TAN XIN PH.D. STUDENT

Email: re.tanxin13@gmail.com
Web: https://txxx926.github.io/

EDUCATION

The Chinese University of Hong Kong — Hong Kong SAR, China
Ph.D. in Computer Science and Engineering — 2022 - 2026 *(expected)*
- Advisor: Prof. Hong Xu
- Research area: Machine Learning System

Northwestern Polytechnical University — Xi'an, China
B.E. in Computer Science and Technology — 2018 - 2022
- GPA: 93.37/100, Rank: 1/247.

RESEARCH INTEREST

I am broadly interested in System Design for Machine Learning (MLSys), including the following topics:

1. **Distributed Training:** Developing and optimizing strategies for efficient, scalable training of large models..
2. **Efficient Serving Systems:** Designing new architectures and algorithms for high-performance serving of large models and related applications, e.g. LLMs, diffusion models.

PUBLICATIONS

1. **Xin Tan**, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, Hong Xu, **DSV: Exploiting Dynamic Sparsity to Accelerate Large-Scale Video DiT Training**. *ArXiv Preprint*, 2025.
2. **Xin Tan**, Yimin Jiang, Yitao Yang, Hong Xu, **Towards End-to-End Optimization of LLM-based Applications with Ayo**. *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, 2025.
3. **Xin Tan**, Jiamin Li, Yitao Yang, Jingzong Li, Hong Xu, **Arlo: Serving Transformer-based Language Models with Dynamic Input Lengths**. *ACM International Conference on Parallel Processing (**ICPP**)*, 2024.

INTERNSHIPS

System Group, StepFun | Beijing, China — 2024.08 - present
- Pinpointed the attention bottleneck with lengthy video input and analyzed the sparse attention patterns in video DiTs.
- Designed two-stage training algorithm via dynamic sparse attention with kernel optimizations to mitigate attention bottleneck while preserving performance.
- Developed a hybrid and sparsity-aware context parallelism scheme for more efficient scaling.

Network Research Group, Microsoft Research Asia | Remote — 2021.10 - 2022.03
- Developed a real-time, non-intrusive system for collecting key AI infrastructure metrics (e.g., GPU utilization, network and NVLink bandwidth).
- Designed and built a data analytics platform to analyze machine learning workloads across large-scale clusters.
- Analyzed six months of datacenter workload data to characterize resource usage and network patterns of various AI tasks, providing actionable recommendations to optimize cloud infrastructure and software stack.

Awards and Honors

- **Student Travel Grant**, ASPLOS 2025 — 2025.4
- **Full Postgraduate Scholarship**, The Chinese University of Hong Kong — 2022-2026
- **Outstanding Graduate**, Northwestern Polytechnical University — 2022
- **National Scholarship**, Ministry of Education (China) — 2020
- **National Scholarship**, Ministry of Education (China) — 2019
- **Champion**, International Underwater Robot Competition — 2020

Skills

Languages: Chinese, English

Programming: `Python`, `Pytorch`, `Megatron`, `Ray`, `Triton`, `CUDA`, `C++`

Academic Services

Reviewers: *IEEE Transactions on Network Science and Engineering*,

Artifact Evaluation Committee: *USENIX OSDI/ATC 2025, ACM CoNeXT 2025, ACM EuroSys 2025 Spring/Fall, USENIX OSDI/ATC 2024*,

Teaching

Teaching Assistant: *CSCI 3150, Introduction to Operating Systems, CUHK. 2023 Spring*, *CSCI 1120, Introduction to Computing Using C++, CUHK. 2022 Fall.*