

# 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

[Arxiv 2025] Full Paper	<a href="#">SageAttention3</a> : 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
[ICML 2025] Full Paper, CCF-A	<a href="#">SpargeAttention</a> : Accurate Sparse Attention Accelerating Any Model Inference. [ <a href="#">code (0.7K Stars)</a> ] <b>Jintao Zhang</b> , Chendong Xiang, Haofeng Huang, Haocheng Xi, Jia Wei, Jun Zhu, Jianfei Chen.
[ICML 2025] Full Paper, CCF-A	<a href="#">SageAttention2</a> : 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.
[ICLR 2025] Full Paper, CCF-A	<a href="#">SageAttention</a> : Accurate 8-Bit Attention for Plug-and-play Inference Acceleration.[ <a href="#">code (2K+ Stars)</a> ] <b>Jintao Zhang</b> , Jia Wei, Pengle Zhang, Jun Zhu, Jianfei Chen.
[ICDE 2025] Full Paper, CCF-A	<a href="#">SAGE</a> : A Framework of Precise Retrieval for RAG. <b>Jintao Zhang</b> , Guoliang Li, Jinyang Su.
[SIGMOD 2024] Full Paper, CCF-A	<a href="#">PACE</a> : Poisoning Attacks on Learned Cardinality Estimation. <b>Jintao Zhang</b> , Guoliang Li, Chao Zhang, Chengliang Chai.
[ICDE 2023] Full Paper, CCF-A	<a href="#">AutoCE</a> : An Accurate and Efficient Model Advisor for Learned Cardinality Estimation. <b>Jintao Zhang</b> , Chao Zhang, Guoliang Li, Chengliang Chai.
[TKDE 2025] Full Paper, CCF-A	A Lightweight Learned Cardinality Estimation Model. Yaoyu Zhu, <b>Jintao Zhang</b> <sup>#</sup> ( <i>corresponding author</i> ), Guoliang Li <sup>#</sup> , Jianhua Feng.
[VLDB 2022] Full Paper, CCF-A	<a href="#">Learned Cardinality Estimation</a> : Design Space Exploration and Comparative Evaluation. Ji Sun*, <b>Jintao Zhang</b> * ( <i>equal contribution</i> ), Zhaoyan Sun, Guoliang Li, Nan Tang.
[ICMLW 2025] Short Paper	<a href="#">SageAttention2++</a> : A More Efficient Implementation of SageAttention2. <b>Jintao Zhang</b> , 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.