Mingyu Lee

Klaus Advanced Computing Building, Room 3306, 266 Ferst Dr NW, Atlanta, GA 30332
+1-470-883-5460 | mlee864@gatech.edu | mingyulee864 | davidlee1229

RESEARCH VISION & INTERESTS

Research Vision: My research targets efficient AI systems through *cross-layer HW-SW co-design*, bridging the gap between algorithmic theory and hardware deployment. I focus on resolving workload bottlenecks to create scalable, real-world solutions under physical constraints.

Research Interests: Agentic AI, Efficient AI Systems, HW-SW Co-design, Model Compression (Quantization, Sparsity), Emerging Data Formats, and Hardware-efficient ML Architectures (SSM, Linear Attention).

EDUCATION

Georgia Institute of Technology, Atlanta, GA — Aug 2026 – Present
Ph.D. in Electrical and Computer Engineering
Advisor: Tushar Krishna

Korea Advanced Institute of Science and Technology, South Korea — Mar 2019 – Aug 2026
Bachelor of Science in Electrical Engineering

PUBLICATIONS (*): Equal contribution

[1] OuroMamba: A Data-Free Quantization Framework for Vision Mamba Models
Mingyu Lee*, Akshat Ramachandran*, Huan Xu, Souvik Kundu, and Tushar Krishna.
IEEE/CVF International Conference on Computer Vision (ICCV), Oct 2025. (*Accept. Rate: 24%*) [Paper]

[2] RECAP: Training-Free Compensation for Coarse Activation Channel Pruning in Compressed LLMs
Mingyu Lee*, Akshat Ramachandran*, and Tushar Krishna.
ML for Computer Architecture and Systems (MLArchSys) Workshop in ISCA, Jun 2025. [Paper]

[3] Polestar-Cache: Reconciling Parallel Decoding and Accuracy in Diffusion LLMs via Token Drift-Aware KV Cache Recalibration
Mingyu Lee*, Akshat Ramachandran*, Souvik Kundu, and Tushar Krishna.
Workshop on Latent & Implicit Thinking (LIT) in ICLR, Mar 2026. [Paper]

[4] Polestar: Drift-Aware Cache Calibration and Token Commitment for Efficient Inference of Diffusion LLMs
Mingyu Lee*, Akshat Ramachandran*, Souvik Kundu, and Tushar Krishna.
arXiv, May 2026. [Paper]

EXPERIENCE

Synergy Lab, Atlanta, GA — Aug 2026 – Present
Graduate Research Assistant (Advisor : Tushar Krishna)
- Conducting research on efficient algorithms, architectures and systems for accelerating Generative/Agentic AI applications.

Synergy Lab, Atlanta, GA — Jan 2025 – Jul 2026
Undergraduate Researcher (Advisor : Tushar Krishna)

- Developed a Data-free Quantization framework for Vision Mamba Models using **cyclic outlier detection** and **mixed-precision CUTLASS kernels**, achieving **2.36× speedup** with lossless accuracy; published as **co-first author at *ICCV 2025***. [Github]
- Proposed a Training-free Error Compensation for channel pruning, improving **34% BoolQ accuracy at 70% sparsity on LLaMA3-8B**; published as **co-first author at *MLArchSys @ ISCA 2025***.
- Taped out a **TSMC 65 nm, 5.56 TOPS/mW weight-stationary systolic GEMM accelerator**, supporting Memory, External, and LFSR-based Built-In Self-Test modes. [Github]

HyperAccel Corporation, Republic of Korea Jun 2024 – Dec 2024
FPGA Engineering Intern (Advisor : Joo-Young Kim)
- Collaborated with Samsung to optimize CXL-based LLM accelerator for **Qwen1.5-MoE** by **extending RISC-V ISA** with adaptive PEs, achieving **3× memory reduction** via FP8 quantization.
- Designed a lightweight data probe IP enabling selective runtime signal capture via AXI-Stream, reducing **BRAM usage by 45%** and **dynamic power by 15%**.
- Designed and optimized the system Soft-Reset protocol, resolving AXI-Stream/DMA deadlocks in multi-instance server environments and improving system reliability under asynchronous resets.

CAST Lab, Republic of Korea Jan 2024 – Jun 2024
Undergraduate Researcher (Advisor : Joo-Young Kim)
- Designed a **fine-grained DMA** in SystemVerilog using APB protocol for Single-Port SRAM access, implemented and verified on **Xilinx VCU118 FPGA** at 200 MHz with ILA probing.
- Developed a CNN accelerator based on **Output Stationary and Weight Stationary systolic array** architectures, optimized for on-chip data reuse and parallel computation.
- Developed a **Sparse ViT accelerator** using a novel **Compressed Sparse Diagonal (CSD)** format, achieving **2× speedup and compression** compared to conventional formats.

PRESENTATIONS

- *Polestar-Cache: Reconciling Parallel Decoding and Accuracy in Diffusion LLMs via Token Drift-Aware KV Cache Recalibration,* Poster, **LIT@ICLR 2026**, Rio de Janeiro, Brazil.
- *OuroMamba: A Data-Free Quantization Framework for Vision Mamba Models,* Poster, **ICCV 2025**, Honolulu, Hawai'i. [Poster] [Slides]
- *RECAP: Training-Free Compensation for Coarse Activation Channel Pruning in Compressed LLMs,* Oral, Poster, **MLArchSys@ISCA 2025**, Tokyo, Japan. [Poster] [Slides]
- *CLAMP-ViT: Contrastive Data-Free Learning for Adaptive Post-Training Quantization of ViTs,* Poster, **CRNCH Summit 2025**, Atlanta, GA.
- *Orion: LLM Serving Product at HyperAccel*, **Supercomputing 2024 (SC24)**, Atlanta, GA.

AWARDS, HONORS, AND SCHOLARSHIPS

- Georgia Tech ECE Fellowship 2026
- ICCV (Int. Conf. on Computer Vision) Broadening Participation Award 2025
- uArch Workshop/Grant in ISCA (Int. Symp. on Computer Architecture) 2025
- HPCA (Int. Symp. on High-Performance Computer Architecture) Student Travel Grant 2025
- Korea-U.S. Student Exchange Program Scholarship 2024

SKILLS, ACTIVITIES

Languages*:* C/C++, SystemVerilog, Python, Triton, CUDA, TCL, Makefile.
Frameworks, Tools*:* PyTorch, HuggingFace, LM Eval, Vivado/Vitis, Synopsys ICC2/PT, VCS, Verdi.
Relevant Coursework: VLSI Design: Theory to Tapeout, HW–SW Co-Design for ML Systems, Computer Architecture, Digital System Design, Digital Signal Processing.