

Jiseung Jeong

B.S. student @ UNIST, Ulsan, South Korea

wjdwltmd1151@unist.ac.kr

GitHub: github.com/jiseungjeong LinkedIn: linkedin.com/in/jiseung-jeong

RESEARCH INTERESTS

Efficient and Interpretable On-Device Foundation Models

Currently focusing on Vision-Language-Action (VLA) systems

EDUCATION

Ulsan National Institute of Science and Technology (UNIST)

Ulsan, Korea

B.S. in Computer Science and Engineering

Mar 2021 — Present

GPA: 3.96/4.3 (Major: 4.01/4.3)

RESEARCH AND WORK EXPERIENCE

Ubiquitous AI Lab @ UNIST

Ulsan, Korea

Undergraduate Research Intern

Apr 2026 — Present

- Reduced end-to-end latency by 13.6% in a Vision-Language-Action (VLA) model by pre-computing inference during user query input, using recognized target-object text and vision tokens.

Ubiquitous AI Lab @ UNIST

Ulsan, Korea

Undergraduate Research Intern

Mar 2025 — Jul 2025

- Researched on-device chain-of-thought (CoT) reasoning, focusing on CoT profiling.

LABIS (Startup team, PSD file translation service)

Ulsan, Korea

Software Engineer

Oct 2024 — Jul 2025

- Developed text-layer and style extraction for Photoshop Documents; implemented parallel translation processing with Django Celery and minimized data transferred to the server.
- Optimized post-translation rendering via dynamic wait-time detection, reducing loading delay by approximately 50%.

Republic of Korea Army

Cheorwon, Korea

Military service

Mar 2023 — Sep 2024

PUBLICATIONS

P: Preprint

[P.1] “Premover: Fast Vision-Language-Action Control by Acting Before Instructions Are Complete”

Joonha Park, **Jiseung Jeong**, Taesik Gong

Preprint, 2026.

HONORS AND AWARDS

Manifold Sponsor Track Award (Hackathon)

Apr 2022

- Awarded for developing an AI-based service at the hackathon hosted by Passfinder Inc.
- Built a web-based application with an interactive interface for configuring CNN model layers and hyperparameters.

PROJECTS

Fused Dequantization for Quantized LLM Attention (1st / 13 teams): Course project for CSE412 (UNIST). Designed CUDA kernels that fuse block-wise 4-bit dequantization (NF4, MXFP4, NVFP4) into the QKV projection, achieving 20–25% lower attention latency for NF4/MXFP4 on an RTX 3090.

QuantShift (1st / 21 students): Course project for CSE402 (UNIST). Showed that self-speculative decoding preserves acceptance under 4/8-bit quantization, and that its speedup is bounded by the AR baseline—so NF4, whose single-token decode is already fast, cannot be accelerated by tuning the exit depth.

Shadow Boxing Tracking App with 3 Devices (1st / 8 teams, final presentation): Course project for CSE465 (UNIST). Recognized and logged boxing motions by fusing IMU data from an earbud, phone, and smartwatch.

Lane Detection with Depth Completion — BTS (Brain-to-Society Program, UNIST): Sensor fusion of stereo camera and LiDAR for depth completion and lane detection, implemented in PyTorch.

YOLOkemon App: Course project for CSE465 (UNIST). A YOLO-based Pokédex object-collection app using a post-training-quantized (PTQ) model.

RELEVANT COURSES

UNIST (*all completed with an A+ grade*)

- AI Programming II
- System Programming
- Algorithm
- Mathematical Foundations of Machine Learning
- Computer Network
- Parallel Computing
- Mobile Computing
- Natural Language Processing

SKILLS

- **Certification:** COS PRO Level 1 (C++ Data Structures & Algorithms), 2022.07
- **Programming languages:** Python, C/C++