Muyang Li

muyangli@mit.edu · lmxyy · Muyang Li · lmxyy.me

Education

Massachusetts Institute of Technology Sep. 2023 – Jan. 2026
Ph.D. at EECS, advised by Prof. Song Han Cambridge, MA

Carnegie Mellon University Aug. 2021 – May 2023
Master of Science in Robotics, advised by Prof. Jun-Yan Zhu Pittsburgh, PA

Shanghai Jiao Tong University Sep. 2016 – Jun. 2020
Bachelor of Engineering in Computer Science Shanghai, China

- Member of ACM Class, an elite CS program for the top 5% talented students.

Research Interests

My research interest is in the intersection of machine learning, system, and computer graphics.
I am currently working on building efficient and hardware-friendly generative models with its applications in computer vision and graphics.

Publications Google Scholar

[1] Xingyang Li*, Muyang Li*, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, and Song Han, *Radial Attention: $O(n \log n)$ Sparse Attention with Energy Decay for Long Video Generation* (**NeurIPS 2025**)

[2] Shuo Yang*, Haocheng Xi*, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, Jianfei Chen, Song Han, Kurt Keutzer, and Ion Stoica, *Sparse VideoGen2: Accelerate Video Generation with Sparse Attention via Semantic-Aware Permutation* (**NeurIPS 2025 Spotlight**)

[3] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala, *Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models* (**NeurIPS 2025 Spotlight**)

[4] Haocheng Xi*, Shuo Yang*, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, and Song Han, *Sparse VideoGen: Accelerating Video Diffusion Transformers with Spatial-Temporal Sparsity* (**ICML 2025**)

[5] Enze Xie*, Junsong Chen*, Yuyang Zhao†, Jincheng Yu†, Ligeng Zhu†, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Cai Han, Bingchen Liu, Daquan Zhou, and Song Han, *SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer* (**ICML 2025**)

[6] Muyang Li*, Yujun Lin*, Zhekai Zhang*, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han, *SVDQuant: Absorbing Outliers by Low-Rank Components for 4-Bit Diffusion Models* (**ICLR 2025 Spotlight**)

[7] Enze Xie*, Junsong Chen*, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, Song Han, *SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers* (**ICLR 2025 Oral**)

[8] Junyu Chen*, Han Cai*, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han, *Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models* (**ICLR 2025**)

[9] Muyang Li*, Tianle Cai*, Jiaxin Cao, Qinsheng Zhang, and Han Cai, Junjie Bai, Yangqing Jia, Ming-Yu Liu, Kai Li, and Song Han, *DistriFusion: Distributed Parallel Inference for High-Resolution Diffusion Models* (**CVPR 2024 Highlight**)

[10] Han Cai, Muyang Li, Zhuoyang Zhang, Qinsheng Zhang, Ming-Yu Liu, and Song Han, *Condition-Aware Neural Network for Controlled Image Generation* (**CVPR 2024**)

[11] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han and Jun-Yan Zhu, *Efficient Spatially Sparse Inference for Conditional GANs and Diffusion Models* (**NeurIPS 2022 & T-PAMI 2023**)

[12] Yihan Wang, Muyang Li, Han Cai, Wei-Ming Chen and Song Han, *Lite Pose: Efficient Architecture Design for 2D Human Pose Estimation* (**CVPR 2022**)

[13] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han, *GAN Compression: Efficient Architectures for Interactive Conditional GANs* (**CVPR 2020 & T-PAMI 2021**)

[14] Jindong Jiang*, Xiuyu Li*, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, Sungjin Ahn, Jan Kautz, Hongxu Yin, Yao Lu, Song Han, and Wonmin Byeon, *Token-Efficient Long Video Understanding for Multimodal LLMs*

Experiences

NVIDIA May 2024 – Sep. 2025
Research Intern Work with Prof. Song Han Santa Clara, CA

NVIDIA Jun. 2023 – Aug. 2023
Research Intern Work with Prof. Song Han and Ming-Yu Liu Shanghai, China

OmniML Inc. May 2022 – Aug. 2022
Summer Intern Work with Prof. Song Han San Jose, CA

Dawnlight Inc. Jul. 2020 – Jul. 2021
Data Scientist Work with Prof. Song Han and Prof. Jia Li Shanghai, China

MIT HAN Lab Jul. 2019 – Jan. 2020
Research Assistant Advisor: Prof. Song Han and Prof. Jun-Yan Zhu Cambridge, MA

Open-sourced Projects

nunchaku-tech/nunchaku (3.6K Stars) May 2024 – Present
Python/CUDA 4-bit diffusion model inference engine.

nunchaku-tech/ComfyUI-nunchaku (2.7K Stars) Mar. 2025 – Present
Python/CUDA ComfyUI plugin for Nunchaku.

nunchaku-tech/deepcompressor(700+ Stars) May 2024 – Present
Python Quantization library for LLMs and diffusion models.

mit-han-lab/radial-attention (500+ Stars) Mar. 2025 – Present
Python $\mathcal{O}(n \log n)$ sparse attention for video generation.

mit-han-lab/gan-compression (1.1K Stars) Jul. 2019 – Apr. 2020
Python A general conditional GAN Compression framework.

mit-han-lab/distrifuser (700+ Stars) Oct. 2023 – Feb. 2024
Python A distributed framework to accelerate diffusion models with multiple GPUs.

lmxyy/sige Jul. 2021 – Nov. 2022
Python/C++/CUDA/Metal A sparse engine to accelerate image editing for GANs and diffusion models.

mit-han-lab/litepose Mar. 2021 – Jun. 2022
Python A light-weighted pose estimation model that could run on mobile devices.

Honors and Awards

Best Pitch Award, Award on MARC 2025 — Jan. 2025
Seneff-Zue CS Fellowship ($98K) — Sep. 2023
Gold Medal, Award on CCPC2017 Harbin Regional, Ranked 10th — Oct. 2017
Gold Medal, Award on ICPC2017 Qingdao Regional, Ranked 5th — Nov. 2017
3rd Runner-up, Award on ICPC2017 Jakarta Regional — Nov. 2017
1st Runner-up, Award on Singing Competition of Zhiyuan College in SJTU — Dec. 2017
Jin Long Yu Fellowship, Award for top 1% students — Dec. 2017
1st Runner up's Coach, Award on ICPC 2018 Pathom Regional — Nov. 2018
A-Class School-level Scholarship, Award for top 1% students — Dec. 2018
Zhiyuan Honorary Scholarship (3 times), Award for top 5% students — 2016, 2017, 2018
Honorable Mention, Award for 2019 American Interdisciplinary Contest in Modeling (ICM) — Jan. 2019

Academic Services

- Conference Reviewer: ICML, ICLR, NeurIPS (Top Reviewer in 2025), ICCV, CVPR, SIGGRAPH Asia
- Journal Reviewer: T-PAMI, IJCV, TVCJ, TCSVT

Teaching

SJTU ACM-ICPC Coach — Jun. 2018 – Apr. 2019
TA at SJTU Data Structure (CS147) — Mar. 2018 – May 2018

Skills

Programming Languages: C++/C/CUDA = Python > Java
Deep Learning Packages: PyTorch, TensorFlow, TVM, TensorRT
Languages: English - Proficient, Mandarin - Native speaker, Japanese - Amateur
Other: Pop Singing