

Daniel Jiang

zhuoyuan.jiang625@gmail.com | New Haven, CT

[LinkedIn](#) | [Github](#)

EDUCATION

Yale University

September 2024 – Expected May 2026

Master of Science (MS) Honors (Highest Grade Level), Biostatistics - Data Science Pathway, Department Student Representative

- Relevant Courses: Advanced Machine Learning, Deep Learning, NLP, Computer Vision, Trustworthy AI, Software Engineering, Building AI Infra System, Computer Network, Computer Architecture

University of California, Santa Barbara (UCSB)

September 2019 – June 2023

Bachelor of Science (BS) High Honors (Top 6% of class), Psychological & Brain Sciences and Financial Mathematics & Statistics

- Cumulative GPA: 3.94/4.0

SKILLS

- **AI & LLM & VLM:** Fine-tuning (QLoRA, SFT, RLHF/DPO), RAG, Multimodal systems (text/audio/video)
- **Deep Learning & ML:** Transformers, Diffusion, Flow Matching, 3DGS, NeRF, GANs, LSTMs; Classical ML (e.g. XGBoost)
- **Frameworks, Inference & Deployment (MLOps):** PyTorch, TensorFlow, Lightning; Hugging Face, vLLM, Triton; W&B
- **Full Stack Software:** Python, C++, HTML, JavaScript, TypeScript, React, Next.js, Vite, Flask, FastAPI, Docker, Git, CI/CD
- **Cloud & Database:** AWS, GCP, Supabase, BigQuery, PostgreSQL, MongoDB; **Dev Tools:** Claude Code, Codex, Cursor

PROJECTS

Lingo: AI Chrome extension for Real-time Slang & Cultural-Reference Learning | *Startup Project*

April 2025 – Present

- Spearheaded “Lingo” from ideation to launch to help non-native English speakers bridge cultural nuances and speak more natural and idiomatic English by auto-surfacing slang, phrasal verbs, cultural references & wordplay in their day-to-day content feed.
- Built data → eval pipelines: curated 122GB raw + synthetic benchmark sets and automated evaluation to select cost-effective solutions across open-source models (e.g. Hugging Face) and LLM APIs; trained a lightweight in-house model as a baseline.
- Engineered the production stack & DevOps layer: React + TypeScript front end, FastAPI + PostgreSQL back end, Dockerized on GCP with automated CI/CD for Chrome Web Store deployment.

Audio Fashion Retrieval Agent: Two-Tower Architecture for Composed Fashion Search [Code](#) | [Model](#)

2026

- Built an audio-native multi-modal fashion item retrieval system achieving **SOTA Recall@10 = 0.637** on a 59K-item benchmark, beating SOTA fashion encoder Marqo-FashionCLIP (0.533) by **+19.5%** and Qwen3-Embedding-8B (0.522) by **+22.0%**.
- Pioneered a two-tower joint embedding architecture for multi-modal retrieval, replacing traditional separate-encoder pipelines; evaluated 3 retrieval architectures and 15+ model/training variants, improving R@10 from a 0.240 baseline to 0.637 (**+165%**).
- Synthesized 56.7K gender-balanced, multi-accent audio training triplets across 110 speakers, providing a high-quality audio corpus for training models robust to speaker, gender, and accent variation.
- Built distributed training and eval infrastructure, scaling PEFT LoRA + bf16 contrastive learning across multiple 8× A6000 GPU nodes with dynamic database re-encoding, W&B tracking, and a Gradio demo UI for qualitative inspection.

Nutrition-Table-Detector: End-to-End Multimodal Localization System [Code](#) | [Model](#)

2025

- Built an end-to-end pipeline to fine-tune Qwen2-VL-7B for nutrition-table localization on 8x RTX 6000 Ada using 4-bit NF4 QLoRA to power scalable inventory digitization, raising detection rate **from 18% to 100%** and IoU by **+780% (0.09 to 0.86)**.
- Built 4 training recipes (LLM-only LoRA, vision-only full fine-tuning, two-stage, joint LoRA) + 3 custom collators and conducted systematic ablations; delivering an extra **+30.3%** IoU gain (**0.66 to 0.86**) while cutting training time **6× (6 to 1 hr/epoch)**.
- Implemented 4-bit and 8-bit quantization; selected GPTQ INT4 (W4A16, LLM-only) and served via vLLM on RTX 3090, achieving **2.4×** compression, **64%** faster decode (TPOT), and **+34%** throughput (c=1) with negligible accuracy loss (−0.74% IoU).
- Deployed an end-to-end multimodal serving stack using Triton with a vLLM backend, achieving **2.9×** throughput at c=8 (3.17 vs 1.09 req/s) and **1.31×** lower single-request latency at c=1 (907 vs 1189 ms) compared with Hugging Face.

Omni-Qwen: Unified Audio-Vision-Language Omni Model [Code](#) | [Model](#)

2025

- Architected an omni model by extending Qwen2-VL-7B with an audio-token pathway, achieving unified audio-vision-language understanding and a **7.90%** WER on the LargeScaleASR benchmark.
- Designed a two-stage training pipeline (modality alignment → LoRA) and scaled training to 6× RTX 6000 Ada GPUs with DDP, improving WER from 14.5% after alignment to 7.90% after LoRA while reducing epoch time 8× across 107K audio samples.

SmolLM-GEC-135M: Style Preserving Grammatical Error Correction(GEC) LLM [Code](#) | [Model](#)

2025

- Built an end-to-end GEC pipeline (SFT → custom dataset creation → DPO/IPO) for faster, style-preserving grammar correction; scaled hyperparameter search on 8×RTX3090 GPUs, delivering a **+880%** BLEU improvement (**0.051 to 0.500**) over baseline.
- Designed optimization strategies — SFT + padding; sequence- and batch-level packing; prompt engineering; data collator (full sequence vs. response only); DPO variants (DPO, IPO) — improving BLEU score over the initial fine-tuned model by **+190%**.

EXPERIENCE

UCM Vision and Learning Lab, Merced, CA

Machine Learning Research Intern

June 2025 – Present

- Developing a feed-forward, audio-driven 4D avatar system that generates personalized talking avatars from a single face image and speech, without per-subject optimization.
- Implemented multi-view training and evaluation pipelines for view-consistent talking-head generation, targeting identity preservation, facial dynamics, and cross-view consistency.
- Built a scalable >13TB data pipeline with resumable processing and auto-validation, reducing data prep ~80% and QA time ~90%.

UCSB Center for Financial Mathematics and Actuarial Research, Santa Barbara, CA

Machine Learning Researcher

September 2022 – May 2023

- Led mortality regression analysis for ages 60-84 across 50+ states & 58 CA counties to address the challenge of predicting more granular state-/county-level mortality rates and analyzing risk drivers, informing public health strategies and actuarial pricing.
- Innovated a novel ML framework of Multi-Output Gaussian Process (MOGP) model beyond standard SOGP model, which is better at capturing data dependencies and fusing data, yielding more realistic & interpretable predictions.
- Built the end-to-end ML workflow: developed ETL pipelines (R, Python); optimized ML training pipeline, reducing iteration time by **95%** (5+ hrs saved/experimentation); and deployed findings via R-shiny app, empowering interactive insight extraction.