Kaiqin Kong
http://h1yori233.github.io/
Email : k1kong@ucsd.edu
Tel: +1 858-214-8023

Education

- **University of California San Diego** — La Jolla, CA, USA
  Master of Science in Computer Science and Engineering — *Sep 2025 – Jun 2027 (expected)*
- **Zhejiang University** — Hangzhou, Zhejiang, China
  Bachelor of Engineering in Industrial Design — *Sep 2020 – Jun 2024*

Projects

- **LM from Scratch**
  Tiny Language Model Implementation
  - **Architecture & Training**: Implemented a decoder-only transformer LM using PyTorch with custom Tokenizer, Multi-Head Causal Attention, and FFN, training from scratch to achieve functional performance on small-scale language tasks.
  - **GPU Optimization**: Developed custom Triton kernels for Flash Attention 2, reducing forward pass latency by up to 2.2x compared to PyTorch implementations while maintaining numerical accuracy.
  - **Performance Analysis**: Conducted extensive benchmarking across configurations up to 65K sequence length with detailed profiling of forward/backward pass performance.
- **InnoWeaver** — ZJU International Design Institute
  AI-Powered HCI Research Platform — *Nov 2024 – May 2025*
  - **Research Pipeline**: Architected AI-powered platform bridging HCI academia and practical design, employing LLMs to synthesize academic papers into actionable design concepts.
  - **Full-Stack Architecture**: Built scalable platform with async FastAPI backend, Next.js 14 frontend, and multi-database architecture using MongoDB, Redis, and Meilisearch for RAG capabilities.
  - **Knowledge Discovery**: Implemented intelligent card-based gallery interface enabling researchers to explore HCI innovations through AI-generated insights from academic literature.
- **Needle**
  Micro Deep Learning Framework Implementation
  - **Automatic Differentiation**: Implemented computational graph-based automatic differentiation system supporting gradient computation and backpropagation.
  - **Operator Implementation**: Developed various operators including basic arithmetic, linear algebra, and reduction operations for building fundamental modules.
  - **CUDA Backend**: Implemented NumPy-like array computation backend using CUDA for efficient tensor operations.

Experience

- **International Design Institute, Zhejiang University** — Hangzhou, China
  Research Assistant Intern — *Jun 2024 – Jun 2025*
  - **Multi-agent Research**: Developed LLM-driven autonomous agents for design ideation and evaluation.
- **Zhejiang University** — Hangzhou, China
  Teaching Assistant — Computer Game Programming — *Sep 2023 – Jan 2024*
  - **Course Support**: Led weekly labs and graded assignments for students.

Programming Skills

- **Languages**: Python, C++, TypeScript, C#
  Technologies: CUDA, PyTorch, Triton, FastAPI, Next.js

Relevant Coursework

- Fundamentals of Data Structures, Deep Learning, Operating System, Deep Learning System, Language Modeling from Scratch