

Kevin Xiang Li

<https://kevinx.li>

Email: kevinx.li@outlook.com

TEL: 734-510-0189

SKILLS

Machine Learning: SGLang, vLLM, Unsloth, Nsight, Gymnasium, HuggingFace, PyTorch, Python

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
 - **Benchmarked VLMs on large GPU clusters:** Measured throughput and latency of the Qwen 2.5 VL family (3B–72B) across H200 and B200 clusters.
 - **Pinpointed and reported inference bottlenecks:** Used Nsight, NVTX markers, and PyTorch Profiler to pinpoint kernel-level bottlenecks in SGLang and vLLM; dissected framework performance gaps under varying concurrency and provided detailed reports well received by both SGLang and vLLM multimodal teams.
 - **Submitted 5 PRs to SGLang that boosted Qwen 2.5 VL throughput by 1.6x end-to-end on MMMU:** (1) Doubled Qwen 2.5 VL vision prefill speed via automatic attention backend selection, (2) Accelerated rotary embedding with CUDA rotary kernels, boosting vision prefill throughput by 21%, (3) Identified and removed redundant device-to-host visual feature transfers to enable accelerated GPU hashing, yielding 7.5% end-to-end speedup on MMMU, (4) Fused SwiGLU in ViT to double peak TensorCore utilization, resulting in 4.5% vision prefill throughput gain, (5) Unified VLM benchmarking to support reliable cross-framework comparisons.
- **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%.
- **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.
 - **Featured by Google DeepMind staff on LinkedIn:** Project featured by Google's hackathon judge on LinkedIn, received 171 upvotes.