Huanyu WANG

Email: whyisverysmart@sjtu.edu.cn | Tel: (86) 176-1218-9497 | GitHub: whyisverysmart.github.io

Education

Shanghai Jiao Tong University, Shanghai, China Sept. 2022 - Jun. 2026 (Expected)

B.S. in Computer Science & B.S. in Applied Mathematics

GPA: 91.3/100 (Major), 90.3/100 (Overall)

Courses: Game Theory (97), Operating System (95), Programming Languages and Compilers (95), Data Structure (95), Principles and Methods of Program Design (95), Design and Analysis of Algorithms (95), Program Design Practice (94.5), Machine Learning (93), Computer Vision (93), Computer Networks (93).

Publications

- **Huanyu Wang**, Ziyu Xia, Zhuoming Chen, Beidi Chen.
  WWW.Serve: Interconnecting Global LLM Services through Decentralization.
  ICML 2026 (Under Review).
- **Huanyu Wang**, Jushi Kai, Haoli Bai, Lu Hou, Bo Jiang, Ziwei He, Zhouhan Lin.
  Fourier-VLM: Compressing Vision Tokens in the Frequency Domain for Large Vision-Language Models.
  arXiv:2508.06038.
- Chengyang Hu, Xinyu Zhou, **Huanyu Wang**, Danyu Shen, Ran Yi, Mengtian Li, Lizhuang Ma.
  LoMo: Longer and More Videos Benchmark for Understanding and Temporal Grounding Tasks.

Research Experiences

Research Assistant, InfiniAI Lab, Carnegie Mellon University Oct. 2025 - Present

Advisor: Prof: Beidi Chen.

- Reframed LoRA-based RL training as a **service-oriented system**, modeling each rollout-training step as a schedulable request to enable multi-tenant RL-as-a-Service.
- Designed a **router-based framework** that orchestrates requests across multiple rollout servers and trainers, balancing heterogeneous resource demands and improving overall hardware utilization.
- Leveraged the stateless property of LoRA adapters to enable fire-and-forget scheduling, flexible decoupling of rollout and training, and dynamic participation of training services.

Research Assistant, InfiniAI Lab, Carnegie Mellon University Mar. 2025 - Oct. 2025

Advisor: Prof: Beidi Chen.

- Tackled scalability bottlenecks, scheduling inefficiencies, and privacy concerns in centralized multi-LLM serving systems through a novel decentralized approach.
- Designed *WWW.Serve*, a **fully decentralized framework** for trustless collaboration across heterogeneous LLM servers, featuring autonomous workload balancing, flexible policies, and privacy protection.
- Evaluated the framework under diverse configurations and workloads, demonstrating up to **1.5**× improvement in global SLO attainment and **27.6**% reduction in request latency.

Research Assistant, LUMIA Lab, Shanghai Jiao Tong University Oct. 2024 - Aug. 2025

Advisor: Prof. Zhouhan Lin

- Addressed the challenge of computational redundancy in Vision-Language Models (VLMs), significantly improving the efficiency of inference while maintaining competitive accuracy.
- Developed *Fourier-VLM*, which integrates the **2-dimensional Discrete Cosine Transform** with LLaVA and Qwen-VL architectures, achieving a **93.75**% reduction in vision token count.
- Conducted comprehensive evaluations across multiple benchmarks, maintaining **96.1**% performance while achieving **83.8**% FLOPs reduction, **86.4**% less KV cache usage, and **31.2**% faster inference speed.

Skills

Programming: Python, PyTorch, Transformers, Git, Linux, Hugging Face, C++, LaTeX.

Mathematics: Linear Algebra, Discrete mathematics, Information Theory, Probability & Statistics, ODE.

Language: English - TOEFL: 107 (Speaking 24).