

JINTAO ZHANG

✉ zhang-jt24@mails.tsinghua.edu.cn | 📞 +86 185-1911-5711 (WeChat ID) | 🌐 <https://jt-zhang.github.io>

EDUCATION

Tsinghua University Ph.D. of Science in Computer Science and Technology Supervisor: Prof. Jun Zhu 🔗 Homepage , and Prof. Jianfei Chen 🔗 Homepage .	Aug. 2024 – Dec.2026 (Expected) Beijing, China
Tsinghua University Master of Science in Computer Science and Technology Supervisor: Prof. Guoliang Li 🔗 Homepage .	Aug. 2021 – Jun. 2024 Beijing, China
Xidian University Bachelor of Science in Computer Science and Technology	Sep. 2017 – Jun. 2021 Shaanxi, China

SELECTED PUBLICATIONS

[Arxiv 2025] Technical Report	TurboDiffusion: Accelerating Video Diffusion Models by 100-200 Times. [Code (3.1K Stars)] Jintao Zhang , Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, Jun Zhu.
[Arxiv 2025] Full Paper	SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention. Jintao Zhang , Haoxu Wang, Kai Jiang, ···, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, Jun Zhu. [Code (0.3K Stars)]
[NeurIPS 2025] Spotlight, CCF-A	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, Jianfei Chen, Jun Zhu.
[ICML 2025] Full Paper, CCF-A	SpargeAttention: Accurate Sparse Attention Accelerating Any Model Inference. [Code (0.9K 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 (3K 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 [#] (<i>corresponding author</i>), Guoliang Li [#] , Jianhua Feng.
[VLDB 2022] Full Paper, CCF-A	Learned Cardinality Estimation: Design Space Exploration and Comparative Evaluation. Ji Sun*, Jintao Zhang * (<i>equal first contribution</i>), Zhaoyan Sun, Guoliang Li, Nan Tang.
[PrePrint] Full Paper	A Survey of Efficient Attention Methods: Hardware-efficient, Sparse, Compact, and Linear Attention Jintao Zhang , 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

Tencent Project Up Scholarship (Less than 15 globally/year)	2026
Bytedance ScholarShip (Less than 20 globally/year)	2025
Siebel Scholars Scholarship (Less than 100 globally/year)	2023
China National Scholarship	2019 & 2020 & 2025