Xinyu **Yang**

Carnegie Mellon University

xinyuya2@andrew.cmu.edu | xinyuyang.me/ | Hanyuezhuohua | Xinyu Yang

Education

Carnegie Mellon University — *Pittsburgh, US*

Ph.D. student in Electrical and Computer Engineering — *Aug. 2023 - Present*

Shanghai Jiao Tong University — *Shanghai, China*

Bachelor of Engineering in Computer Science, ACM Honors Class — *Sept. 2019 - Jun. 2023*

- **ACM Honors Class** is an elite CS program for students ranked in the top 5% with aspirations in research.
- Overall GPA: **91.26 / 100**, Ranking: **1 / 29**

Research Interests

Machine Learning and System, Scalable and Generalizable Foundation Models for In-the-wild Applications

Publications

** indicates equal contributions*

APE: Faster and Longer Context-Augmented Generation via Adaptive Parallel Encoding — *Under Review*

Xinyu Yang, *Tianqi Chen, Beidi Chen*

VcLLM: Video Codecs are Secretly Tensor Codecs — *Preprint*

Ceyu Xu, Yongji Wu*, **Xinyu Yang***, Beidi Chen, Matthew Lentz, Danyang Zhuo, Lisa Wu Wills*

S^2FT: Efficient, Scalable and Generalizable LLM Fine-tuning by Structured Sparsity — *NeurIPS 2024*

Xinyu Yang, *Jixuan Leng, Geyang Guo, Jiawei Zhao, Ryumei Nakada, Linjun Zhang, Huaxiu Yao, Beidi Chen*

It Takes Two: On the Seamlessness between Reward and Policy Model in RLHF — *Under Review*

Taiming Lu, Lingfeng Shen*, **Xinyu Yang***, Weiting Tan, Beidi Chen, Huaxiu Yao*

Blindness and Hallucinations: Revisiting Multi-modal Alignment in Vision-Language Large Models — *Under Review*

Xinyu Yang, *Chenhang Cui, Jaehong Yoon, Yiyang Zhou, Yi-Lin Sung, Mohit Bansal, Beidi Chen, Huaxiu Yao*

Zeroth-Order Fine-Tuning of LLMs with Extreme Sparsity — *Preprint*

*Wentao Guo, Jikai Long, Yimeng Zeng, Zirui Liu, **Xinyu Yang**, Yide Ran, et al.*

TriForce: Lossless Acceleration of Long Sequence Generation with Hierarchical Speculative Decoding — *COLM 2024*

*Hanshi Sun, Zhuoming Chen, **Xinyu Yang**, Yuandong Tian, Beidi Chen*

Get More with LESS: Synthesizing Recurrence with KV Cache Compression for Efficient LLM Inference — *ICML 2024*

*Harry Dong, **Xinyu Yang**, Zhenyu Zhang, Zhangyang Wang, Yuejie Chi, Beidi Chen*

Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges — *Preprint*

*Chenhang Cui, Yiyang Zhou, **Xinyu Yang**, Shirley Wu, Linjun Zhang, James Zou, Huaxiu Yao*

Improving Out-of-Domain Generalization with Domain Relations — *ICLR 2024 (Spotlight)*

Huaxiu Yao, **Xinyu Yang***, Xinyi Pan, Shengchao Liu, Pang Wei Koh and Chelsea Finn*

Multi-domain Long-Tailed Learning By Augmenting Disentangled Representations — *TMLR*

Xinyu Yang, *Huaxiu Yao*, Allan Zhou and Chelsea Finn*

FlatFormer: Flattened Window Attention for Efficient Point Cloud Transformer — *CVPR 2023*

Zhijian Liu, **Xinyu Yang***, Haotian Tang, Shang Yang and Song Han*

BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation — *ICRA 2023*

Zhijian Liu, Haotian Tang*, Alexander Amini, **Xinyu Yang**, Huizi Mao, Daniela Rus and Song Han*

Variational Inference for Training Graph Neural Networks in Low-Data Regime through Joint Structure-Label Estimation — *KDD 2022*

Danning Lao, **Xinyu Yang***, Qitian Wu and Junchi Yan*