

EDUCATION	<b>The Chinese University of Hong Kong</b> <i>Ph.D. in Computer Science and Engineering</i> <ul style="list-style-type: none"> <li>• Advisor: Prof. Hong Xu</li> <li>• Research area: Machine Learning System</li> </ul>	Hong Kong SAR, China 2022 - 2026 ( <i>expected</i> )
	<b>Northwestern Polytechnical University</b> <i>B.E. in Computer Science and Technology</i> <ul style="list-style-type: none"> <li>• GPA: 93.37/100, Rank: 1/247.</li> </ul>	Xi'an, China 2018 - 2022
RESEARCH INTEREST	<p>I am broadly interested in System Design for Machine Learning (MLSys), including the following topics:</p> <ol style="list-style-type: none"> <li>1. <b>Distributed Training:</b> Developing and optimizing strategies for efficient, scalable training of large models..</li> <li>2. <b>Efficient Serving Systems:</b> Designing new architectures and algorithms for high-performance serving of large models and related applications, e.g. LLMs, diffusion models.</li> </ol>	
PUBLICATIONS	<ol style="list-style-type: none"> <li>1. <b>Xin Tan</b>, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, Hong Xu, <b>DSV: Exploiting Dynamic Sparsity to Accelerate Large-Scale Video DiT Training</b>. <i>ArXiv Preprint</i>, 2025.</li> <li>2. <b>Xin Tan</b>, Yimin Jiang, Yitao Yang, Hong Xu, <b>Towards End-to-End Optimization of LLM-based Applications with Ayo</b>. <i>ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)</i>, 2025.</li> <li>3. <b>Xin Tan</b>, Jiamin Li, Yitao Yang, Jingzong Li, Hong Xu, <b>Arlo: Serving Transformer-based Language Models with Dynamic Input Lengths</b>. <i>ACM International Conference on Parallel Processing (ICPP)</i>, 2024.</li> </ol>	
INTERNSHIPS	<b>System Group, StepFun</b>   Beijing, China <ul style="list-style-type: none"> <li>• Pinpointed the attention bottleneck with lengthy video input and analyzed the sparse attention patterns in video DiTs.</li> <li>• Designed two-stage training algorithm via dynamic sparse attention with kernel optimizations to mitigate attention bottleneck while preserving performance.</li> <li>• Developed a hybrid and sparsity-aware context parallelism scheme for more efficient scaling.</li> </ul>	2024.08 - present
	<b>Network Research Group, Microsoft Research Asia</b>   Remote <ul style="list-style-type: none"> <li>• Developed a real-time, non-intrusive system for collecting key AI infrastructure metrics (e.g., GPU utilization, network and NVLink bandwidth).</li> <li>• Designed and built a data analytics platform to analyze machine learning workloads across large-scale clusters.</li> <li>• 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.</li> </ul>	2021.10 - 2022.03

AWARDS AND HONORS	• <b>Student Travel Grant</b> , ASPLOS 2025	2025.4
	• <b>Full Postgraduate Scholarship</b> , The Chinese University of Hong Kong	2022-2026
	• <b>Outstanding Graduate</b> , Northwestern Polytechnical University	2022
	• <b>National Scholarship</b> , Ministry of Education (China)	2020
	• <b>National Scholarship</b> , Ministry of Education (China)	2019
	• <b>Champion</b> , International Underwater Robot Competition	2020
SKILLS	<b>Languages:</b> Chinese, English	
	<b>Programming:</b> Python, Pytorch, Megatron, Ray, Triton, CUDA, C++	
ACADEMIC SERVICES	<b>Reviewers:</b> <i>IEEE Transactions on Network Science and Engineering</i> ,	
	<b>Artifact Evaluation Committee:</b> <i>USENIX OSDI/ATC 2025, ACM CoNext 2025, ACM EuroSys 2025 Spring/Fall, USENIX OSDI/ATC 2024,</i>	
TEACHING	<b>Teaching Assistant:</b> <i>CSCI 3150, Introduction to Operating Systems, CUHK. 2023 Spring, CSCI 1120, Introduction to Computing Using C++, CUHK. 2022 Fall.</i>	