Kengyuan Qiu

kennyqiu0312@gmail.com | GitHub: github.com/kenny0312 | Personal Website: kenny0312.github.io | LinkedIn

Education

MS, **Electrical and Computer Engineering** *New York University* — Aug 2024 – May 2026
BS, **Computer Engineering** *South China Agricultural University* — Sep 2020 – Jun 2024

Work Experience

Machine Learning Engineer | VortexNet, Covina, CA — March 2026 – June 2026

- Built a production **video-understanding agent** based on **Gemini 3.5**, integrating a self-designed **agent loop** and **multi-turn dialogue** to answer open-ended questions over raw video content with playable clips and charts as evidence.
- Designed a **probe-and-step agent loop** with multi-turn memory, context management, and an extensible **MCP-based tool framework** supporting video retrieval, database operations, evidence generation, and custom tool execution.
- Deployed the system with **FastAPI** on **Google Cloud Run**, implemented **SSE streaming** for real-time responses, and added per-request cost telemetry using **BigQuery** and Cloud Logging.
- Built a **τ^2-bench-style evaluation suite** for multi-turn video-agent reliability and improved **pass@k by 8%** through **GEPA-based prompt evolution** and tool-use policy optimization.

Research Assistant | FAMS Lab, NY — Nov 2025 – Jan 2026

- Built a legal consultation chatbot based on LLMs, integrating **intent recognition**, **query rewriting**, and **RAG**, and improved response accuracy through **instruction tuning** and **reinforcement learning**.
- Developed an intent recognition module with **Qwen2.5-1.5B** to filter non-legal queries, perform query rewriting and keyword extraction, improving input quality for the downstream retrieval module.
- Constructed a legal knowledge base using recursive and sliding-window chunking strategies, generating vector indexes with **BGE-Large** embeddings.
- Built a RAG-based QA system with **Qwen2.5-7B**, implementing dense + sparse **dual-recall** retrieval, BGE **reranking**, and an **Agent** module to trigger web search when needed, improving accuracy from **74%** $\rightarrow$ **83.6%** (+9.6%) over the baseline.
- Constructed a 100K instruction dataset from public data, legal QA data, and LLM-generated data, applied **LoRA SFT** (+6%), then built a 5K preference dataset via multi-model scoring for **DPO** fine-tuning (+4% over SFT).

Selected Projects

Natural-Language Image Retrieval System | FAMS Lab, NY — Sep 2025 – Dec 2025

- Built a natural-language image-retrieval system with **LLM-based intent recognition** (Qwen2.5-3B) and multimodal image–text matching, letting users pull specific categories and quantities from large-scale datasets.
- Designed a unified text-to-image and image-to-image retrieval framework; benchmarked the CLIP family (CLIP, SigLIP, Taiyi-CLIP) and added image–text feature fusion, raising retrieval F1 by **10–14%**.
- Fine-tuned a fine-grained retrieval module (Qwen2.5-VL-7B) with **LoRA** on a 4K multimodal dataset, reaching **89% / 92.5%** F1 (text-to-image / image-to-image) in complex scenarios.

DINOv2-GraphNAV: Graph-based Visual Navigation System | NYU — Sep 2024 – Jun 2025

- Designed a scalable graph-based visual navigation pipeline leveraging **DINOv2** global and local features to construct topological maps from offline exploration trajectories.
- Built a KNN similarity graph over 10K+ keyframes with weighted edges combining global embeddings and **PCA-reduced** local patch descriptors, enabling viewpoint-invariant localization and robust spatial reasoning.
- Implemented real-time online localization using **FAISS** approximate nearest neighbor search with a lightweight DINOv2 backbone, achieving **sub-100ms** inference latency per query.

Skills

- **Programming & Frameworks:** Python, PyTorch, NumPy, OpenCV, CUDA, SQL, Git, Docker
- **Retrieval & Multimodal AI:** FAISS, BM25, reranking, BGE embeddings, CLIP, SigLIP, Qwen-VL, DINOv2, ViT
- **LLM & Agent Systems:** RAG, function calling, MCP, prompt optimization, multi-turn dialogue, agent evaluation
- **Model Training & Fine-Tuning:** LoRA, SFT, DPO, RLHF, GEPA, instruction tuning, preference optimization
- **Backend & Cloud:** FastAPI, REST APIs, Google Cloud Run, BigQuery, Cloud Logging, Vertex AI, CI/CD, pytest