

KAIQIN KONG

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

EDUCATION

University of California San Diego

Master of Science in Computer Science

La Jolla, CA

Sep 2025 – Jun 2027

Zhejiang University

Bachelor of Engineering in Industrial Design

Hangzhou, Zhejiang

Sep 2020 – Jun 2024

SKILLS

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

Web: React, Next.js, Tailwind CSS

Machine Learning: PyTorch, FastAPI, LangChain

GPU: CUDA, Triton, Nsight, WebGPU

Tools: Linux, Git, Bash, L^AT_EX, CMake, Docker, Unity, VS Code, Cursor, Claude Code, Figma

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

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.

CUDA Path Tracer

Mar 2025 – Apr 2025

- Built a GPU-accelerated Monte Carlo path tracer in **CUDA** with next-event estimation and depth-of-field, optimizing warp efficiency through ray sorting by material and stream compaction.
- Added an optimized **G-buffer** with **à-trous wavelet denoiser**, compressing normals via **octahedral encoding**.

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.