Kunjun Li

Singapore | kunjun@u.nus.edu | kunjun-li.github.io | kunjun-li

Education

National University of Singapore Aug 2022 – Present
Bachelor of Engineering in Computer Engineering GPA: 4.8/5.0 (Top 5%)

University of Washington Mar 2025 – Jul 2025
Visiting Student Academic Standing: Dean's List

Research Interests

Efficient Deep Learning, with focus on optimizing training and inference of **LLMs, Diffusion and Multimodal Models**. I have been working on sparse attention, network pruning and efficient architectures. My work strives to achieve computational breakthroughs, making deep learning affordable and accessible to everyone, everywhere.

Publications

Memory-Efficient Visual Autoregressive Modeling with Scale-Aware KV Cache [PDF] [Project page]
Kunjun Li, Zigeng Chen, Cheng-Yen Yang, Jenq-Neng Hwang
Neural Information Processing Systems (NeurIPS 2025)

TinyFusion: Diffusion Transformers Learned Shallow [PDF] [Project page]
Gongfan Fang*, Kunjun Li*, Xinyin Ma, Xinchao Wang (Co-First Author)
Computer Vision and Pattern Recognition (CVPR 2025), **Highlighted Paper (3%)**

PixelGen: Rethinking Embedded Camera Systems for Mixed-Reality [arXiv] [Award page]
Kunjun Li, Manoj Gulati, Dhairya Shah, Steven Waskito, Shantanu Chakrabarty, Ambuj Varshney
Information Processing in Sensor Networks (IPSN 2024), **Best Demonstration Runner-Up**

Research Experiences

Princeton University, Research Intern 06/2025 – Present
Supervisor: Prof. Zhuang Liu | Goal: CVPR 2026

- Developed compute-optimal training strategies for efficient diffusion transformers with constrained resources and conducted in-depth analysis to compare pruning against random initialization under various settings.

Information Processing Lab, University of Washington, Research Intern 01/2025 – 05/2025
Supervisor: Prof. Jenq-Neng Hwang | Outcome: NeurIPS 2025

- Proposed ScaleKV, a KV cache compression framework that achieved 90% memory reduction (85 GB $\rightarrow$ 8.5GB) and substantial speedup for Visual Autoregressive (VAR) modeling while preserving pixel-level fidelity and facilitating the scaling of VAR models to ultra-high resolutions.

xML Lab, National University of Singapore, Research Intern 06/2024 – 01/2025
Supervisor: Prof. Xinchao Wang | Outcome: CVPR 2025 Highlight

- Proposed TinyFusion, an end-to-end learnable depth pruning framework that achieves comparable performance with 50% model parameters and depth while reducing pre-training costs to under 7%, generalizing effectively across various generative architectures (Diffusion, Flow, and AR).

National Computer Systems, Singapore, Edge-AI Engineer 07/2023 – 05/2024
Supervisor: Prof. Ambuj Varshney | Outcome: IPSN 2024 Best Demonstration Runner-Up

- Developed PixelGen, an innovative Embedded Camera System integrating Language Models and Diffusion Models, to generate High-Resolution RGB Images from monochrome images and sensor data.

Selected Honors

NeurIPS'25 Scholar Award 2025
Dean's List, University of Washington 2025
Best Undergraduate Researcher, National University of Singapore 2025
Innovation & Research Award, National University of Singapore 2025
Dean's List, Top 5% of Cohort, NUS School of Computing 2025
Second Place, 2025 SkiTB Visual Tracking Challenge 2025
IPSN'24 Best Demonstration Runner-Up, ACM/IEEE 2024

Course Projects

Multi-View Winter Sports Tracking with ReID-SAM [arXiv] Dec 2024 – Feb 2025
University of Washington, Supervisor: Prof. Jenq-Neng Hwang

- Proposed ReID-SAM framework that integrates Segment Anything Model 2 (SAM2) with person re-identification (ReID) and Kalman Filter for robust multi-view tracking in challenging winter sports scenarios.
- Achieved **Second Place** in 2025 SkiTB Visual Tracking Challenge.

Hardware-Software Co-Design for Intelligent VR System Aug 2025 – Nov 2025
National University of Singapore, Supervisor: Prof. Li-Shiuan Peh

- Co-designed neural network architectures and FPGA-based hardware accelerator for real-time action recognition in complex VR gaming environments.
- Integrated system achieved 95% accuracy on player evaluation while meeting real-time constraints.

High-Performance CUDA Kernel for Virus Signature Matching [Project page] Aug 2024 – Nov 2024
National University of Singapore, Course: Parallel Computing

- Developed high-performance CUDA kernels for large-scale virus signature matching with asynchronous memory transfers and multi-stream processing on Nvidia A100 and H100 GPUs.
- Utilized shared memory, texture cache, and memory coalescing to maximize memory bandwidth.
- Designed scalable solution for TB-scale datasets with memory management and kernel fusion.

Pipelined General-Purpose RISC-V Processor Design [Project page] Jan 2024 – Jun 2024
National University of Singapore, Course: Computer Architecture

- Designed and implemented 5-stage pipelined RISC-V CPU supporting RV32I instruction set with advanced hazard detection, forwarding and Karatsuba Algorithm optimized matrix multiplication.
- Implemented branch prediction and cache hierarchy to minimize pipeline stalls for throughput.

Technical Skills

Programming Languages: Python, C, C++, CUDA, Java, Shell Script, Assembly

Deep Learning Frameworks: PyTorch, SGLang, vLLM, DeepSeed, Megatron

Hardware AI: FPGA Accelerator, TinyML, Embeded Systems

Teaching

CS2309 CS Research Methodology [Course page] Aug 2025 – Nov 2025
National University of Singapore, Supervisor: Prof. Wynne Hsu

- Teaching assistant at NUS CS2309, advising undergraduates for exploring interest in computer science topics.
- Hosted the first lecture to teach students with advanced topics in efficient generative ai and model compression.

Hobbies

Sports: Volleyball (captain & gold medal), Basketball, Hiking, Table Tennis

Arts: Calligraphy, Seal Carving, Cooking, Electronic Music