SUNGWOONG YUNE

291, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
+82 10 5488 4338 ◇ imwooong@kaist.ac.kr ◇ Google Scholar ◇ LinkedIn

EDUCATION

Ph.D. Candidate in Graduate School of AI Semiconductor March 2025 - February 2029 (Expected)
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

M.S. in Electrical Engineering March 2023 - February 2025
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea GPA: 3.48/4.3

B.S. in Electrical Engineering March 2017 - February 2023
Korea University, Seoul, Korea GPA: 4.11/4.5 (Major GPA: 4.31/4.5)

RESEARCH INTERESTS

AI Hardware Architecture & Digital Chip Design
Design and silicon tape-out of domain-specific accelerators for Generative AI (Vision Autoregressive, Text-to-Motion) and Computer Vision.

Hardware/Software Co-Design & Workload Optimization
Hardware acceleration and algorithm optimizations for advanced AI workloads, including Diffusion Models and Secure AI (Fully Homomorphic Encryption).

RESEARCH PUBLICATION (* EQUAL CONTRIBUTION)

[17] H. Cho*, A. Putra*, **S. Yune**, C. M. Duong, J. Jeon, J.-Y. Kim, "*A 435.86 M-GOPS/W Bootstrappable Multi-Scheme FHE Accelerator with on-Chip Data Generation for Privacy-Preserving Computation*," IEEE Journal of Solid-State Circuits (JSSC) (Under Review) 2026

[16] T. Suh, M. Park, J.-Y. Kim, **S. Yune**, H. Kim, J.-Y. Kim, "*EssenSplat: Eliminating Tile- and Pixel-Level Redundancy in Gaussian-based Rendering Pipeline*," IEEE Transactions on Emerging Topics in Computing (TETC) (Under Review) 2026

[15] **S. Yune***, J. Jeon*, A. Putra, H. Cho, C. D. Manh, Y. Kim, T. Suh, J.-Y. Kim, J.-Y. Kim, "*Trident: A 28nm 1.83-to-9.74 pJ/bit Multi-Parameter FHE Processor with Unified Constant Geometry Datapath for Privacy Preserving Client-Server Computing*," 2026 IEEE Asian Solid-State Circuits Conference (A-SSCC) (Highlight Paper) 2026

[14] T. Suh, J. Lim, J.-Y. Kim, M. Park, H. Choi, J. Min, **S. Yune**, J.-Y. Kim, "*GAREN: A 890.9 FPS, 412 iter/s 3D Gaussian Splatting Rendering and Modeling Accelerator with Hybrid Sorting, Out of Range Filtering, and Gradient Skipping*," 2026 IEEE Asian Solid-State Circuits Conference (A-SSCC) (Highlight Paper) 2026

[13] J. Jeon, W. Shin, **S. Yune**, J.-Y. Kim, "*NetFED: Algorithm/Hardware Co-design for Multi-Key Privacy-Preserving Federated Learning Aggregation*," 2026 59th IEEE/ACM International Symposium on Microarchitecture (MICRO) 2026

[12] **S. Yune***, J. Yoon*, H. Lee, A. Putra, J. Heo, J.-Y. Kim, "*A 11.0 TOPS/W Diffusion Accelerator with Temporal Data Reuse for Real-Time Text-to-Motion Generation*," IEEE Journal of Solid-State Circuits (JSSC) 2026 [Link]

[11] **S. Yune***, S. Jeong*, J.-Y. Kim, "*FastSTAR: Spatiotemporal Token Pruning for Efficient Autoregressive Video Synthesis*," 2026 19th European Conference on Computer Vision (ECCV) 2026 [Link]

[10] A. Putra*, H. Cho*, **S. Yune**, C. M. Duong, J. Jeon, J.-Y. Kim, "*OmniCrypt: A 435.86 M-GOPS/W Bootstrappable Multi-Scheme FHE Accelerator with on-Chip Data Generation for Privacy-Preserving Computation*," 2026 IEEE International Solid-State Circuits Conference (ISSCC) 2026 [Link]

[9] J. Yoon*, **S. Yune***, H. Lee, A. Putra, J. Heo, J.-Y. Kim, "*MoDiff: A 11.0 TOPS/W Diffusion Accelerator with Temporal Data Reuse for Real-Time Text-to-Motion Generation*," 2025 IEEE Asian Solid-State Circuits Conference (A-SSCC) 2025 [Link]

[8] J. Heo, A. Putra, J. Yoon, **S. Yune**, H. Lee, J. H. Kim, J.-Y. Kim, "*Exion: Exploiting inter-and intra-iteration output sparsity for diffusion models*," 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA) 2025 [Link]

[7] J. Heo, A. Putra, **S. Yune**, J. Yoon, H. Lee, J. H. Kim, J.-Y. Kim, "*A 57.3-fps 12.8 TFLOPS/W Text-to-Motion Processor With Inter-Iteration Output Sparsity and Inter-Frame Joint Similarity*," IEEE Journal of Solid-State Circuits (JSSC) 2025 [Link]

[6] **S. Yune**, H. Lee, A. Putra, H. Cho, C. D. Manh, J. Jeon, J.-Y. Kim, "*ABC-FHE: A Resource-Efficient Accelerator Enabling Bootstrappable Parameters for Client-Side Fully Homomorphic Encryption*," 2025 62nd ACM/IEEE Design Automation Conference (DAC) 2025 [Link]

[5] D. Kim, J.-Y. Kim, H. Cho, S. Yoo, S. Lee, **S. Yune**, S. Yang, H. Jeong, K. Park, "*DPIM: A 2T1C eDRAM Transformer-in-Memory Chip With Sparsity-Aware Quantization and Heterogeneous Dense-Sparse Core*," IEEE Journal of Solid-State Circuits (JSSC) 2025 [Link]

[4] J. Heo, A. Putra, **S. Yune**, J. Yoon, H. Lee, J. Kim, J.-Y. Kim, "*HuMoniX: A 57.3 fps 12.8 TFLOPS/W Text-to-Motion Processor with Inter-Iteration Output Sparsity and Inter-Frame Joint Similarity*," 2025 IEEE International Solid-State Circuits Conference (ISSCC) 2025 [Link]

[3] D. Kim, J.-Y. Kim, H. Cho, S. Yoo, S. Lee, **S. Yune**, H. Jeong, K. Park, K.-S. Lee, "*DPIM: A 19.36 TOPS/W 2T1C eDRAM transformer-in-memory chip with sparsity-aware quantization and heterogeneous dense-sparse core*," 2024 IEEE European Solid-State Electronics Research Conference (ESSERC) 2024 [Link]

[2] J. Yang, **S. Yune**, S. Lim, D. Kim, J.-Y. Kim, "*Acane: An efficient fpga-based embedded vision platform with accumulation-as-convolution packing for autonomous mobile robots*," 2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC) 2024 [Link]

[1] J. Heo, Y. Shin, S. Choi, **S. Yune**, J. H. Kim, H. Sung, Y. Kwon, J.-Y. Kim, "*Primo: A full-stack processing-in-dram emulation framework for machine learning workloads*," 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD) 2023 [Link]

RESEARCH EXPERIENCE

Graduate Research Assistant (Ph.D. Candidate), CAST Lab, KAIST March 2023 - Present
Advisor: Prof. Joo-Young Kim

[Sub-lead] Efficient LLM Acceleration with Gated-Attention Mechanism Jan. 2026 - Present

- Conducting in-depth software-level profiling and workload analysis of Large Language Models (LLMs) incorporating Gated-Attention mechanisms.
- Investigating hardware-friendly optimization strategies to address memory bottlenecks and computational complexity in state-of-the-art LLM architectures.

[Project Lead] 28nm LPP Visual Autoregressive (VAR) Accelerator Mar. 2025 - Present

- Serving as the Tape-out Lead for a high-performance image/video generation chip using Samsung 28nm LPP process.

- Architecting the hardware by analyzing software models and proposing novel features, followed by complete RTL-level implementation and functional verification.
- Managing the back-end handoff and post-silicon phases, executing logic synthesis and formality checks, through to post-DB-out packaging and PCB-level validation (DB-out: Jan. 2026).
- Spearheading the architecture design for VAR acceleration, extending the core algorithmic insights from the tape-out into a GPU-accelerated framework for video synthesis (Ref: [11]).

[Full-Chip Tape-out] 28nm FD-SOI Multi-Scheme FHE Accelerator (OmniCrypt) Mar. 2024 - Feb. 2025

- Developed a bootstrappable FHE accelerator in Samsung 28nm FD-SOI, focusing on server-side privacy-preserving computation. (Ref: [10]
- Led the hardware/software co-design from algorithmic verification using Go (Golang) to RTL implementation and functional verification.
- Extended research into resource-efficient parameters for client devices (Ref: [6]).

[Full-Chip Tape-out] 14nm Text-to-Motion Diffusion Processor (HuMoniX) Jun. 2023 - Feb. 2025

- Designed and verified a dedicated compute core for Human Skin Rendering acceleration using Samsung 14nm technology.
- Advanced the architecture (MoDiff) to achieve state-of-the-art energy efficiency (11.0 TOPS/W) (Ref: [4, 7, 8, 9, 12]).

Memory-Centric & FPGA-based Hardware Exploration 2023 - 2024

- DPIM: Designed eDRAM layouts via Cadence Virtuoso; conducted post-layout simulations to resolve power spikes and ensure PVT stability (Ref: [3, 5]).
- ACANE: Proposed and verified a DSP-packing technique for FPGA-based CNN acceleration using Python (Ref: [2]).
- PRIMO: Developed a cycle-accurate DRAM-PIM emulation framework by modifying DRAMsim (C++) to facilitate hardware-software co-analysis (Ref: [1]).

Undergraduate Research Intern, IMPACT Lab, KAIST Dec. 2021 - Mar. 2022

- Designed and simulated low-power analog amplifiers for cochlear implants using Cadence Virtuoso.

INTERNSHIPS

Memory Design Intern, Server Memory Design Team,
Samsung Electronics (Device Solutions - Memory), Hwaseong, South Korea July 2022 - August 2022

- Analyzed the fundamental architecture of 1T1C (1 Transistor - 1 Capacitor) DRAM cells, focusing on the intricate design and operation of bitlines (BL) and wordlines (WL).
- Acquired practical understanding of industry-standard DRAM circuit design methodologies and commercial-scale server memory workflows.

TEACHING EXPERIENCE

Teaching Assistant for EE303, "Digital System Design", KAIST Sept. 2024 - Dec. 2024

Teaching Assistant for SS301, "Basic Semiconductor Electronics Lab", KAIST Mar. 2024 - Jun. 2024

Teaching Assistant for EE305, "Introduction to Electronics Design Lab", KAIST Sept. 2023 - Dec. 2023

PATENTS

[1] **S. Yune**, "*Resource-Efficient Accelerator Enabling Bootstrappable Parameters for Client-Side Fully Homomorphic Encryption*," Republic of Korea Patent App. No. 10-2025-0143441 2025

SKILLS

Languages & Environments: C/C++, Go(lang), Python, Verilog, SystemVerilog, MATLAB, Linux OS, Server Management

HW Tools: ModelSim, Cadence (Virtuoso, PSpice), Synopsys (Memory Compiler, Library Compiler, Design Compiler, VCS, DVE, Formality, PrimeTime, PrimePower, StarRC, ICC2)

HONORS & AWARDS

Deputy Prime Minister's Award (Minister of Science and ICT), 2026 AI Semiconductor Excellence Competition 2026

Semester High Honors, School of Engineering, Korea University 2018, 2021, 2022

INVITED TALKS & PRESENTATIONS

Invited Speaker, AI Semiconductor Graduate School Special Session July 2026
The Semiconductor Engineering Society of Korea (ISEK) Summer Conference, Busan, Korea

Keynote/Awardee Presentation, AI Semiconductor Career Insight Day June 2026
Hosted by Ministry of Science and ICT & IITP, Seoul, Korea