

DONG-WON LEE

 [linkedin.com/in/dong-won-lee](https://www.linkedin.com/in/dong-won-lee) |  dlehddnjs245@khu.ac.kr

EDUCATION

Kyung Hee University

B.S. in Software Convergence | GPA: 3.91 / 4.50

Mar 2021 – Present, Expected Aug 2027

EXPERIENCE

EMIL Lab, Kyung Hee University

Undergraduate Research Intern

Jul 2026 – Present

- **Research Topic:** Memory-efficient KV cache quantization for reasoning LLMs (LRMs) (in progress).
- Studying how to store the KV cache at low precision during long chain-of-thought decoding, where the cache grows with the length of the reasoning itself.
- Aiming to keep final-answer reasoning accuracy close to the full-precision KV cache while reducing KV memory, building on public KV cache quantization work.

ROBROS

Software Engineer Internship

Dec 2025 – Mar 2026

- Built real-time multi-camera streaming for a humanoid robot SDK over CycloneDDS, from an on-robot NVIDIA Jetson through a bridge to client PCs across a mixed LAN/Wi-Fi network (Python, partly C++).
- Set up cross-network routing (IP forwarding, NAT), tuned UDP socket buffers, and shipped it as a systemd service with a Cython-compiled core; validated on the physical robot.

Kyung Hee University

Biometric Data Analysis Research (Research Assistant)

Jul 2025 – Dec 2025

- Social data analysis of ChatGPT use in elementary-school classrooms: built a multimodal classroom-interaction dataset, defined annotation criteria, checked consistency, and ran hypothesis-driven analysis (Python).

Photo Studio Booking Platform

Student Founder & Product Manager

Mar 2024 – Jul 2025

- Founded a student-run startup building a photo studio and photographer search/booking app; led a 5-member team across product, design, and performance marketing.
- Built and shipped the MVP as an iOS app plus an admin web (React Native, Node.js, GCP, GitHub Actions); reached 16 B2B partner studios and photographers and 500+ DAU, and monetized the app on the App Store.

PROJECTS

Memory-Efficient Long-Video Understanding

Capstone Project

Jun 2026 – Present

- On top of InfiniPot-V KV cache compression for streaming long video (Qwen2-VL-7B), storing the kept KV cache in INT4 instead of FP16 to reduce KV memory at the same token budget.
- Evaluating on Video-MME and MLVU against FP16 baselines.

GPU-Accelerated LLM Inference

- Implemented GPU kernels with core optimization techniques to understand hardware-level bottlenecks.
- Built self-hosted LLM serving with vLLM on limited GPU resources.

Mar 2025 – Present

Data-Efficient 3D Reconstruction

- Input-view selection for 3D Gaussian Splatting via dataset distillation, comparing view-scoring strategies against uniform/random baselines.

Mar 2026 – Jun 2026

MILITARY SERVICE

Republic of Korea Army

Completed military service as Army Sergeant

Jul 2022 – Jan 2024

SKILLS

Languages: Python, C++, CUDA

ML / Inference: PyTorch, vLLM, Model Quantization, LLM Serving