

Jianzhong He

HeJianzhong@berkeley.edu | (408) 361-0997 | San Jose, CA | andy-he.com

SUMMARY

Software engineer building an **autonomous video agent** handling **10K+ requests/day**, and the **agent infrastructure** underneath: **multi-agent orchestration**, **RAG**, **memory**, **MCP tools**, **GPU inference**, and **evaluation** systems.

PROFESSIONAL EXPERIENCE

Creatify.ai

Software Engineer

Mountain View, CA

April 2025 – Present

- Developed an **autonomous video agent** and its runtime: a **plan** → **retrieve** → **generate** → **edit** loop with **multi-agent orchestration**, **per-session sandbox isolation**, per-stage tracing, and **self-hosted LLM inference** on a dedicated **8×B300 SGLang** cluster, sustaining **2,000+ videos/day**.
- Designed the **RAG** system behind high-conversion ad script generation and **fine-tuned GPT-4.1** on **30K+ curated historical ads**: dual embedding indexes over scripts and products, category-scoped recall re-ranked on **live ad-performance signals** (ROAS, CPA, CTR), lifting scripts **12%** above competing ad platforms.
- Built an **LLM-as-judge** eval harness with **16 quality checks** and **pairwise model comparison** across 7 dimensions, catching regressions before release and lifting **human-eval good/fair** from **40% to 80%**.
- Built the **MCP** server exposing **40 tools** across Claude, ChatGPT, and Codex, with interactive App cards and per-client telemetry, serving **400+ weekly client authorizations** at a **0.41% tool failure rate**.
- Built **active-speaker detection** from scratch (**SyncNet** + **TalkNet**) on **multiple GPUs** via Modal: MediaPipe face tracking driving per-scene audio-visual sync scoring for **lip-sync QA** on every scene of generated video.

HDWAlabs.ai

Software Engineer

San Francisco, CA

August 2024 – March 2025

- Developed an agentic data operator on **Apache Airflow**, generating standardized ETL pipelines as **versioned DAGs** (extract → transform → validate → load) with reusable operators, dependency management, retries, and SLA monitoring. Automated multi-source onboarding via connector abstractions (schema discovery + lightweight profiling) to recommend extraction strategies and reduce manual pipeline setup.
- Integrated LLMs to translate natural-language data questions into **executable ETL/SQL code**, run governed extraction and transformation jobs, and return **human-readable summaries**—enabling self-serve analytics while enforcing access controls and governance policies.

DQLabs

Software Engineer Intern, API Backend

Pasadena, CA

May 2024 – July 2024

- Developed user-facing and private **REST APIs** in **Django**. Used Swagger UI to generate interactive API documentation. Incorporated internal API tests into the CI/CD pipeline, increasing service availability by 4.5%.
- Developed a **GitHub CI** add-on that automatically tracks downstream dependencies on each **dbt** commit, automating impact analysis and prioritizing impacted assets for testing based on usage, criticality, and data volume. Triggered 50+ times per week.

University of California, Santa Cruz

Data Science Intern, Science Internship Program

Santa Cruz, CA

May 2021 – August 2021

- Worked with Prof. Raja's research group to develop **Python-based machine learning** pipelines. Used **scikit-learn** and **TensorFlow** to classify weak CN stars and carbon stars in the M31, M33, and LMC datasets.
- Refined models through **hyperparameter tuning**, **cross-validation**, and imbalanced-data techniques such as **SMOTE**. Improved classification accuracy by 15%.

EDUCATION

University of California, Berkeley

Bachelor of Arts in Computer Science

Berkeley, CA

June 2023