

# JINTAO ZHANG

✉ 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</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</b>	<b>Sep. 2017 – Jun. 2021</b> <b>Shaanxi, China</b>

## SELECTED PUBLICATIONS

<b>[Arxiv 2025]</b> Technical Report	<b>TurboDiffusion:</b> Accelerating Video Diffusion Models by 100-200 Times. <a href="#">[Code (3.2K Stars)]</a> <b>Jintao Zhang</b> , Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, Jun Zhu
<b>[Arxiv 2025]</b> Full Paper	<b>SLA:</b> Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention. <b>Jintao Zhang</b> , Haoxu Wang, Kai Jiang, · · ·, Ion Stoica, Joseph E. Gonzalez, Jun Zhu, Jianfei Chen
<b>[NeurIPS 2025]</b> Spotlight, CCF-A	<b>SageAttention3:</b> 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
<b>[ICML 2025]</b> Full Paper, CCF-A	<b>SpargeAttention:</b> Accurate Sparse Attention Accelerating Any Model Inference. <a href="#">[Code (0.9K Stars)]</a> <b>Jintao Zhang</b> , Chendong Xiang, Haofeng Huang, Haocheng Xi, Jia Wei, Jun Zhu, Jianfei Chen.
<b>[ICML 2025]</b> Full Paper, CCF-A	<b>SageAttention2:</b> 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.
<b>[ICLR 2025]</b> Full Paper, CCF-A	<b>SageAttention:</b> Accurate 8-Bit Attention for Plug-and-play Inference Acceleration. <a href="#">[Code (3K Stars)]</a> <b>Jintao Zhang</b> , Jia Wei, Pengle Zhang, Jun Zhu, Jianfei Chen.
<b>[ICDE 2025]</b> Full Paper, CCF-A	<b>SAGE:</b> A Framework of Precise Retrieval for RAG. <b>Jintao Zhang</b> , Guoliang Li, Jinyang Su.
<b>[SIGMOD 2024]</b> Full Paper, CCF-A	<b>PACE:</b> Poisoning Attacks on Learned Cardinality Estimation. <b>Jintao Zhang</b> , Guoliang Li, Chao Zhang, Chengliang Chai.
<b>[ICDE 2023]</b> Full Paper, CCF-A	<b>AutoCE:</b> An Accurate and Efficient Model Advisor for Learned Cardinality Estimation. <b>Jintao Zhang</b> , Chao Zhang, Guoliang Li, Chengliang Chai.
<b>[TKDE 2025]</b> Full Paper, CCF-A	<b>A Lightweight Learned Cardinality Estimation Model.</b> Yaoyu Zhu, <b>Jintao Zhang</b> <sup>#</sup> ( <i>corresponding author</i> ), Guoliang Li <sup>#</sup> , Jianhua Feng.
<b>[VLDB 2022]</b> Full Paper, CCF-A	<b>Learned Cardinality Estimation:</b> Design Space Exploration and Comparative Evaluation. Ji Sun*, <b>Jintao Zhang</b> * ( <i>equal first contribution</i> ), Zhaoyan Sun, Guoliang Li, Nan Tang.
<b>[PrePrint]</b> Full Paper	<b>A Survey of Efficient Attention Methods:</b> Hardware-efficient, Sparse, Compact, and Linear Attention <b>Jintao Zhang</b> , Rundong Su, Chunyu Liu, Jia Wei, Ziteng Wang, Pengle Zhang, et al.

## RESEARCH INTERESTS

- 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

<b>Bytedance ScholarShip</b> (Less than 20 globally/year)	2025
<b>Siebel Scholars Scholarship</b> (Less than 100 globally/year)	2023
<b>China National Scholarship</b>	2019 & 2020 & 2025