# JINTAO ZHANG

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** — **Beijing, China**
**Supervisor:** Prof. Guoliang Li Homepage.

**Xidian University** — **Sep. 2017 – Jun. 2021**
**Bachelor of Science in Computer Science and Technology** — **Shaanxi, China**

## SELECTED PUBLICATIONS

**[Arxiv 2025]** TurboDiffusion: Accelerating Video Diffusion Models by 100-200 Times. [Code (1K Stars)]
Technical Report — **Jintao Zhang**, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E. Gonzalez, Jianfei Chen, Jun Zhu

**[Arxiv 2025]** SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention.
Full Paper — **Jintao Zhang**, Haoxu Wang, Kai Jiang, ⋯, Ion Stoica, Joseph E. Gonzalez, Jun Zhu, Jianfei Chen

**[NeurIPS 2025]** SageAttention3: Microscaling FP4 Attention for Inference and An Exploration of 8-Bit Training.
**Spotlight**, CCF-A — **Jintao Zhang**, Jia Wei, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Haoxu Wang, Kai Jiang, Jun Zhu, Jianfei Chen

**[ICML 2025]** SpargeAttention: Accurate Sparse Attention Accelerating Any Model Inference. [Code (0.9K Stars)]
Full Paper, CCF-A — **Jintao Zhang**, Chendong Xiang, Haofeng Huang, Haocheng Xi, Jia Wei, Jun Zhu, Jianfei Chen.

**[ICML 2025]** SageAttention2: Efficient Attention with Thorough Outlier Smoothing and Per-thread INT4 Quantization
Full Paper, CCF-A — **Jintao Zhang**, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, Jianfei Chen.

**[ICLR 2025]** SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration.[Code (3K Stars)]
Full Paper, CCF-A — **Jintao Zhang**, Jia Wei, Pengle Zhang, Jun Zhu, Jianfei Chen.

**[ICDE 2025]** SAGE: A Framework of Precise Retrieval for RAG.
Full Paper, CCF-A — **Jintao Zhang**, Guoliang Li, Jinyang Su.

**[SIGMOD 2024]** PACE: Poisoning Attacks on Learned Cardinality Estimation.
Full Paper, CCF-A — **Jintao Zhang**, Guoliang Li, Chao Zhang, Chengliang Chai.

**[ICDE 2023]** AutoCE: An Accurate and Efficient Model Advisor for Learned Cardinality Estimation.
Full Paper, CCF-A — **Jintao Zhang**, Chao Zhang, Guoliang Li, Chengliang Chai.

**[TKDE 2025]** A Lightweight Learned Cardinality Estimation Model.
Full Paper, CCF-A — Yaoyu Zhu, **Jintao Zhang#** (*corresponding author*), Guoliang Li#, Jianhua Feng.

**[VLDB 2022]** Learned Cardinality Estimation: Design Space Exploration and Comparative Evaluation.
Full Paper, CCF-A — Ji Sun*, **Jintao Zhang*** (*equal first contribution*), Zhaoyan Sun, Guoliang Li, Nan Tang.

[PrePrint] A Survey of Efficient Attention Methods: Hardware-efficient, Sparse, Compact, and Linear Attention
Full Paper — **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

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