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: 4.77/5.0

University of Washington Mar 2025 – Jul 2025
Visiting Student

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%)

On the Role of Network Pruning in Diffusion Transformers
Kunjun Li, Wenhao Chai, Taiming Lu, Yufeng Xu, Jianwen Xie, Jiachen Zhu, Mingjie Sun, Zhuang Liu
Under Review

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 Demo Runner-Up

Research Experiences

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

- Developed **compute-optimal** training strategies for efficient Diffusion Transformers under constrained resources and conducted in-depth analyses comparing pruning against random initialization across multiple settings.
- Built a unified pruning benchmark and codebase implementing magnitude, L2, random, taylor, hessian, diff-pruning, bk-sdm, and tinyfusion to systematically study the role of network pruning in DiT.
- Investigated robustness of pruned and scratch-trained Diffusion Transformers under noisy class-condition embeddings, showing that pruning primarily enhances stability to input condition perturbations.

University of Washington, Research Intern 12/2024 – 05/2025
Supervisor: Prof. Jenq-Neng Hwang | Outcome: NeurIPS 2025

- Proposed ScaleKV, a scale-aware KV cache compression framework for Visual Autoregressive (VAR) modeling that achieves **90%** KV cache memory reduction (85 GB $\rightarrow$ 8.5 GB) while preserving pixel-level fidelity.
- Enabled large-batch inference on a single GPU and facilitated scaling VAR models to ultra-high resolutions such as 4K, which would otherwise be limited by memory bottlenecks and inference latency.
- Validated its effectiveness across a wide range of compression ratios, supported by evaluations on multiple benchmarks and detailed ablations on cache budget allocation and attention-pattern analysis.

National University of Singapore, Research Intern 06/2024 – 12/2024
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 **(2$\times$ sampling speedup)** while reducing pre-training costs by **93%**.
- Conducted large-scale experiments on 100,000 models, showing that common pruning heuristics are unreliable and that final performance depends not on initial loss but on a model's ability to recover with fine-tuning.
- Generalized TinyFusion across various network architectures (Diffusion, Flow-Matching, and Autoregressive).

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

- Designed PixelGen, a low-power embedded camera system that fuses a monochrome image sensor with environmental and motion sensing for rich multimodal scene capture under strict power and bandwidth budgets.
- Implemented an edge pipeline coupling LLM with Diffusion model and ControlNet to generate high-resolution RGB images from low-resolution monochrome frames and sensor data across diverse visual styles.
- Built a VR prototype with **27**× lower uplink bandwidth compared to streaming native-resolution RGB video.

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

- Designed an FPGA-based hardware accelerator for real-time action recognition, incorporating asyncio-driven parallel inference to maximize throughput under VR latency constraints.
- Applied extensive data augmentation and model optimization techniques, enabling the integrated system to achieve 98% accuracy on player evaluation while sustaining real-time performance.

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.

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, vLLM, SGLang, DeepSeed, Megatron

Hardware AI: FPGA Accelerator, TinyML, Embedded Systems, Vitis HLS

Hobbies

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

Arts: Calligraphy, Seal Carving, Cooking, Electronic Music