Xinyuan Li

No.866, Yuhangtang Road, Xihu District, Hangzhou, Zhejiang Province, China
+86 130 3529 0600 xinyuanli2327@outlook.com Website GitHub Blog

Educational Experience

Zhejiang University
Hangzhou, Zhejiang

B.E. in Automation Engineering, Minor in ACEE Honor Class of CKC; Sept 2023 - Jun 2027

- Ranked in the top 5 % of the major in the first academic year.
- GPA: 3.95/4.00

Research Experience

MLL Lab
Jul 2025 - Present

Research Intern Advisor: Prof. Manling Li

- **Project: Cross-View Spatial Consistency for UMMs** *(targeting ECCV 2026)*
- Co-developed a research framework to enhance spatial reasoning in UMMs; iterated methodology with senior PhDs.
- Built the end-to-end data processing pipeline and managed the model training process.
- Established a comprehensive evaluation suite by integrating bencmarks (e.g. MindCube, MMSI) to rigorously assess performance.

CAD&CG Lab
Jan 2025 - Mar 2025

Research Intern Advisor: Prof. Sida Peng

- **Project: Split 4D**
- Established research workflow, including GPU resource management, systematic experimental tracking, and reproduction of SOTA baselines.

Selected Projects

CMU 11-868: Large Language Model Systems
Jul 2025 - Sep 2025

- Developed a **full-stack LLM system** from scratch, including a custom auto-differentiation engine and high-performance **CUDA kernels** (Softmax, Attention, Layernorm) for Transformer components. [Code]
- Implemented distributed training paradigms including **Data, Tensor, and Pipeline Parallelism** to scale model training across multiple GPUs.
- Optimized inference performance by integrating **PagedAttention** (vLLM), **SGLang**, and model **quantization** techniques to enhance KV Cache management and throughput.

NeRF Replication
Jun 2025

- Replicated the Neural Radiance Fields (NeRF) pipeline, implementing volume rendering and positional encoding to achieve high-fidelity 3D scene reconstruction. [Code]

Mathematical Foundations of Reinforcement Learning
Apr 2025 - May 2025

- Conducted in-depth analysis of RL convergence and optimality, implementing core algorithms (Value/Policy Iteration, Q-Learning) with a focus on theoretical guarantees. [Notes]

Skills

- **Programming:** Python, C++, CUDA, Bash

- **Deep Learning:** PyTorch, vLLM, DeepSpeed, FSDP
- **Computer Vision & Robotics:** NeRF, Gaussian Splatting
- **Tools & Infrastructure:** Linux (SSH/Tmux), Git, Slurm, Weights & Biases (W&B), LaTeX