Last updated in Aug 2026

Junhyeok Jeong

Email: spacehyuck@gmail.com |github: www.github.com/jjunhyeok|

SUMMARY

My primary research interest lies in optimizing KV Cache pipelines for LLM inference, with a current focus on designing storage I/O scheduling algorithms to minimize latency. Driven by a passion for system-level efficiency, I aim to expand my expertise toward scalable, high-performance system architectures that support next-gen AI.

EDUCATION

Gyeongsang National University | B.S., Engineering **Jinju, Korea | Mar. 2023 – Ongoing**
GPA : 4.xx/4.5 (TBD)
B.S. in Aerospace and Software Engineering

EXPERIENCES

Big Data System Software lab
Research Intern **Jinju, Korea | Jun. 2024 – Ongoing**
Topics: System Software, Storage System, Machine Learning, OS
Supervisor: Prof. Jaeho Kim

Publications & Presentations

[Faster SSD Failure Prediction with Transformer Architecture using SMART Attributes] *KSC, 2025*
Jobeda Khanam Lia, ***Junhyeok Jeong***, Jaeho Kim

[Poster: Analyzing I/O Characteristics and Latency in Storage-based LLM Inference] *KCS, 2026*
Junhyeok Jeong, Munseok Choe, Jaeho Kim

[Semantic-Aware I/O Isolation Scheduling for VLA Inference in Edge Environments] *ISET, 2026*
Junhyeok Jeong, Jaeho Kim - ***selected as an excellence paper !***

[Counterfactual Explanation-based Failure Boundary Search in Object Detection for Autonomous Flight Software] *KCC, 2026*
Jia Shin, Jinwoo Kim, ***Junhyeok Jeong***, Seonah Lee - ***selected as an excellence paper !***

[Poster: Semantic-Aware Flushing for VLA Inference for Latency and Jitter suppression in continuous VLA inference on Edge Devices] *NVMSA, 2026*
Junhyeok Jeong, Jaeho Kim

Projects

Development of Storage System for OpenVLA (Year 2)

Undergraduate Researcher | 2026.04 - Ongoing

- Designed SAF, an idle-interval KV-Cache flushing and offloading mechanism, reducing mean inference latency by 24.2% and P99 tail latency by 27.7% on Jetson Orin NX edge device
- Achieved 77.8% jitter reduction and completely eliminated OS flusher interventions across 500 evaluation frames
- Submitted the paper on poster session for IEEE NVMSA 2026, as a first author

Research on Optimizing Model Retraining Cycles for AI Cost Reduction

Undergraduate Researcher | 2026. 05 - Ongoing

- Analyzing the correlation between data drift and model performance
- Finding decision-making variables for retraining

Counterfactual Explanation-based Failure Boundary Search in Object Detection for UAV Software

Capstone Design | 2026. 03 - 2026. 06

- Designed a closed-loop XAI-LLM verification framework combining SHAP-guided counterfactual analysis with LLM-based scenario generation for UAV object detection safety validation
- Best Presentation Paper Award at KCC 2026, as a first author

Development of Storage System for OpenVLA (Year 1)

Undergraduate Researcher | 2025. 08 - 2026. 02

- Conducted eBPF-based latency breakdown analysis on Flexgen storage offloading, revealing that app-level software overhead accounts for 93.7% of total latency vs. < 0.4% at the kernel level

- Quantified that ~43% of total I/O requests originate from KV cache access pattern
- Best Presentation Paper Award at ISET 2026, as a first author

Honors and Awards

Presidential Science Scholarship
Ministry of Science and ICT / Korea Student Aid Foundation (KOSAF) 2026

Excellence Presentation Paper Award **[Semantic-Aware I/O Isolation Scheduling for VLA Inference in Edge Environments]**
The 21th IeMeK Symposium on Embedded Technology (ISET 2026)

Outstanding Presentation Paper Award **[Counterfactual Explanation-based Failure Boundary Search in Object Detection for Autonomous Flight Software]**
Korea Computer Congress (KCC 2026)

Dean-selected Scholarships
Seven successive semesters (2023 Spring - ongoing)

Best Buddy Mentors
Exchange Student Program of Gyeongsang National University (2025 Spring)

Skills

Languages
Korean: Native
English: Advanced
- OPIc Advanced Low | Feb 2026
- TOEIC 890/990 | April 2024

Technical Skills
- Coding Languages: Python, C
- Frameworks & Tools: PyTorch, bpftrace, FlexGen, DSPy