

XUANLEI ZHAO

xuanlei@comp.nus.edu.sg ◇ +65 81485063 ◇ Singapore

[Homepage](#) ◇ [GitHub](#) ◇ [Google Scholar](#) ◇ [Twitter](#)

EDUCATION

National University of Singapore Ph.D. in Computer Science Supervisor: Yang You	01.2024 - Present
National University of Singapore M.S. in Computer Science	08.2022 - 12.2023
Huazhong University of Science and Technology B.Eng. in Computer Science & Electronic Information	09.2018 - 06.2022

RESEARCH INTEREST

- **Machine Learning System:** Parallelism, Scheduling, Offloading, Compiler.
- **Efficient Video Generation:** Efficient Training and Inference, Algorithm-System Co-Design.

RESEARCH EXPERIENCE

PAB: The **First Real-Time** and **Most Cited** cache-based video generation acceleration method.

- [ICLR 2025] Real-Time Video Generation with Pyramid Attention Broadcast
- [Xuanlei Zhao](#)^{*}, [Xiaolong Jin](#)^{*}, [Kai Wang](#)^{*†}, [Yang You](#)[†]

VideoSys: The **First** and **Most Starred** open-source project for system speedup of video training and inference.

- VideoSys: An Easy and Efficient System for Video Generation
- Project lead.

DCP: The **First Practical** parallel method for efficient variable sequences training (*e.g.*, videos).

- Training Variable Sequences with Data-Centric Parallel
- [Geng Zhang](#)^{*}, [Xuanlei Zhao](#)^{*}, [Kai Wang](#)[†], [Yang You](#)[†]

DSP: The **Most Efficient** sequence parallel for multi-dim transformers (*e.g.*, spatial-temporal video models).

- [ICML 2025] DSP: Dynamic Sequence Parallelism for Multi-Dimensional Transformers
- [Xuanlei Zhao](#), [Shenggan Cheng](#), [Chang Chen](#), [Zangwei Zheng](#), [Ziming Liu](#), [Zheming Yang](#), [Yang You](#)

HeteGen: Accelerate LLM offloading inference by heterogeneous computing between CPU and GPU.

- [MLSys 2024] HeteGen: Heterogeneous Parallel Inference for Large Language Models on Resource-Constrained Devices
- [Xuanlei Zhao](#)^{*}, [Bin Jia](#)^{*}, [Haotian Zhou](#)^{*}, [Ziming Liu](#), [Shenggan Cheng](#), [Yang You](#)

AutoChunk: A compiler to reduce activation memory by over 80% for long sequences (*e.g.*, videos).

- [ICLR 2024] AutoChunk: Automated Activation Chunk for Memory-Efficient Long Sequence Inference
- [Xuanlei Zhao](#), [Shenggan Cheng](#), [Guangyang Lu](#), [Jiarui Fang](#), [Haotian Zhou](#), [Bin Jia](#), [Ziming Liu](#), [Yang You](#)

FastFold: The **First** and **Most Cited** system optimization method for AlphaFold by parallel and computing.

- [PPoPP 2024] FastFold: Optimizing AlphaFold Training and Inference on GPU Clusters
- [Shenggan Cheng](#), [Xuanlei Zhao](#), [Guangyang Lu](#), [Jiarui Fang](#), [Tian Zheng](#), [Ruidong Wu](#), [Xiwen Zhang](#), [Jian Peng](#), [Yang You](#)

INDUSTRY EXPERIENCE

Pika, Inc.

05.2024 - 08.2024

Research Intern | Supervised by Chenlin Meng

Palo Alto, CA

- Optimize distributed system on thousands of GPUs for efficient large-scale training of video models.
- Improve training performance with hybrid parallel, I/O optimization, and dynamic activation checkpointing.
- Improve generation efficiency with sequence parallel, adaptive computing, efficient kernel and distillation.

HPC-AI TECH, Inc. (Colossal-AI)

05.2022 - 12.2023

Research Intern | Supervised by Jiarui Fang

Singapore

- Contribute 48k lines of code as a core contributor ([rank 5th by 2023](#)) and help it gain 35k stars on Github.
- Propose AutoChunk, a compiler to reduce the activation memory by 80% for long sequences inference.
- Participate in the development of various parallelism strategies including sequence parallel, tensor parallel, ZeRO, offloading, auto parallelism and efficient kernels.