyang Su 41th IEEE International Conference on Data Engineering, 2025. <a href="#">[PDF]</a>
<b>[SIGMOD 2024]</b> Full Paper	PACE: Poisoning Attacks on Learned Cardinality Estimation. <b>Jintao Zhang</b> , Guoliang Li, Chao Zhang, Chengliang Chai ACM SIGMOD International Conference on Management of Data, 2024. <a href="#">[PDF]</a>
<b>[ICDE 2023]</b> Full Paper	AutoCE: An Accurate and Efficient Model Advisor for Learned Cardinality Estimation. <b>Jintao Zhang</b> , Chao Zhang, Guoliang Li, Chengliang Chai 39th IEEE International Conference on Data Engineering, 2023. <a href="#">[PDF]</a>
<b>[VLDB 2022]</b> Full Paper	Learned Cardinality Estimation: Design Space Exploration and Comparative Evaluation. Ji Sun*, <b>Jintao Zhang*</b> , Zhaoyan Sun, Guoliang Li, Nan Tang. 48th International Conference on Very Large Databases, 2022. <a href="#">[PDF]</a>

## 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

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