

Hyunwoo Oh

🏠 Irvine, CA +1 (949) 571-0953 ✉️ hyunwooo@uci.edu 🌐 hyun-woo-oh.github.io 🔗 linkedin.com/in/hyunwooooh

EDUCATION

Ph.D. in Computer Science

University of California, Irvine

Sep. 2024 – Jun. 2029 (expected)

Irvine, CA

M.S. in Electronic Engineering

Seoul National University of Science and Technology

2023

Seoul, Korea

B.S. in Electronic Engineering

Seoul National University of Science and Technology

2021

Seoul, Korea

TECHNICAL SKILLS

Research Interests: Hardware/Software Co-design for Machine Learning, Computer Architecture, On-Device AI

Programming: C/C++, Python, SIMD (x86 AVX2), DSP (TI C66x)

HDLs & Generators: Verilog/SystemVerilog, Chisel

ML & Frameworks: PyTorch, NumPy, CUDA

Arch/EDA Tools: gem5, Vivado/Vitis, Quartus, Synopsys (DC, VCS, Verdi, PrimeTime, Formality, ICC2), Cadence Genus

RESEARCH EXPERIENCE

Quantization-Aware SW/HW Co-design for CPU-Only LLM Inference | Python, C++, SIMD Intrinsics, gem5, SystemVerilog Mar. 2025 – Present

- Co-designed ternary LLM GEMM/GEMV algorithms, a ternary ISA extension, AVX2 kernels, and SIMD-unit RTL for LLM inference (e.g., BitNet-b1.58).
- Repurposed SIMD register files as ternary LUTs to remove LUT DRAM traffic, achieving **5.6–24.5× lower GEMM latency**, **1.1–86.2× higher GEMV throughput**, and **2.5–4.9× better energy** than Jetson AGX Orin with only **3.2% power / 1.4% area overhead** for the SIMD unit redesign.
- Developed an intra-layer mixed-precision quantization and compilation framework with activation-aware per-channel 2/3/4/8/16-bit weights and AVX2-friendly kernels.
- Achieved effective **3–6 b** scaling on Falcon-H1-3B, Llama2-13B, and Qwen3-32B, reducing perplexity by **2.4–7.8% vs AWQ** and **10.9–17.0% vs GPTQ** at similar latency/energy.
- Publications:** DATE 2026 (accepted); DAC 2026 (under review).

ASIC Accelerator for Emerging Object Detection Models | PyTorch, CUDA, Chisel, SystemVerilog, Verilator, Synopsys DC Sep. 2024 – Present

- Designed QUILL, a deformable-attention accelerator for DETR-style detection, co-optimizing attention schedules with on-chip memory.
- Achieved up to **7.3× throughput** and **47× energy efficiency vs RTX 4090** at comparable accuracy by reordering queries and prefetching regions to make sparse deformable-attention accesses cache-local in a 28nm fused core.
- Designed a **128×128 8-bit systolic array** with flexible weight/input/output-stationary dataflows and on-chip scheduling, removing SRAM and host tensor-reordering overheads.
- Achieved **30.3 TOPS @ 1.85 GHz (28nm)** with up to **3.5× speedup** and **40% lower energy** vs GPU baselines.
- Publications:** DATE 2026 (accepted); DAC 2025.

FPGA Accelerator for Multimodal AI Workloads | Vivado, PyTorch, Chisel, SystemVerilog Dec. 2024 – Present

- Designed an FPGA accelerator and compiler for multimodal AI (ViT, GNN, CNN, NLP) under tight latency and resource budgets.
- Implemented mode-switchable compute engines, scalable top-k token pruning, and dependency-aware offloading to support diverse workloads without bitstream reconfiguration.
- Developed a Chisel-based RTL generator targeting Xilinx Alveo U50 and ZCU104, achieving up to **22.6× lower latency vs RTX 4090** and **6.9× vs Jetson Orin Nano**, outperforming prior FPGA accelerators.
- Publications:** FCCM 2025; DATE 2026 (accepted).

Additional Research (prior work) Mar. 2021 – Dec. 2022

- Streaming FPGA sorter generator with up to **68% lower latency** and **1.26× higher throughput** vs prior designs (IEEE TCAS-II).
- RISC processor with posit arithmetic and modified GCC toolchain, up to **60× speedup** vs software posit and **11% reduced LUT (ISLPED 2023)**.
- Participated in **7 ASIC tape-outs** (Samsung 65–28nm, TSMC 180nm) with roles across RTL design/verification, back-end, PCB bring-up, and test.

INDUSTRY EXPERIENCE

FPGA/Embedded SW Engineer (Full-time), Core H/W Team, Hanwha Systems, South Korea Jan. 2023 – Aug. 2024

High-Resolution Thermal Image Processor for 1280×1024 IR Cameras | C/C++, Verilog, Vivado, Zynq Ultrascale+ MPSoC Jan. 2023 – Aug. 2024

- Built an end-to-end thermal image pipeline (NUC + CLAHE) on Zynq SoC FPGAs with PL accelerators and ARM cores sharing DDR via AXI.

- Achieved real-time 640×480 @ 60 FPS on XC7Z020 using <40% LUT/BRAM/DSP; extended to Zynq Ultrascale+ MPSoC for 1280×1024 @ 60 FPS in a fully on-device sensor-to-display pipeline.
- Work published at **Euromicro DSD 2023** and used in internal high-resolution prototypes.

Real-Time Thermal Imaging on Heterogeneous SoC (TI TDA3x) | C/C++, DSP programming (C66x), SIMD/VLIW, RTOS Jan. 2023 – Aug. 2024

- Developed NUC, contrast enhancement, and noise filtering on Cortex-M4, dual C66x DSPs, and vision engine under memory/power limits.
- Hand-optimized fixed-point SIMD/VLIW kernels and cache-aware task scheduling to reach 640×480 @ 60 FPS with <2.2 W system power and 57.5% peak core load.
- **Publications: RTCSA 2024.**

SELECTED PUBLICATIONS

Quantization-Aware SW/HW Co-design for CPU-Only LLM Inference

- **H. Oh et al.**, "T-SAR: A Full-Stack Co-design for CPU-Only Ternary LLM Inference via In-Place SIMD ALU Reorganization", Design, Automation & Test in Europe Conference (**DATE'26**, accepted).
- **H. Oh et al.**, "PolyQ: Codesigning End-to-End Quantization Framework for Scalable CPU-Only LLM Inference", DAC'26 under review.

ASIC Accelerator for Emerging Object Detection Models

- **H. Oh et al.**, "QUILL: An Algorithm-Architecture Co-Design for Cache-Local Deformable Attention", Design, Automation & Test in Europe Conference (**DATE'26**, accepted).
- S. Jeong, H. Barkam, **H. Oh et al.**, "iTaskSense: Task-Oriented Object Detection in Resource-Constrained Environments", IEEE/ACM Design Automation Conference (**DAC'25**).

FPGA Accelerator for Multimodal AI Workloads

- **H. Oh et al.**, "RIFT: A Single-Bitstream, Runtime-Adaptive FPGA-Based Accelerator for Multimodal AI", Design, Automation & Test in Europe Conference (**DATE'26**, accepted).
- **H. Oh et al.**, "A Multimodal AI Acceleration with Dynamic Pruning and Run-time Configuration", IEEE International Symposium on Field-Programmable Custom Computing Machines (**FCCM'25**).
- H. Chen, Y. Ni, W. Huang, **H. Oh et al.**, "LVLN_CSP: Accelerating Large Vision Language Models via Clustering, Scattering, and Pruning for Reasoning Segmentation", ACM International Conference on Multimedia (**MM'25**).
- H. Chen, Y. Ni, W. Huang, **H. Oh et al.**, "Revisiting Reconfigurable Acceleration of Vision Transformer with Patch Pruning", IEEE/ACM International Conference on Low Power Electronics and Design (**ISLPED'25**).

Real-Time Thermal Imaging on Heterogeneous SoC

- **H. Oh et al.**, "A Compact Real-Time Thermal Imaging System Based on Heterogeneous System-on-Chip", IEEE International Conference on Embedded and Real-Time Computing Systems (**RTCSA'24**).

Additional Research (prior work)

- **H. Oh et al.**, "DL-Sort: A Hybrid Approach to Scalable Hardware-Accelerated Fully-Streaming Sorting", IEEE Transactions on Circuits and Systems II (**IEEE TCAS-II**).
- **H. Oh et al.**, "RF2P: A Lightweight RISC Processor Optimized for Rapid Migration from IEEE-754 to Posit", IEEE/ACM International Conference on Low Power Electronics and Design (**ISLPED'23**).

CHIP TAPE-OUTS (SELECTED, 4 OF 7 TOTAL)

32-bit Processor with Posit Arithmetic Coprocessor for Embedded Systems

Jul. 2021

- Technology: Samsung 28nm RFCMOS (1-poly 8-metal)
- Role: RTL Design & Verification, ASIC Design Front-end/Back-end, Firmware, PCB Design & Chip Test



In-Vehicle Network Processor based on Cortex-M0

Mar. 2022

- Technology: TSMC 180nm RFCMOS (1-poly 6-metal)
- Role: System Verification SW, RTL Verification, Pre/Post-Layout Simulation



A RISC-V Processor Supporting AMBA AXI Protocol for Embedded Systems

Jul. 2022

- Technology: Samsung 28nm RFCMOS (1-poly 8-metal)
- Role: RTL Verification



Implementation of Lossless Decompression Accelerator Based on Inflate Algorithm

Sep. 2020

- Technology: Samsung 65nm RFCMOS (1-poly 8-metal)
- Role: System Verification SW, PCB Design & Chip Test

