Clark Peng

San Jose, CA (U.S. Citizen) | cipeng@ucla.edu | linkedin.com/in/clarkpeng | github.com/clarkipeng | clarkipeng.github.io

Education

University of California, Los Angeles — Los Angeles, CA
B.S. in Computer Science (GPA: 3.95) — *Expected June 2028*

Research and Engineering Experience

UCLA NLP Group — Los Angeles, CA
Student Researcher — *Oct 2024 – Present*

- Co-first author of **VideoPhy-2**, a video-physics benchmark (ICLR 2026, ICML workshop Best Paper); 170+ citations.
- Built its data-generation pipeline, expanding it 6x, and generated all 6,800 videos across 7 models, quantizing open models and parallelizing inference with Ray and Ring/Ulysses sequence parallelism.
- Trained VideoPhy-2-AutoEval, a judge that correlates 81–236% better with human ratings than Gemini-2.0-Flash-Exp.
- Co-authored **DialectGen** on dialect robustness in image and video generation (ResponsibleFM, NeurIPS 2025).
- Second author of **Cross-Lingual Alignment for Decoder-Only Models using MoE Routers**, aligning MoE routing so models up to Qwen3.5-VL-30B learn new languages with fewer samples; packed teacher and student batches for 2.1x faster training, now the team default.

Catapult (YC S26) — San Francisco, CA
Co-founder and CTO — *Jun 2026 – Present*

- Built the game API for a benchmark that ran 19 frontier LLMs from 7 labs through 1K games, producing 257M tokens of reproducible gameplay.
- Built the core of a 30-game platform on Postgres and Durable Objects, with idempotent moves and change-aware CI.
- Built the first version of Music Arena, where AI models make music in FL Studio, plus its plugin and flpkit (PyPI).

Camfer (YC S24) — San Francisco, CA
Software Engineering Intern — *Aug 2025 – Dec 2025*

- Made Qwen2.5-VL and Qwen3-VL mid- and post-training 10x faster through multimodal sequence packing and data preparation, then optimized further with FlashAttention-3, Triton kernels, quantization, and FSDP.
- Built a sharded synthetic-data pipeline turning 800K CAD parts into 600M+ tokens for pretraining and fine-tuning; added Onshape support, doubling the addressable user base.
- Built async EC2 rollouts for model error correction and a CAD replay layer that caught errors tests missed.

Scale AI — Los Angeles, CA
Technical Intern — *Nov 2024 – Aug 2025*

- Designed judges and rubrics for RL evaluation of coding agents and built deterministic Docker environments; wrote 100+ expert competitive-programming solutions used as training data for frontier-lab LLMs.

HMC Music Retrieval Lab — Claremont, CA
Research Intern — *Jun 2025 – Aug 2025*

- Proposed self-distillation and DPO adaptations for steering MusicGen, tested across 1M+ audio clips.

Selected Projects

Event Detection / Kaggle | *First-Author Paper, Code, 5+ Gold Notebooks (20K+ Forks)* — 2023 – 2025

- Probability-density regression for event detection, adopted by nearly all top teams of 1,877; Silver Medal.

Graphics | *C++, OpenGL, Unity* — 2022 – Present

- Built a 4D game engine, a GPU slime simulation of 1M agents at 120 FPS, and raytracers for light and sound.

Technical Skills

Languages: Python, C++17/20, TypeScript, SQL, Bash, Mojo, C#
Machine Learning: PyTorch, FSDP, FlashAttention, Triton, vLLM/SGLang, CUDA/Mojo, Ray, quantization
Infrastructure: Linux, Docker, AWS, Azure, PostgreSQL, Cloudflare Durable Objects, GDB, CMake

Leadership and Distinctions

UCLA ACM AI | *Co-President* — Nov 2025 – Present

- Lead an 800-member community and 40-officer team; events draw thousands of participants; secured club sponsors.
- ICPC SoCal Regional Top 10 | USACO Platinum | Codeforces Expert | USAPhO Silver | Kaggle Top 1%