Soomin Lee

lee5782@purdue.edu | linkedin.com/in/soominleeee

Education

Purdue University Jan. 2025-Present
M.S., Electrical and Computer Engineering West Lafayette, IN
GPA: **3.93/4.0** | Expected graduation: May 2027 | Advisor: Prof. Younghyun Kim
Relevant coursework: Hardware for Machine Learning; Digital System Design Automation; Embedded Systems; MOS VLSI.

Chung-Ang University Mar. 2018-Aug. 2022
B.S., Electrical and Electronics Engineering Seoul, Korea
GPA: **3.6/4.0**

Research Interests

Efficient AI systems and hardware–software co-design; memory-efficient inference and temporal reuse for streaming multimodal models; computer architecture, ML accelerators, and embedded inference; RTL/ASIC design and verification.

Publications & Manuscripts

† Equal contribution (co-first authors).

TaRo: Temporal Reuse of Top-Ranked KV Cache Entries for Efficient Streaming Video-Language Models
Soomin Lee, Omkar Prabhune, Arghadip Das, Vijay Raghunathan, and Younghyun Kim.
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2027. Under review.

EdgeVQA: Sustaining Frame and Token Rates in Streaming-Video LLMs on Edge Devices
JoonSeok Kim†, **Soomin Lee**†, Donghwan Yoon, Younghyun Kim, and Seokhyeong Kang.
Design Automation Conference (DAC), 2027. Under review.

HEED: Lightweight Early Workload Prediction for Energy-Efficient Streaming Audio Event Classification
Soomin Lee, Taewoong Yoon, Arghadip Das, Vijay Raghunathan, and Younghyun Kim.
Unpublished manuscript, 2026.

TESSERA: A Workload-Driven Simulation and Design-Space Exploration Framework for Heterogeneous NPUs
Arghadip Das†, Hoseok Kim†, **Soomin Lee**, Arnab Raha, Deepak A. Mathaikutty, and Vijay Raghunathan.
IEEE International Symposium on Workload Characterization (IISWC), 2026. Published.

Research Experience

Purdue University Aug. 2025-Present
Research Assistant, Networked and Embedded Intelligent Systems Lab West Lafayette, IN

TaRo | Temporal reuse for streaming VLM attention
First author

- Designed and implemented reuse of top-ranked KV candidate indices to reduce repeated full-history searches during streaming video prefill; retained the complete KV history and rescored candidates with current queries.
- Developed position-wise candidate banks, grouped-query-attention-aware refresh decisions, and bounded refresh intervals; combined reused candidates with the current frame's causal keys and used query similarity to trigger full-history search.
- Implemented fused Triton kernels for candidate gathering, rescoring, selection, and attention. Evaluated LLaVA-OneVision-0.5B and Qwen2.5-VL-3B on an NVIDIA H100 using OVO-Bench and StreamingBench subsets.
- Achieved **1.37-3.58× LLM-prefill speedups** over ReKV across evaluated models and context-length bins, with model/benchmark-dependent accuracy changes of -4.60 to +6.59 percentage points. On LLaVA, vision-encoding plus LLM-prefill latency fell by **30.6-31.0%**, with **3.45-6.59 percentage-point higher accuracy** than ReKV.
- Limited early-to-late prefill latency growth to **2.7-4.9%**, versus 38.9-99.1% for Exact Top-k, on fixed video sets within the evaluated context ranges (up to 32K/16K KV entries for LLaVA/Qwen).

EdgeVQA | Streaming-video LLM systems co-design
Co-first author; collaboration with POSTECH

- Co-designed joint frame prefill and in-sweep speculative verification to share target-model weight reads between video ingestion and answer generation under DRAM bandwidth constraints.
- Co-designed acceptance-paced frame chunk sizing using delivered-output lead, observed token yield, and estimated sweep time, with compute/frame-freshness limits and verification-only execution when answer delivery falls behind.
- Contributed to execution and hardware-support design for shared weight tiles, separate frame/answer attention contexts and rotary positions, full frame commitment despite rejected drafts, and partial answer-KV commitment.
- Implemented the C++ LLaVA-OneVision execution foundation for numerical validation, spanning vision encoding, multimodal projection, LLM prefill/decode, and explicit KV-cache handling.
- Joint system evaluation reduced modeled bandwidth for **10 FPS and 3 tokens/s by 58.6%** (55.5 to 23.0 GB/s) versus Sarathi-style piggybacking for LLaVA-OneVision-7B. In an eight-video real-model study at 25 GB/s, the system delivered **3.46 vs. 1.61 tokens/s** and **78% vs. 1% stall-free answers**.
- The project's ZCU104 FPGA evaluation showed **63.4% lower estimated area-time product** than Sarathi-style hardware; the verification and partial-KV-commit extension added 544 LUTs with no additional DSPs or BRAMs.

Research Experience (continued)

Purdue University | Research Assistant (continued)

HEED | Early workload prediction for energy-efficient streaming audio

First author

- Proposed lightweight workload prediction for streaming audio event classification by exploiting temporal similarity between adjacent audio tensors before classifier execution.
- Designed a two-stage exit-depth predictor combining pre-classifier L1 tensor-distance prediction with an in-classifier fallback to preserve accuracy during event transitions; integrated proactive dynamic voltage and frequency scaling (DVFS) under per-segment real-time deadlines.
- Deployed on an STM32U575ZI Cortex-M33 MCU and evaluated ESC-50-derived streams. Evaluation using measured MCU power and latency yielded **16.17-17.55% energy savings** versus EENet at matched accuracy, with **zero deadline misses** in the evaluated streams.
- Improved computation savings by **14.83-16.73 percentage points** at matched accuracy, with approximately **6 KiB** of additional runtime memory.

Deep learning intrusion detection for in-vehicle CAN traffic

Embedded inference and automotive security

- Developed a real-time intrusion detection pipeline using sliding-window CAN message representations and lightweight CNNs, including Reduced Inception-ResNet.
- Applied timing-aware features to detect spoofing and injection attacks; evaluated a pipeline with **over 99% accuracy**, **80 ms end-to-end latency**, and **4,000 messages/s throughput**; studied throughput/latency trade-offs for real-time inference.

TESSERA | Preliminary heterogeneous NPU exploration prototype Jan. 2026-May 2026

Course research project, Purdue University West Lafayette, IN

- Built an early design-space exploration framework to investigate underutilization in homogeneous AI accelerators and evaluate heterogeneous multi-tile NPU configurations.
- Developed synthesis-calibrated latency, energy, area, and energy-delay-area product (EDAP) models across MAC-array size, precision, SRAM capacity, and tile composition; compared configurations under iso-area and iso-latency constraints.
- Achieved up to **17.6% EDAP improvement** and **34% energy savings** over homogeneous baselines in the preliminary course-prototype evaluation.
- The final collaborative TESSERA framework identified an approximately $200\,\mathrm{mm}^2$ heterogeneous NPU with **46.91% mean modeled energy savings** over the best iso-area homogeneous baseline across 20 workloads. Coauthor of the published IISWC 2026 paper.

Seoul National University Jul. 2021-Dec. 2021

Undergraduate Researcher, Computer Architecture & Parallel Processing Lab Seoul, Korea

- Deployed and profiled CNN inference on NVIDIA Jetson and TPU-class hardware, examining end-to-end latency including I/O and preprocessing costs.
- Prototyped quantized FPGA inference for embedded object recognition; explored hardware–software co-design for vision models on FPGA and embedded accelerator platforms.

Industry Experience

LG Display, OLED Algorithm Team Jul. 2022-Aug. 2025

Research Engineer Seoul, Korea

- Developed hardware-oriented video-processing algorithms through microarchitecture definition, Verilog RTL implementation, FPGA validation, and ASIC transition/integration.
- Performed block-level verification and debugging using SystemVerilog testbenches, ModelSim, Verdi, and SpyGlass lint/rule checks.
- Received the LG Display 2023 4Q Vision Incentive for proposing an algorithmic idea adopted in production.

Hardware-aware compression for OLED compensation maps

Jan. 2023-Jun. 2024

- Redesigned vector-quantization/K-means-style compression for memory-bound compensation maps, replacing multiplier-heavy distance computation and iterative centroid updates with shift-based approximate distance and incremental codebook updates.
- Co-designed the memory dataflow to store compact codebooks in SRAM and index maps in DDR, enabling real-time per-frame reconstruction under external memory bandwidth constraints.
- Implemented and validated the encoder in Verilog RTL through simulation, synthesis, FPGA testing, and ASIC transition; achieved **8:1 data reduction**, reduced DDR requirements from **two channels to one**, and saved approximately **$3 per panel**.

Micro LED temperature compensation

Jun. 2024-Aug. 2024

- Modeled panel temperature characteristics and designed a color compensation algorithm with a power-optimized RTL implementation.
- Conducted FPGA testing and verified performance using Quartus.

AI-based OLED degradation prediction

Aug. 2024-Mar. 2025

- Developed AI models to predict OLED panel degradation and generated degradation parameter maps for residual-image compensation.
- Applied explainable AI techniques to analyze the factors contributing to OLED degradation and inform compensation modeling.

Content-driven thermal estimation for OLED panels

Feb. 2025-Aug. 2025

- Developed a pixel-wise thermal prediction algorithm based on a 2D heat-diffusion model and video-derived current maps.
- Optimized the algorithm for real-time hardware implementation, achieving **15% improvement in temperature prediction accuracy** and completing RTL handoff before tape-out.

Academic Projects

FPGA-based sign language gesture classification Jan. 2025-May 2025

Course project, Purdue University West Lafayette, IN

- Developed and augmented an American Sign Language gesture dataset and trained compact CNNs for hardware deployment.
- Applied fixed-point quantization and pruning, and mapped weights to ROM initialization formats.
- Designed a Verilog RTL CNN accelerator and validated inference on a Basys3 Artix-7 FPGA with an ESP32 preprocessing pipeline.

Presentations

ISLPED Women in Technology Aug. 2026

Lightning talk and poster, Evanston, IL.
Exploiting Temporal Similarity for Energy-Efficient Audio Event Classification.

DAC Young Fellow Jul. 2026

Poster presentation, Long Beach, CA.
Exploiting Temporal Similarity for Energy-Efficient Audio Event Classification.

Midwest Machine Learning Symposium Jun. 2026

Poster presentation, West Lafayette, IN.
Exploiting Temporal Similarity for Energy-Efficient Inference on MCUs.

Purdue AI Research Showcase Apr. 2026

Poster presentation, West Lafayette, IN.
Exploiting Temporal Similarity for Energy-Efficient Audio Event Classification.

Awards & Scholarships

LG Electronics Global Industry-Academic Scholarship Spring 2026-Spring 2027

DAC Young Fellow, Design Automation Conference Jul. 2026

LG Display 2023 4Q Vision Incentive Dec. 2024

LG Display Circuit & Algorithm Research Division

Technical Skills

Programming & HDL: Python, C/C++, Verilog, SystemVerilog, MATLAB.

ML & GPU Systems: PyTorch, TensorFlow, Triton; GPU profiling, KV-cache management, quantization, embedded inference.

RTL & Verification: ModelSim, Verdi, SpyGlass; SystemVerilog testbenches, waveform debugging, lint/rule checks, regression support, synthesis, block-level verification.

FPGA & Design Tools: Vivado, Vitis HLS, Quartus, SignalTap, PSpice, Virtuoso; fixed-point implementation and FPGA validation.

Digital Design: Microarchitecture, pipelining, finite-state machines, datapath design, timing constraints and closure, memory/bandwidth optimization, and power, performance, and area trade-offs.

Development Environment: Linux, Git.