KAIQIN KONG

+1 (858) 214-8023 | k1kong@ucsd.edu | h1yori233.github.io | LinkedIn

Education

University of California San Diego — La Jolla, CA
Master of Science in Computer Science — *Sep 2025 – Jun 2027*

Zhejiang University — Hangzhou, Zhejiang
Bachelor of Engineering in Industrial Design — *Sep 2020 – Jun 2024*

Skills

Languages: Python, C, C++, C#, TypeScript

Machine Learning: PyTorch, CUDA, Triton, FastAPI

Tools: Linux, Git, Bash, LaTeX, CMake, Docker

Coursework: Data Structures, Operating Systems, Computer Graphics, GPU Programming and Architecture, Deep Learning Systems, Language Modeling from Scratch, Systems for LLMs and AI Agents

Open-Source Experiences

FastVideo

- Work on action-conditioned video generation.

Projects

Language Model from Scratch — *Jul 2025 – Aug 2025*

- Implemented a decoder-only transformer LM using **PyTorch** with a custom BPE-Tokenizer, MHA, and FFN. The model was trained from scratch on TinyStories to a **val loss 0.968**.
- Optimized GPU performance with a **Triton** implementation of **FlashAttention-2**, using IO-aware attention via tiling Q/K/V blocks and online-softmax for on-chip computation; achieving up to **2.2×** lower latency on short sequences and nearly **4×** on long sequences, evaluated on an RTX 4090.
- Benchmarked sequence lengths up to 65K with detailed profiling of forward/backward throughput & latency.

InnoWeaver - AI-Powered HCI Research Platform — *Nov 2024 – May 2025*

- Architected an HCI research platform that generates academic reports for design from custom input, powered by LLMs an async **FastAPI** backend integrated with **LangChain**.
- Developed a frontend webpage with user management, live chat, and a design gallery for exploring AI-generated insights from HCI papers, using **Next.js** and **Tailwind CSS**.
- Implemented data management with **MongoDB** for user interaction data and **Meilisearch** for RAG, and optimized the backend with **Redis**.

Needle - Mini-PyTorch Implementation from Scratch — *Jun 2025 – Jul 2025*

- Built a lightweight computational graph and reverse-mode **autodiff engine**.
- Implemented arithmetic, linear algebra, and reduction operations for core modules, following the design of **PyTorch**.
- Added a **CUDA**-based array engine for efficient tensor operations, achieving up to **115×** speedup on matrix multiplication and 3–6× on reductions.

Experience

Research Intern | Full-Stack & LLM Research — Hangzhou, Zhejiang
International Design Institute, Zhejiang University — *Jun 2024 – Jun 2025*

- Developed AI workflow powered by RAG over an HCI paper database to generate designs and evaluations.
- LLM fine-tuning, prompt engineering, and data processing pipeline development.

Teaching Assistant | Computer Game Programming — Hangzhou, Zhejiang
Zhejiang University — *Sep 2023 – Jan 2024*

- Guided student from initial proposal to final implementation of game.
- Led weekly labs and graded assignments for students.