LIANG MI
First-year PhD Student,
School of Computer Science,
Nanjing University

Nanjing University,
163 Xianlin Avenue,
Nanjing, Jiangsu Province,
China.
liangmi@smail.nju.edu.cn

Bio & Education

Sep. 2024 – present	:	PhD in Computer Science at Nanjing University Advisor: Prof. Guihai Chen, Assoc. Prof. Haipeng Dai
Sep. 2021 – Jun. 2024	:	M.S. in Computer Science at Nanjing University Advisor: Assoc. Prof. Haipeng Dai
Sep. 2017 – Jun. 2021	:	B.S. in Software Engineering at Shandong University

Intern Ship

Apr. 2023 – present	:	Research Intern in Institute for AI Industry Research, Tsinghua University Advisor: Prof. Yunxin Liu

Research Interests

I am interested in video analytics and LLM serving systems, and embodied AI. Besides, I have a broad interest in systems, with most of my focus on LMM and high-performance systems.

Publications

Conference Papers (* = equal contributions):

- Empower Vision Applications with LoRA LMM
  Liang Mi, Weijun Wang, Wenming Tu, Qingfeng He, Rui Kong, Xinyu Fang, Yazhu Dong, Yikang Zhang, Yuanchun Li, Meng Li, Haipeng Dai, Guihai Chen and Yunxin Liu, **ACM EuroSys**, 2025.
- Region-based Content Enhancement for Efficient Video Analytics at the Edge
  Weijun Wang*, **Liang Mi***, Shaowei Cen, Haipeng Dai, Yuanchun Li, Xiaoming Fu, Yunxin Liu, Guihai Chen, **USENIX NSDI**, 2025.
- BiSwift: Bandwidth Orchestrator for Multi-Stream Video Analytics on Edge
  Lin Sun*, Weijun Wang*, Tingting Yuan, **Liang Mi**, Haipeng Dai, Liuyun Xin, Xiaoming Fu, **IEEE INFOCOM**, 2024.
- AccDecoder: Accelerated Decoding for Neural-enhanced Video Analytics
  Tingting Yuan*, **Liang Mi***, Weijun Wang, Haipeng Dai, Xiaoming Fu, **IEEE INFOCOM**, 2023.

Journal Papers:

- Accelerated Neural Enhancement for Video Analytics with Video Quality Adaptation
  Liang Mi, Tingting Yuan, Weijun wang, Haipeng Dai, Lin Sun, Jiaqi Zheng, Guihai Chen, Xiaoming Fu, **IEEE Transactions on Networking**.

Manuscript:

- Bandwidth Orchestrator for Multi-Stream Video Analytics on Edge,
  Haipeng Dai, **Liang Mi**, Weijun Wang, Yuanchun Li, Tingting Yuan, Lin Sun, Yuben Qu, Yunxin Liu, Xiaoming Fu, Guihai Chen, Transactions on Networking, submitted

Research Experience

Video Analytics System

In this project, we concentrate on optimizing the performance of the content-enhanced video analytics system.

- We presented AccDecoder, a novel accelerated decoder for real-time and neural-enhanced video analytics.
- We presented region-based content enhancement that enhances only the important regions in videos to improve analytical accuracy.
- We presented BiSwift, a bi-level framework that scales real-time video analytics using an adaptive hybrid codec, multi-level pipelines, and a global bandwidth controller for multiple streams.

Large Multimodal Models

In this project, we present VaLoRA, an end-to-end solution that empowers diverse vision tasks and enriches vision applications with LoRA LMMs.

KV Cache Compression with GPU-native Codec Chips for Fast LLM Serving

KV Cache is crucial for boosting performance for the LLM serving system. Many systems reuse the KV Cache via remote storage, yet this approach inevitably introduces network transmission overhead. To address this, we design a KV Cache encoding and decoding method via codec chips in the GPU that is independent of CUDA/Tensor Core. Our method achieves highly efficient KV compression while not affecting the LLM serving system.