SHENG LI

✉ shl188@pitt.edu 🏠 Homepage in LinkedIn 🎓 Google Scholar

Education

University of Pittsburgh, Pittsburgh, PA, USA 08/2021 – Present
Ph.D. in Computer Science
(Field: Efficient Machine Learning, GenAI, Diffusion, Video Generation, LLM, Self-Supervised Learning)
Advisor: Dr. Xulong Tang

Sichuan University, Chengdu, Sichuan, China 09/2016 – 06/2020
B.Eng. in Software Engineering, Graduated with Honor
Advisor: Dr. Ziliang Feng

Experience

Adobe Research 05/2025 – 11/2025
Research Scientist/Engineer Intern *Mentor: Dr. Yifan Gong*

- **Dynamic Patchification for Video Diffusion:** Developed a content-aware patchification router that adaptively adjusts patch size (i.e., process granularity) across different spatiotemporal regions of video, reducing redundant tokens and achieving **1.8× speedup** for video generation without affecting visual quality.
- Submitted a first-author paper for this project and is accepted to **CVPR 2026**!

University of Pittsburgh 2020 – Present
Research Assistant / Teaching Assistant *Advisor: Dr. Xulong Tang*

- **Video Diffusion Pruning:** Proposed a training-free temporal-aware token pruning method for video generation that preserves temporal coherence during pruning and achieves **1.5× speedup** without quality loss.
- **Efficient Self-Supervised Learning (SSL):** Developed a similarity-based pruning method to discard less important input image regions in SSL, cutting training costs by over **40%**.
- **Efficient LLM Serving:** Accelerated RAG-enabled large language model serving by deduplicating and reordering retrieved RAG chunks and improving KV-cache reuse, reducing redundant chunk transfers by ∼**25%** and end-to-end latency by ∼**18%** (Llama-3.1-8B on vLLM).
- **Layer Freezing for Training:** Proposed an attention-based layer freezing approach that reduces computation by **40%** and memory cost by **50%** for ML model training.
- **3DGS Acceleration:** Leveraged tensor cores to accelerate the 3D Gaussian Splatting rendering.

Selected Publications *(*equal contribution)*

[1] **Sheng Li**, Connelly Barnes, Mamshad Nayeem Rizve, Hongwu Peng, Zhengang Li, Ohi Dibua, Alireza Ganjdanesh, Xulong Tang, Yan Kang, and Yifan Gong. “Content-Aware Dynamic Patchification for Efficient Video Diffusion.” *In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition 2026.* **(CVPR 2026). This is the intern project in Adobe!**

[2] **Sheng Li***, Qitao Tan*, Yue Dai, Zhenglun Kong, Tianyu Wang, Jun Liu, Ao Li, Ninghao Liu, Yufei Ding, Xulong Tang, and Geng Yuan. “Mutual Effort for Efficiency: A Similarity-based Token Pruning for Vision Transformers in Self-Supervised Learning.” *In Proceedings of 13th International Conference on Learning Representations.* **(ICLR 2025)**

[3] **Sheng Li**, Chao Wu, Ao Li, Yanzhi Wang, Xulong Tang, and Geng Yuan. “Waxing-and-Waning: a Generic Similarity-based Framework for Efficient Self-Supervised Learning.” *In Proceedings of 12th International Conference on Learning Representations.* **(ICLR 2024)**

[4] **Sheng Li***, Geng Yuan*, Yue Dai*, Youtao Zhang, Yanzhi Wang, and Xulong Tang. “SmartFRZ: An Efficient Training Framework using Attention-Based Layer Freezing.” *In Proceedings of 11th International Conference on Learning Representations.* **(ICLR 2023, Honored with Spotlight Award)**

[5] **Sheng Li**, Yang Sui, Junhao Ran, Bo Yuan, Yue Dai, and Xulong Tang. “TAPE: Temporal Aware Pruning for Efficient Diffusion-based Video Generation.” **(Submitted to ECCV 2026)**

[6] Huidong Ji, **Sheng Li**, Yue Cao, Chen Ding, Jiawei Xu, Qitao Tan, Ao Li, Jun Liu, Xulong Tang, Lirong Zheng, Geng Yuan, and Zhuo Zou. “A Computation and Energy Efficient Hardware Architecture for SSL Acceleration.” *In Proceedings of 30th Asia and South Pacific Design Automation Conference.* **(ASP-DAC 2025)**

[7] Geng Yuan*, Yanyu Li*, **Sheng Li**, Zhenglun Kong, Sergey Tulyakov, Xulong Tang, Yanzhi Wang, and Jian Ren. “Layer Freezing & Data Sieving: Missing Pieces of a Generic Framework for Sparse Training”. *In Proceedings of the 36th Conference on Neural Information Processing Systems.* **(NeurIPS 2022)**

[8] **Sheng Li**, Geng Yuan, Yawen Wu, Yue Dai, Tianyu Wang, Chao Wu, Alex K. Jones, Jingtong Hu, Yanzhi Wang, and Xulong Tang. “ETuner: A Redundancy-Aware Framework for Efficient Continual Learning Application on Edge Devices.” **(Preprint)**

[9] Tianyu Wang*, **Sheng Li***, Bingyao Li, Yue Dai, Ao Li, Geng Yuan, Yufei Ding, Youtao Zhang, and Xulong Tang. “Improving GPU Multi-Tenancy Through Dynamic Multi-Instance GPU Reconfiguration.” **(Preprint)**

Selected Honors and Awards

Spotlight Award, The 11th ICLR Conference	2023
National Scholarship, Ministry of Education of China	2018, 2019
Comprehensive Scholarship, Sichuan University	2017

Skills

Python, PyTorch; Generative AI, Diffusion, Video Generation; LLM Serving (RAG, KV-cache reuse); Computer Vision, Self-Supervised Learning; Efficient ML Methods (Token Pruning, Layer Freezing); CNN, Transformer; GPU Programming (CUDA), GPU Architecture, Nsight GPU Profiling, ML Kernel; 3DGS

Teaching

Teaching Assistant: Computer Organization and Assembly Language; Introduction to Systems Software; Algorithms and Data Structures

Professional Service

Reviewer of NeurIPS, ICLR, ICML, CVPR, ECCV, AAAI, TPAMI, TICPS, ACML, and NPL