Hyesung Jeon

✉Email 🏠Homepage LinkedIn Github Google Scholar (Last updated: Mar. 2026)

Research Interests

Keywords: Efficient AI Serving, Agentic AI Systems, KV Cache Compression, PEFT, Model compression

My primary research focus is building efficient algorithms and serving systems for generative models, such as large language models, large multi-modal models, and diffusion models. My work mainly spans the post-training stack, from quantization-aware fine-tuning to KV cache management of agentic LLMs. I am drawn to the emerging efficiency challenges with long-context inference and agentic workloads.

In particular, my research interests lie in:

- Model compression (quantization and pruning) and parameter-efficient fine-tuning
- KV cache compression and efficient attention mechanisms
- Efficient inference systems for LLM-based agents
- Hardware-software co-design for low-precision training and inference

Education

Seoul National University

Ph.D. Student in Electrical and Computer Engineering Mar. 2023 – (Present)
Advisor: Prof. Jae-Joon Kim

B.S. in Electrical and Computer Engineering Mar. 2019 – Feb. 2023
Graduated Summa Cum Laude (GPA 4.04/4.30)

Gyeonggi Science High School

Math and Science Specialized High School Mar. 2016 – Feb. 2019

Work Experiences

MangoBoost Apr. 2022 – Jun. 2022

Internship
RDMA System Architecture Design for Data Processing Unit
Mentor: Prof. Jangwoo Kim

SK Hynix Solution Center

Internship Jun 2021 – Aug. 2021
Deep Learning Network Design for SoC-NAND Validation Indicator
The Encouragement Prize in the Internship Workshop
Mentor: Dr. Yong Lee

Seoul National University
Graduate School of Convergence Science and Technology Jan. 2021 – Feb. 2021

Student Researcher
Deep Learning Network Design for Nano-optical Layer Architecture
Mentor: Prof. Changsoon Kim

Department of System Semiconductor Engineering For AI Sep. 2020 – Feb. 2021
Student Researcher
End-to-End Development of an English Subtitle Generation Web Service
End-to-End Development of a Visitor Counter Application with a Face Detection Model
Mentor: Dr. Seokkyu Kwon

Publications

[7] **Hyesung Jeon**, Hyeongju Ha, Jae-Joon Kim, "LRAgent: Efficient KV Cache Sharing for Multi-LoRA LLM Agents", International Conference on Machine Learning (**ICML**), Jul. 2026.

[6] **Hyesung Jeon***, Seojune Lee*, Beomseok Kang, Yulhwa Kim, Jae-Joon Kim, "QWHA: Quantization-Aware Walsh-Hadamard Adaptation for Parameter-Efficient Fine-Tuning on Large Language Models", The Fourteenth International Conference on Learning Representations (**ICLR**), April. 2026.

[5] Jehun Lee, **Hyesung Jeon**, Juchan Lee, Jae-Joon Kim, "PRESTE: Preserving Tiny Exponent Precision for Efficient Sub-8-bit LLM Inference and Fine-Tuning", preprint, Jan. 2026.

[4] **Hyesung Jeon**, Yulhwa Kim, Jae-Joon Kim, "L4Q: Parameter Efficient Quantization-Aware Fine-Tuning on Large Language Models", The 63rd Annual Meeting of the Association for Computational Linguistics (**ACL**), Jul. 2025.

[3] Yulhwa Kim, Dongwon Jo, **Hyesung Jeon**, Taesu Kim, Daehyun Ahn, Hyungjun Kim, Jae-Joon Kim, "Leveraging Early-Stage Robustness in Diffusion Models for Efficient and High-Quality Image Synthesis", Conference on Neural Information Processing Systems (**NeurIPS**), Dec. 2023.

[2] Jiwoong Choi, Minkyu Kim, Daehyun Ahn, Taesu Kim, Yulhwa Kim, Dongwon Jo, **Hyesung Jeon**, Jae-Joon kim, Hyungjun Kim, "Squeezing Large-Scaling Diffusion Models for Mobile", International Conference on Machine Learning (**ICML**) Workshop on Challenges of Deploying Generative AI, Jul. 2023.

[1] **Hyesung Jeon**, Jae-Joon Kim, "AND-Net Based Multi Precision Neural Network Accelerator Design", B.S. Graduate Paper, Feb. 2023.

Honors and Awards

Qualcomm Innovation Fellowship Korea (Winner) 2025
L4Q: Parameter Efficient Quantization-Aware Fine-Tuning on Large Language Models

Samsung SAIT Computer Engineering Challenge (3rd Prize) 2023
vLLM-based LLM Inference Acceleration on Multi-GPU Systems

SNU Social Responsibility+ Competition (The Encouragement Prize) 2020
Arduino-based Storage Assistance Agent Embedded System Design

SNU Electrical Circuit Design Workshop (The Excellence Prize) 2020

Korea Presidential Science Scholarship (Full Tuition Scholarship) 2019 – 2023

Intel International Science and Engineering Fair (Finalist) 2018

Invited Talks

Efficient AI Meetup Korea (Speaker) 2024, 2026

Skills

Language	English, Korean
Programming Language	Python, C/C++
Deep Learning Frameworks	PyTorch, vLLM, SGLang, CUDA, Triton, Verilog, HLS

Leadership & Volunteering

Reviewer & Committee
ICLR 2026, NeurIPS 2026, ARR 2026 (Reviewer)

SNU Tomorrow's Engineers Membership (Chairman) 2021-2023
Organized and participated undergraduate mentorship and scholarship programs.

References

Prof. Jae-Joon Kim
Department of Electrical and Computer Engineering
Seoul National University
E-mail: kimjaejoon@snu.ac.kr