Shengguang Cui

(310)-694-4245 | lilil@ucla.edu | www.linkedin.com/in/shengguang-cui | Los Angeles, CA, 90095

Education

University of California, Los Angeles *Los Angeles, CA*

M.S. in Electronic and Computer Engineering *Sept 2024 - Jun 2026 (expected)*

- **Cumulative GPA:** 4.0/4.0
- **Core Curriculum:** Information Theory, Computational Robotics, Deep Learning, Trustworthy AI, AI on Chips, Neural Signal Processing

The Chinese University of Hong Kong *Shenzhen, China*

B.Eng. in Electronic Information Engineering - Computer Engineering *Sept 2020 - May 2024*

- **Cumulative GPA:** 3.7/4.0 (Rank 5/100 in major)
- **Core Curriculum:** Computer Architecture, Operating System, Parallel Computing, Database Systems, Algorithm Analysis, Machine Learning

Experience

Videospace, Inc. *Jun 2025 – Jan 2026*

Intern, Engineering and AI Research *Los Angeles, CA*

- Shipped a **live captioning system** for Fortune 500 clients which **reduced caption latency from 6s to 1s** by engineering a **hybrid Java-Python pipeline** on Wowza Streaming Engine to pipe resampled audio to speech-to-text services and convert the output to ID3 tags for player rendering
- Enhanced live webcasts with **real-time translation** and **live AI highlights** by developing a **Python** generation module utilizing **OpenAI** and **Gemini APIs**; added dynamic user controls for language and vocabulary by routing configuration updates through Wowza's **REST API endpoints**
- Scaled production inference by deploying in-house **Whisper**, **PaddleOCR**, and **LLaMA** models on **Lightning AI** with GPU auto-scaling and API serving capabilities, while evaluating **LoRA** and **RAG** strategies to implement continuous knowledge injection in production
- Built a **Slack integration module** with **Django** models, views, and **REST API** endpoints to manage OAuth authorization and asynchronous message retrieval, enabling users to search and view LLM-generated summaries of relevant discussions from Slack workspaces
- Architected an asynchronous **Message Processing Engine** using **Celery** to execute tasks dispatched from the Django API endpoints; engineered a distributed pipeline to fetch slack messages, segment conversations into topic clusters, and generate structured summaries and titles

University of California, Los Angeles *Mar 2025 - May 2025*

Research Assistant *Los Angeles, CA*

- Researched hallucination reduction in **Large Vision-Language Models (LVLMs)**, focusing on inference-time methods
- Proposed a contrastive decoding approach (based on On-the-Fly Preference Alignment) to steer LVLM outputs toward factual content
- Re-implemented Activation Steering Decoding (ASD) as a baseline and developed a **PyTorch** implementation of the proposed method; benchmarked on **LLaVA-1.5** with **POPE** and **LLaVA-Bench**, demonstrating the proposed method reduced hallucinations relative to ASD

Duke Kunshan University *Feb 2023 - Dec 2023*

Research Intern of Federated Learning *Shenzhen, China*

- Designed HeteroPruneFL, a heterogeneity-aware **Federated Learning** framework that assigns each edge device a client-specific subnetwork via importance-based pruning, enabling training within restricted compute/memory budgets while preserving comparable model performance
- Introduced dynamic sparse training (prune–regrow) in client-side training to adapt per-client network topology to local data within memory caps
- Engineered a reproducible **PyTorch** stack (server–client orchestration, autorun scripts, analysis pipeline) and benchmarked on 4 datasets vs. 3 baselines, showing consistent accuracy gains compared to all the baselines under different resource constraints

Projects

Automated Privacy Testing for LLM through fuzzing *Jan 2025 - Mar 2025*

Project for course *Trustworthy AI, UCLA* *Los Angeles, CA*

- Built a **privacy harness** in **Python** to test **LLM**s' PII extraction from HTML profiles; extended PROMPTFUZZ with privacy-oriented mutators (context injection, persona/role prompts, obfuscation, pseudocode) and HTML-aware templating to generate adversarial prompts automatically
- Uncovered a **+7 pp** lift in attack success rate **(85%→92%)** on GPT-4o through iterative mutation and response analysis

Campus Image Generative Models *Mar 2023 - Apr 2023*

Project for course *Advanced Machine Learning, CUHK* *Shenzhen, China*

- Collected and curated a CUHK-Shenzhen campus image dataset; applied **preprocessing** and **augmentation** to build a high-quality training set
- Trained **GAN**, **DCGAN**, and a **diffusion model** (with ImageNet pretraining) in **PyTorch**; conducted a comparative study using direct and **t-SNE/LLE** visualizations, confirming diffusion outperformed GAN and DCGAN in realism and mode coverage

Technical Skills

Programming Languages: Python, C/C++, Java, SQL, Javascript, MATLAB

Frameworks & Tools: PyTorch, CUDA, HuggingFace, Scikit-Learn, MySQL, Django, REST APIs, Git, Celery, Lightning/LitServe