April (Yutong) Yang

650-269-9357 | aprilyangyyt@gmail.com | linkedin.com/in/yutongayang | github.com/april-yyt | april-yang.com

Experience

NVIDIA Santa Clara, CA

Research Engineer (Multimodal Data Understanding & Retrieval) *Feb. 2025 – Present*

VLM-Based Multimodal Retrieval

- Developed a VLM-based video search pipeline from scratch and scaled it from prototype into the team's production data-search platform: VLMs caption clips with structured driving descriptions, which are embedded as text vectors for semantic retrieval over a petabyte-scale autonomous-vehicle video datalake to mine long-tail, safety-critical driving scenarios.
- Developed production video-captioning pipelines on 7B/8B video-reasoning VLMs, including the parallel multi-round structured captioner published at a CVPR 2026 workshop (pSVLMs, below) and a unified captioning library with pluggable inference backends.
- Shipped an LLM-based query-understanding layer in the search pipeline that rewrites user queries into caption-aligned variants and fuses retrieval across the rewrites, achieving a **42% relative** precision gain over the caption-as-query baseline on complex driving-behavior queries.
- Shipped an embedding-map explorer (UMAP/HDBSCAN with LLM topic extraction) that lets users understand and categorize the scenario classes in any custom slice of data and surfaces under-covered scenario clusters; presented the coverage gaps to steer safety-critical data collection.
- Built the team's retrieval evaluation from scratch: two hand-curated benchmarks, Qwen3 embedding and reranker baselines, and an **MTEB-style leaderboard** reporting per-scenario AP, P@K, and MRR.

VLM Post-Training & Evaluation

- Drove evaluation for the team's VLM fine-tuning effort, which adapts 7B/8B-sized video VLMs (LoRA SFT on curated AV caption data) for driving-scene understanding; owned the evaluation module end to end, from metric design to deciding which fine-tuned checkpoints ship to production.
- Built the team's end-to-end **VLM fine-tuning evaluation pipeline** (distributed batch inference, LLM-as-judge and reranker scoring with degenerate-output guardrails, results dashboard), covering 8 benchmarks spanning caption quality, retrieval performance, and embedding scenario separability.
- Provided production support for live fine-tuning runs: **agent-orchestrated** watch loops monitored per-checkpoint inference and scoring workflows, kept long jobs resumable, and compared candidates against the in-production model before release.
- Codified the entire evaluation workflow as reusable **LLM-agent skills**: from a single request, an agent launches inference across all eval dataset splits, runs scoring, publishes metrics, and registers the checkpoint in the results dashboard, making full checkpoint evaluations hands-off and reproducible.
- Designed the evaluation **metric suite**: per-section factual correctness (LLM-judge and multi-agent LLM-verifier with written rationale, plus a learned reranker relevance score), temporal correctness, external NLG benchmarks (BLEU/METEOR/CIDEr, LingoJudge on LingoQA), and caption-embedding-driven metrics, including nearest-centroid scenario separability and retrieval metrics over a densely labeled internal dataset.
- Connected evaluation back to training: prototyped GRPO-style RL fine-tuning on an SFT checkpoint with eval-derived rewards (embedding similarity plus reranker score) to test which eval metrics are robust enough to serve as training signals.

Agentic Search & Data Curation

- Built the MVP of the team's **agentic search-refinement loop**, automating a scenario-curation process that previously took **hours of manual search and review per scenario**: a **rewriter** agent reformulates the query, a **search** agent retrieves candidates, a **judge** agent verifies them, and a **refiner** trains a linear-probe classifier on the verdicts or refines the query, looping until precision converges.
- Hardened the loop into a data-curation **agent SDK** the team uses to run scenario curation end to end: each agent stage behind typed contracts with deterministic and LLM tool-calling orchestrators, every run traced, seedable, and testable offline; rebuilt it **67%** smaller with API-drift CI.
- Built an **agentic hybrid-search system** that turns a plain-English data request into a curated clip set: an LLM decomposes the request into a multi-source retrieval plan with server-side rank fusion, then iteratively tunes ranking weights and queries against a VLM relevance judge until per-source precision converges.
- Developed its **evaluation harness**: manifest-backed **retrieval benchmarks** composed into regression and capability suites with pass-rate expectations, per-stage evaluators (retrieval nDCG/MRR, rewriting, decomposition, judge precision), and seeded repeated trials reporting **pass@k** reliability; exported the same benchmarks as containerized **Harbor** tasks, scoring coding agents (Claude Code, Codex, oracle baselines) end to end with the identical judge and metrics so component-level and agent-level results are directly comparable.
- Curated **13,800+ clips across 84 long-tail scenarios** via agentic search (LLM planner + VLM judge); hardened the judge with object-detector cross-checks, improving precision on **4 of 6** failure scenarios in a controlled A/B study.
- Shipped an **agent-console observability UI** (FastAPI + React) with live run traces, a per-step execution timeline, and playback of the judged video evidence, making agent runs and failures inspectable end to end.

NVIDIA Santa Clara, CA

Software Engineer Intern, LLMs *May 2024 – Aug. 2024*

- Built and deployed a multi-agent RAG system for AV engineering Q&A, test, and code generation (LangChain, LangGraph, GraphRAG) on Kubernetes, serving **3,000+ engineers**, with an internally measured **50%** efficiency gain on that workflow.

Publications & Research

Scalable Parallel Prompting for Complex AV Video Captioning | *CVPR 2026 Workshop, co-first author* Jun. 2026

- Proposed **pSVLMs**, a parallel multi-round prompting framework in which small VLMs caption the scene, road entities, and key driving actions in parallel and consolidate them into one structured caption, curbing the compounding hallucinations and cost of large-VLM captioning; improved top-10 retrieval precision by up to **11 points** over baselines. Powers the production captioner behind the team's video search.

SIL-Wheel: A Multi-Modal Search and Curation Platform for Physical AI Systems | *NVIDIA whitepaper* 2026

- Authored the embedding-space analysis experiments (**4.55M** embeddings, three encoder families): showed mean-centering corrects encoder anisotropy (silhouette 0.10 $\rightarrow$ 0.75) and that cluster-diversified retrieval beats relevance-ranked sampling on a 1M-clip benchmark; includes a dual-lens coverage methodology that cross-checks caption and driving-dynamics embeddings, surfacing an under-covered scenario cluster at $\sim$**15**$\times$ the average failure rate.

Distilling Robot Foundation Models into Lightweight Policies | *Stanford CS224R (Deep RL) project* 2026

- Designed the **student policy architecture** (dual-camera and proprioception encoders with an MLP action-chunk head) and implemented the **IL/RL algorithms** (behavior cloning, DAgger, **AWAC**) that distill a robot-foundation-model teacher, using teacher rollouts as demonstrations and teacher value estimates as a dense reward for sparse-reward tasks; students reached **98%** success on 15 LIBERO manipulation tasks at $\sim$**400**$\times$ the teacher's inference speed.

Research-Tracker Agent & Conference Paper Search MCP Server | *FastMCP, SQLite, GitLab CI* 2026

- Built a live **MCP server** indexing conference papers (crawler, typed search tools) plus a Slack agent that tracks and answers questions about AV research papers via **conference-paper MCP catalogs** (CVPR, NeurIPS); ships continuously via self-hosted CI/CD.

Education

Stanford University Stanford, CA
Graduate coursework, Robotics & Autonomous Systems program (incl. Deep Reinforcement Learning) *2026 – Present*

Carnegie Mellon University Mountain View, CA
M.S. in Software Engineering, Electrical & Computer Engineering; GPA: 3.8/4.0; TA for Intro. to ML *Aug. 2023 – Dec. 2024*

University of California, Berkeley Berkeley, CA
M.Eng. in Industrial Engineering & Operations Research (FinTech); GPA: 3.98/4.0 *Aug. 2022 – May 2023*

Renmin University of China Beijing, China
B.S. in Management Science (Business Analytics); GPA: 3.82/4.0; National Scholarship (Top 1%) *Sep. 2018 – Jul. 2022*

Technical Skills

Languages: Python, SQL, C/C++, TypeScript/JavaScript, Swift
ML Training & Inference: PyTorch, HuggingFace, LoRA/QLoRA SFT, GRPO/DPO/AWAC post-training, vLLM serving
Evaluation: LLM/VLM-as-judge, benchmark & eval-harness design, agent evaluation (Harbor/NemoGym)
Retrieval & Agents: vector search (FAISS, Milvus), embedding models & rerankers, rank fusion, RAG, agentic search, MCP
Data Curation: multimodal (video/vision-language) curation, captioning, retrieval, clustering (UMAP/HDBSCAN)
Infrastructure: Docker, Kubernetes, Slurm, GitLab CI/CD, AWS S3, Databricks/Spark, FastAPI, React, Streamlit, Plotly