

Kevin Xiang Li

<https://kevinx.li>

Email: kevinx.li@outlook.com

TEL: 734-510-0189

SKILLS

Machine Learning: PyTorch, HuggingFace Transformers, Unsloth, Nsight, Gymnasium, Python, C++

Mobile Dev: Flutter, SQLite, Rust **AR/VR:** Unreal, Unity, C# **Web:** TypeScript, JavaScript, HTML, CSS

EDUCATION

- **Stanford University** Stanford, CA, U.S.A
M.S. in Computer Science; GPA: 4.0 2024. 9 – 2026. 6
 - **Course Highlights:** Reinforcement Learning, Deep RL, Spoken Language Processing, Machine Learning with Graphs, Infrastructure at Scale, Computer Networking
- **University of Michigan** Ann Arbor, MI, U.S.A
B.S. in Computer Science, Minor in Linguistics; GPA: 3.87, Summa Cum Laude. 2020. 9 – 2024. 5
 - **Course Highlights:** Intro to ML, Intro to NLP, Computer Vision, XR & Society, Programming Languages, Compiler Construction, Intro to Operating Systems, Computer Security

EXPERIENCES

- **LLM Inference Workload Performance** Santa Clara, U.S.A.
ML Engineer at Nvidia 2025. 6 - 2025. 9 (Expected)
 - **Benchmarking leading VLMs on large scale GPU clusters:** Measuring throughput and latency of leading open source VLMs like Llama 4 and Qwen 2.5 VL on large scale H100, H200, and B200 clusters.
 - **Analyzing VLM inference bottlenecks:** Pinpointing kernel-level performance bottlenecks with Nsight and PyTorch Profiler on SGLang; calculating and comparing with theoretical FLOP estimates to identify perf gaps.
 - **Contributing to Open Source Inference Frameworks:** Working with SGLang team to profile and optimize VLM performance; Submitting PR to improve visual processing and profiling support for latest Llama 4 model.
- **VideoMultiAgents: A Multi-Agent Framework for Video QA** Stanford University, U.S.A.
Researcher, in collaboration with Panasonic 2024. 10 - 2025. 3
 - **Designed multi-agent framework with modality-specific agents for video QA:** Enhanced video understanding by leveraging strengths of video, text, and graph modalities through multi-agent collaboration.
 - **Discovered that modality-specific multi-agent architectures benefit from structure and independence:** Showed that our Report architecture performs the best by aggregating opinions from independent modality-specific agents through an organizer agent and weighing strength of evidence from each modality.
 - **Achieved SOTA accuracy on popular video QA benchmarks:** Improved previous SOTA on Intent-QA by +6.2%, EgoSchema subset +3.4%, and NExT-QA by +0.4%.
- **LLM Hub Supporting Fine-tuning, Inference, and Evaluation** Shanghai, China
ML Engineer at GienTech Technology 2024. 6 - 2024. 8
 - **Devised Evaluations for LLMs:** Evaluated LLMs on metrics like BLEU, ROUGE, Levenshtein Distance, and LLM-as-a-Judge methods. Incorporated evaluation module into existing PoC product.
 - **Implemented Instruction Selection and Generation:** Leveraged latest techniques like CaR (Clustering and Ranking) and Self-Instruct to select and generate instructions for more efficient and performant fine-tuning.
 - **Fine-tuned LLMs on Multiple GPUs:** Utilized popular frameworks like LLaMA-Factory and DeepSpeed to fine-tune 16 open LLMs on multiple GPUs.
- **Live-It: Image-to-3D Scene Generation** Berkeley, U.S.A.
Team Lead, UC Berkeley AI Hackathon 2025 2025. 6
 - **Led a team of 3 to win at the world's largest AI Hackathon:** Developed Live-it, a project that transforms any image into an explorable 3D world by combining the strengths of video diffusion model with Gaussian Splatting. Won the Nitrode Turbo Mode Award amongst 1,400+ hackers and 350 projects.
 - **Architected a real-time image-to-3D pipeline integrating multiple models:** Orchestrated a workflow using Veo 3 for video generation and 3D Gaussian Splatting for reconstruction. Leveraged Visual Geometry Grounded Transformer (VGGT) for its fast feed-forward inference of camera trajectories and 3D point clouds, enabling an instant 3D scene preview which was then continuously refined in real-time.