Kunjun Li

Singapore | +65 89423624 | kunjun@u.nus.edu | kunjun-li | kunjun

Education

National University of Singapore Aug 2022 – Present
Bachelor of Engineering in Computer Engineering **GPA:** 4.8 / 5.0 **(Top 5%)**
Research Interest: Efficient Generative Models

Publications

Memory-Efficient Visual Autoregressive Modeling with Scale-Aware KV Cache
Kunjun Li, Zigeng Chen, Cheng-Yen Yang, Jenq-Neng Hwang
Under review, Arxiv 2505.19602
Lossless VAR KV Cache Compression | From 85 GB to 8.5 GB

TinyFusion: Diffusion Transformers Learned Shallow
Gongfan Fang*, **Kunjun Li***, Xinyin Ma, Xinchao Wang *(Equal-first author)*
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '25 Highlight)
CVPR Highlight (3%) | Tiny DiTs at 7% Training Costs | 2x Faster Inference

PixelGen: Rethinking Embedded Camera Systems for Mixed-Reality
Kunjun Li, Manoj Gulati, Dhairya Shah, Steven Waskito, Shantanu Chakrabarty, Ambuj Varshney
The 30th Annual International Conference on Mobile Computing and Networking (MobiCom '24)

Professional Experience

Undergraduate Researcher 07/2025 – Present
Working with Prof. Zhuang Liu

- Explored compute-optimal training strategies for efficient diffusion transformers.

UW Information Processing Lab Seattle
Undergraduate Researcher 01/2025 – 05/2025
Supervisor: Prof. Jenq-Neng Hwang

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

NUS xML Lab Singapore
Undergraduate Researcher 06/2024 – 01/2025
Supervisor: Prof. Xinchao Wang

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

NCS Group Singapore
Edge-AI Developer 07/2023 – 05/2024

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

Selected Honors

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

Project Experience

Parallel Virus Scanning with CUDA NUS

- Developed a CUDA program with asynchronous kernel and memory transfers on A100 and H100.

High-Performance RISC-V Processor Design NUS

- Designed a pipelined RISC-V CPU with Karatsuba Algorithm-optimized matrix multiplication.