

SHANG YANG

✉ shangy@mit.edu · ⚡ ys-2020 · Ⓜ Shang Yang · ⌐ ys-2020.github.io

🎓 EDUCATION

Massachusetts Institute of Technology (MIT)

Ph.D. Student in EECS, advised by Prof. Song Han.

Sep. 2023 - Present

Cambridge, MA

Tsinghua University

Bachelor of Engineering in Electronic Information Science and Technology

Aug. 2019 - Jul. 2023

Beijing, China

- Overall GPA: 3.99 / 4.0 Rank: 1 / 256

📄 SELECTED PUBLICATIONS [GOOGLE SCHOLAR](#)

[1] Ji Lin*, Jiaming Tang*, Haotian Tang[†], **Shang Yang**[†], Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, Song Han. **AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration.** (*Algorithm co-lead, [†]System co-lead. The first four authors have equal contributions.) (**MLSys 2024 Best Paper Award**)

[2] Yujun Lin*, Haotian Tang*, **Shang Yang**^{*}, Zhekai Zhang, Guangxuan Xiao, Chuang Gan, Song Han. **QServe: W4A8KV4 Quantization and System Co-design for Efficient LLM Serving.** (* indicates equal contribution) (**MLSys 2025**)

[3] **Shang Yang**^{*}, Junxian Guo*, Haotian Tang, Qinghao Hu, Guangxuan Xiao, Jiaming Tang, Yujun Lin, Zhijian Liu, Yao Lu, Song Han. **LServe: Efficient Long-sequence LLM Serving with Unified Sparse Attention.** (**MLSys 2025**)

[4] Qinghao Hu*, **Shang Yang**^{*}, Junxian Guo, Xiaozhe Yao, Yujun Lin, Yuxian Gu, Han Cai, Chuang Gan, Ana Klimovic, Song Han. **Taming the Long-Tail: Efficient Reasoning RL Training with Adaptive Drafter.** (**ASPLoS 2026**)

[5] Haotian Tang*, **Shang Yang**^{*}, Zhijian Liu, Ke Hong, Zhongming Yu, Xiuyu Li, Guohao Dai, Yu Wang, Song Han. **TorchSparse++: Efficient Training and Inference Framework for Sparse Convolution on GPUs.** (**MICRO 2023**)

[6] Yuxian Gu, Qinghao Hu, **Shang Yang**, Haocheng Xi, Junyu Chen, Song Han, Han Cai. **Jet-Nemotron: Efficient Language Model with Post Neural Architecture Search.** (**NeurIPS 2025**)

[7] Haotian Tang*, Yecheng Wu*, **Shang Yang**, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, Song Han. **HART: Efficient Visual Generation with Hybrid Autoregressive Transformer.** (**ICLR 2025**)

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

[9] Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, **Shang Yang**, Haotian Tang, Yao Fu, Song Han. **DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads.** (**ICLR 2025**)

[10] Zhijian Liu*, Ligeng Zhu*, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, **Shang Yang**, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, Yao Lu. **NVILA: Efficient Frontier Visual Language Models.** (**CVPR 2025**)

[11] Yukang Chen*, Fuzhao Xue*, Dacheng Li[†], Qinghao Hu[†], Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, **Shang Yang**, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, Song Han. **LongVILA: Scaling Long-Context Visual Language Models for Long Videos.** (**ICLR 2025**) (*Algorithm co-lead, [†]System co-lead. The first four authors have equal contributions.)

EXPERIENCES

NVIDIA	<i>Research Intern</i>	Work with Prof. Song Han	Jun. 2025 - Aug. 2025
Topic:	Efficient Training for Reasoning Large Language Models	Santa Clara, CA	
NVIDIA	<i>Research Intern</i>	Work with Prof. Song Han	Jun. 2024 - Jan. 2025
Topic:	Efficient Systems for Large Language Models and Multimodal Models	Cambridge, MA	
MIT	<i>Research Intern</i>	Advised by Prof. Song Han	Jul. 2022 - Aug. 2023
Topic:	Efficient Machine Learning Systems for 3D Point Clouds	Cambridge, MA	

PROJECTS

NVlabs/VILA (**3.6K Stars**)

A Family of State-of-the-Art Vision Language Models (VLMs) for Diverse Multimodal AI Tasks.

mit-han-lab/llm-awq (**3.3K Stars**)

Effective Low-bit Weight Quantization Algorithm for LLMs with Efficient System Support.

mit-han-lab/torchsparse (**1.4K Stars**)

High-performance Neural Network Library for Point Cloud Processing.

mit-han-lab/omniserve (**> 700 Stars**)

Efficient and Accurate LLM Serving System on GPUs with W4A8KV4 Quantization and Unified Sparse Attention.

TEACHING

Teaching Assistant for TinyML Course ([MIT 6.5940](#))

Sep. 2024 - Dec. 2024

SKILLS

- **Programming Languages:** Python, CUDA, C++
- **Languages:** English - Proficient, Mandarin - Native