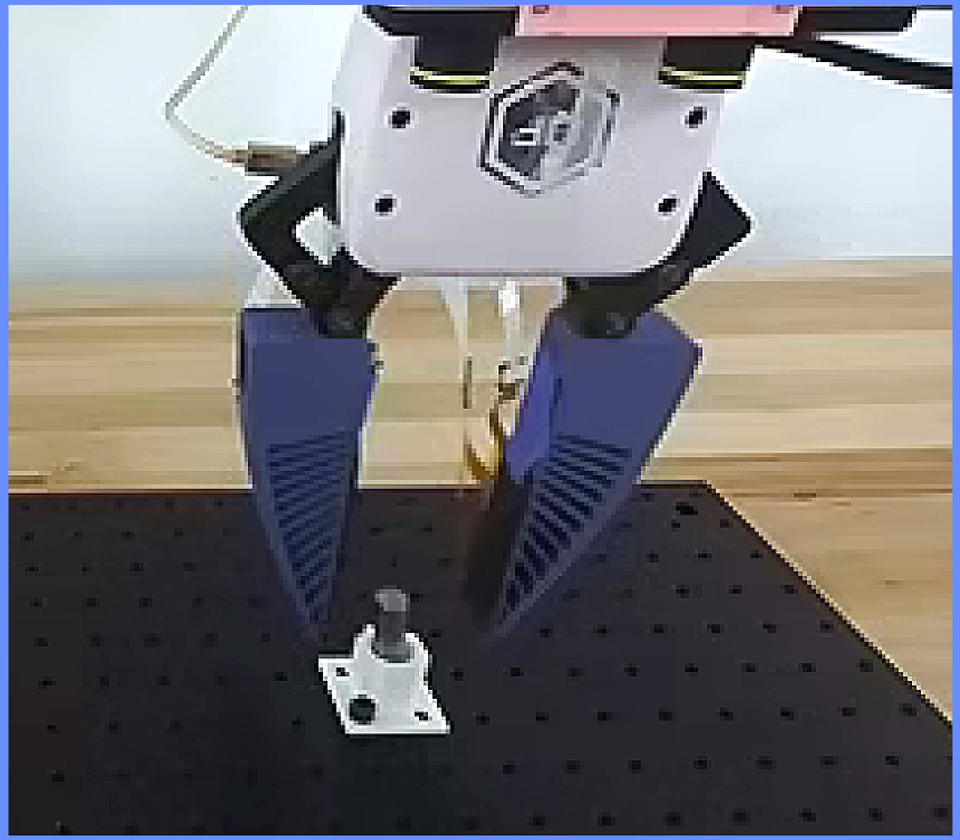
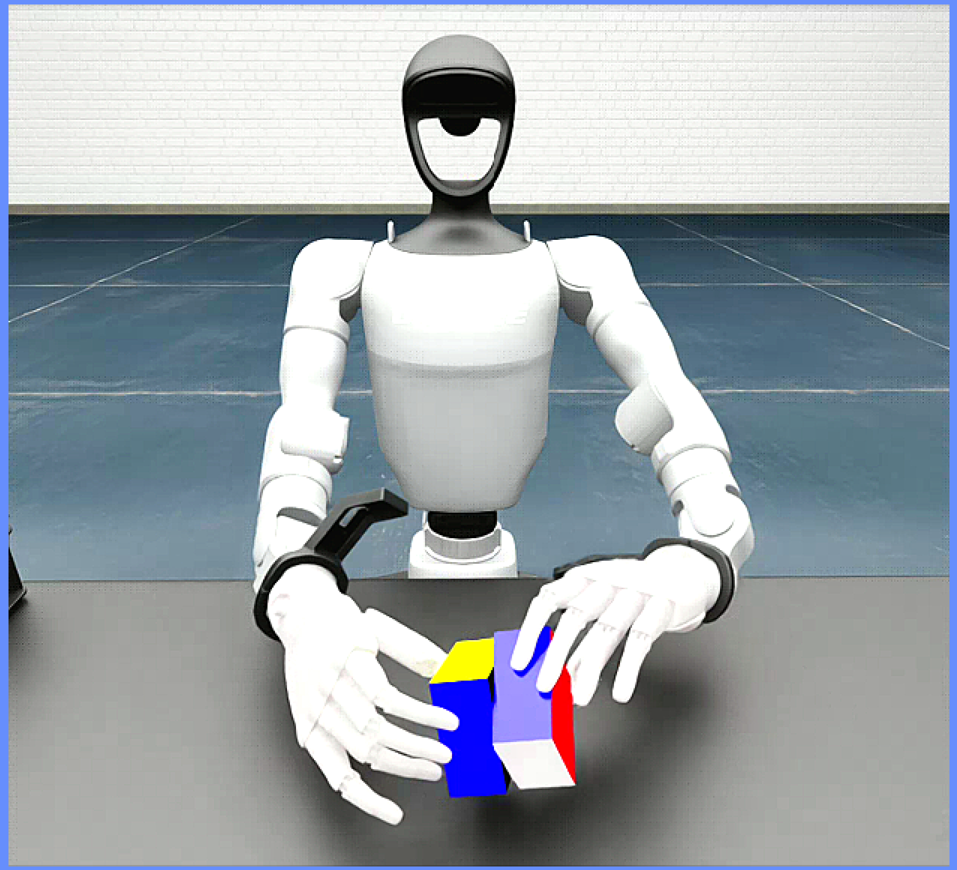
Seongmin Jung

Building robots that see, feel, and understand

sm18570@snu.ac.kr | +82-10-8073-6228

[Homepage] [LinkedIn] [GitHub]





Who I am

I build end-to-end robot vision & learning systems, from research and algorithm development, to real-world deployment and testing.

Background

- M.S. in AI @ Seoul National University (2024-2026)
- Research intern @ NYU AI4CE Lab
- One paper under review @ CVPR 2026

\+ Multiple startup experiences

Core Expertise

Multimodal Sensor Fusion

• Vision + Tactile • Multi-rate synchronization

Foundation Model Development

• VLM/VLA fine-tuning • Diffusion policies • 8×A100

Real-time Robot Deployment

• C++ optimization • Embedded systems

Vision

To empower robots to navigate and interact with the world *even better than we humans do.*

Visual-Tactile Diffusion Policy

New York University (Remote)

01

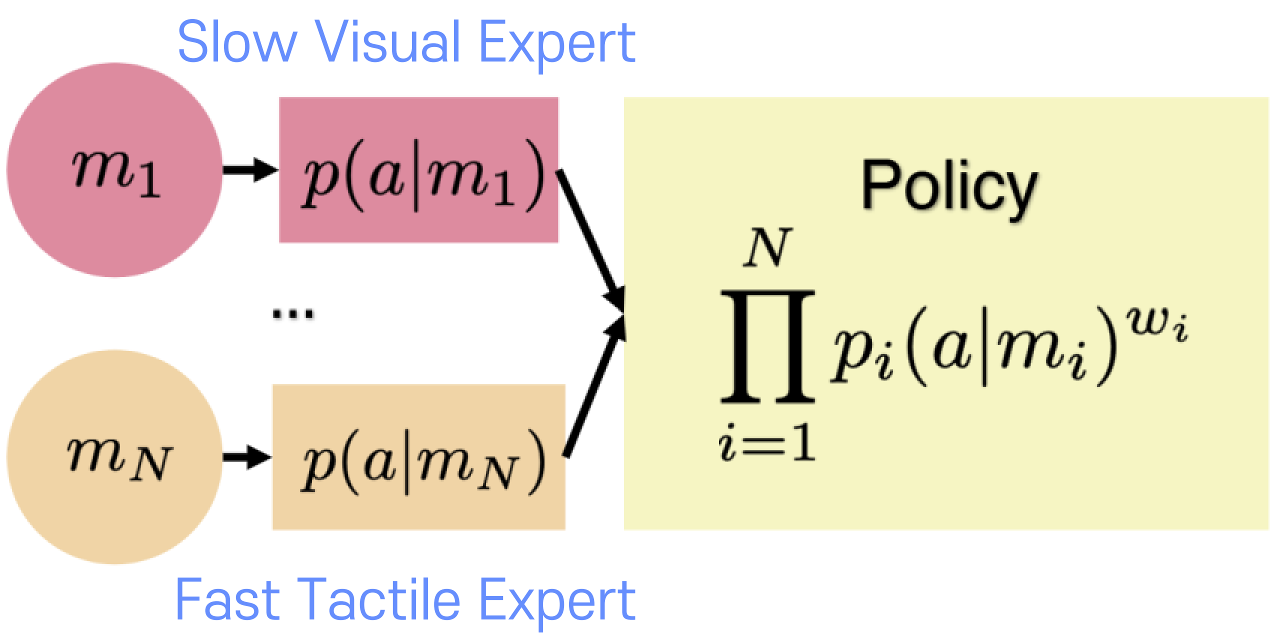
2025.11 - Present

The Challenge

Vision: 30Hz / Tactile: 1000Hz
Vision-only policies miss critical, momentary event
(e.g. slips / perturbations)

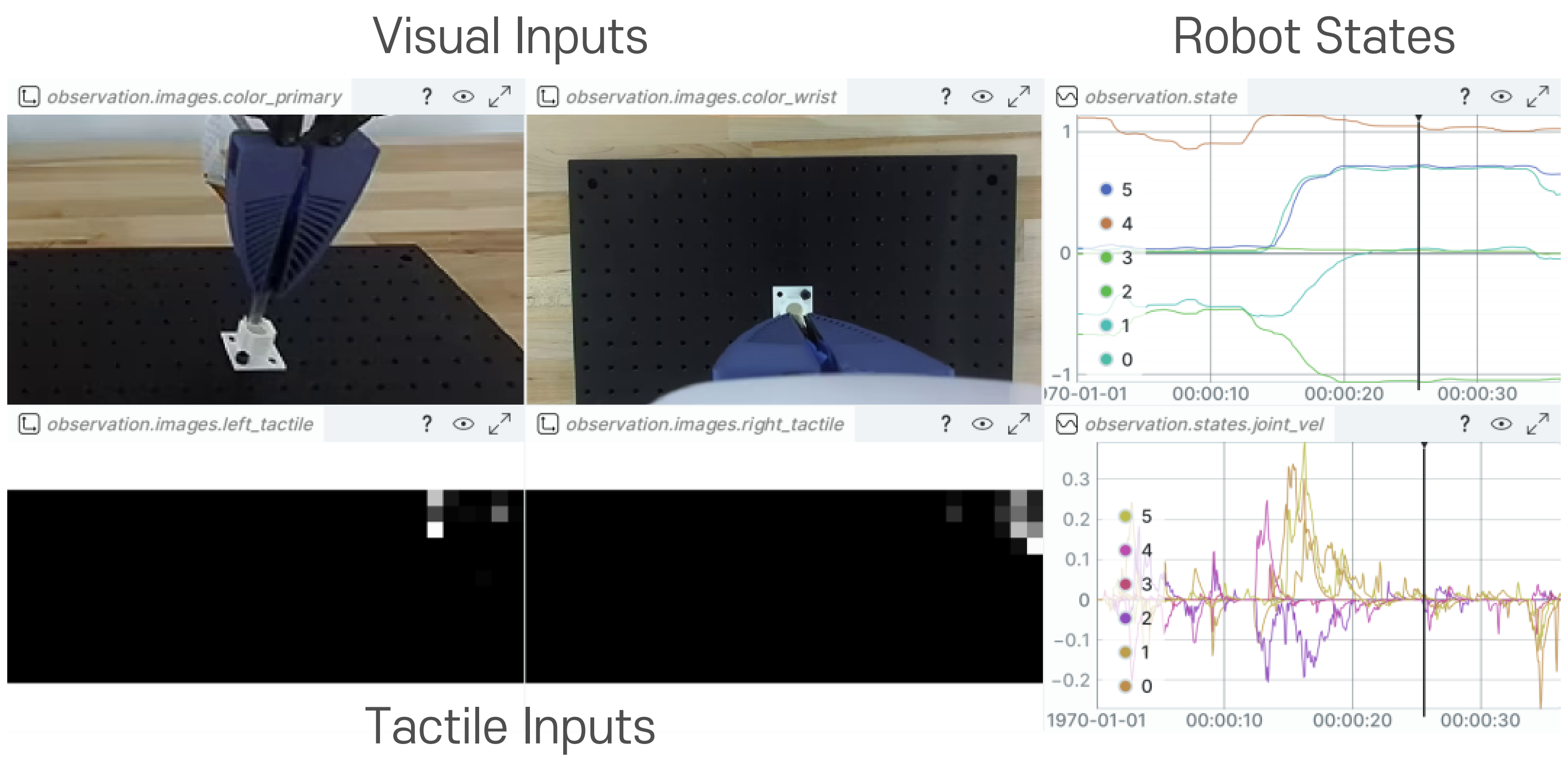
Our Approach

Compositional Diffusion Policy
→ Real-time tactile feedback



Progress So Far

- Built compositional diffusion policy baseline.
- Optimizing architecture and training objectives.
- Expanding to cable routing and other contact-rich tasks.



Peg-in-hole data example

My Contribution

- Training framework and baseline implementation (RDP, Minimal Iterative Policy).
- LeRobot customization (dataloader, models).
- Sensor sync pipeline (30Hz ↔ 1kHz alignment).

Tech Stack

- ML: PyTorch, LeRobot, Diffusion Models.
- Hardware: Piezo tactile (~1kHz), RGB-D cameras (30Hz).

PanoGrounder

CVPR 2026 Under Review
[ArXiv] [Project Page]

02

Seoul National University

2025.01 - 2025.11

3D Visual Grounding Challenge

Task: "Find the brown desk in the corner"
→ Locate object in 3D scene

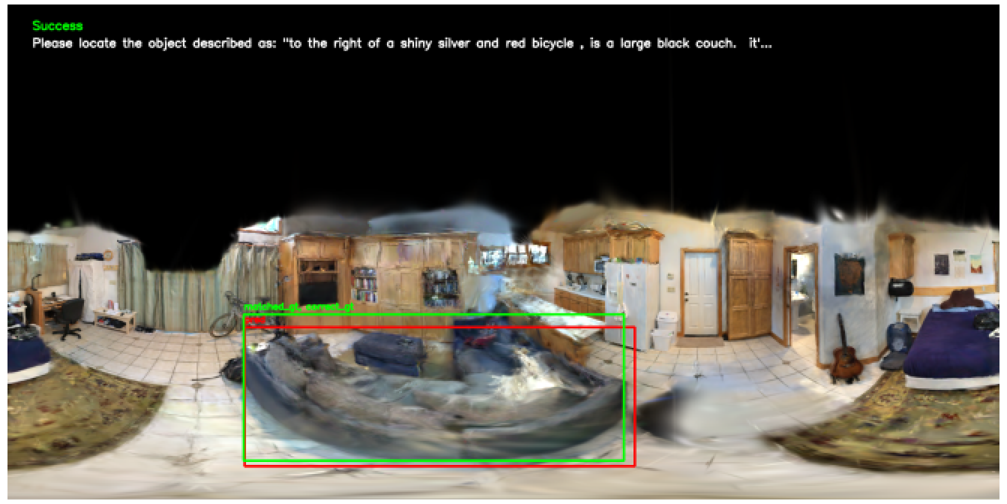
Prior Work Issues:

- Limited complex text & spatial reasoning
- Poor cross-dataset generalization

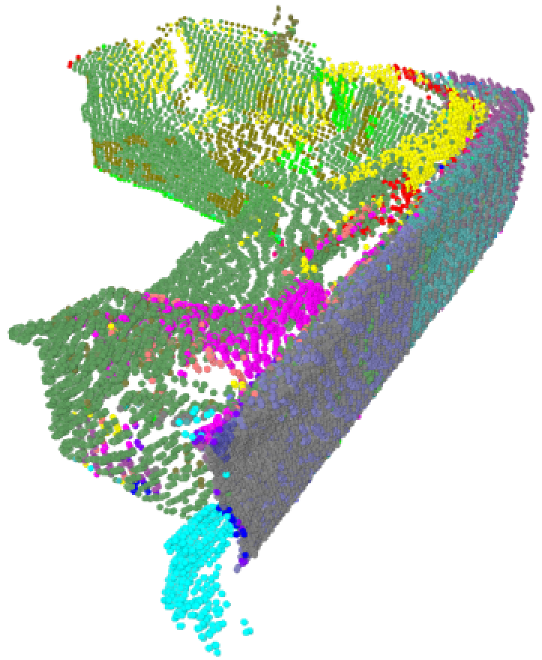
Our Approach

3DGS + VLM

Key Insight → Leverage VLM's strong 2D reasoning



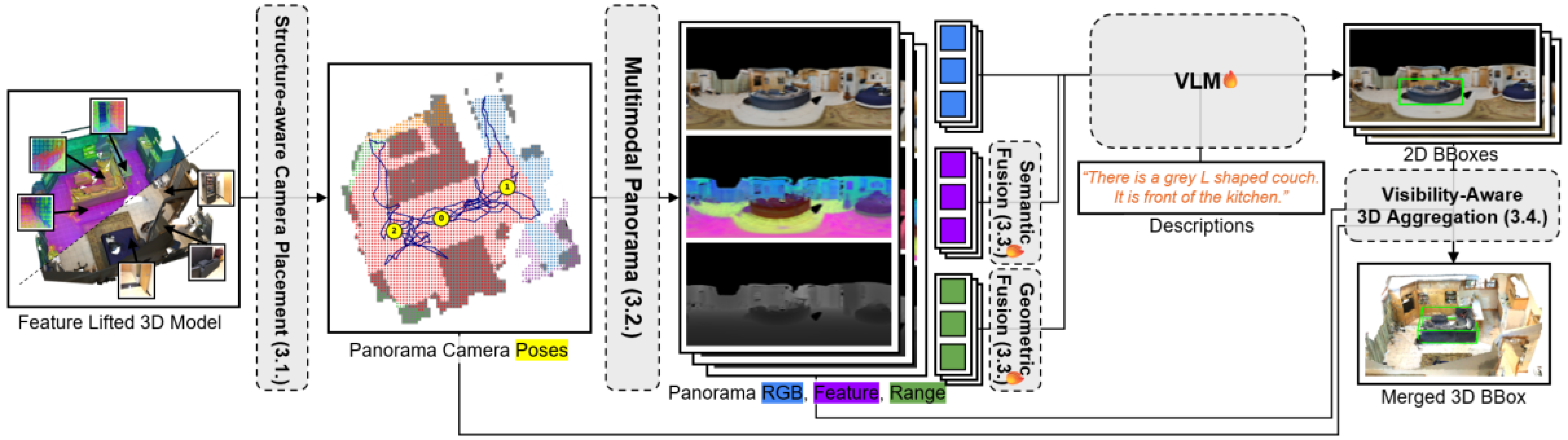
Panoramic rendering of 3DGS



Lift VLM'S 2D output back into 3D

Method

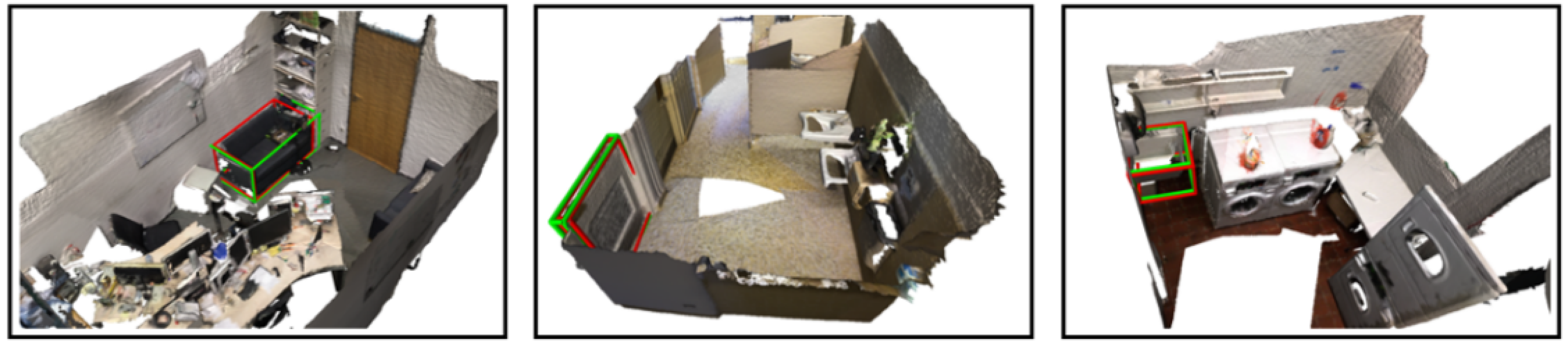
1. Multimodal Panorama Rendering
    - RGB + DINO features + depth from 3D mesh/3DGS.
2. VLM Inference
    - DINO injection via custom adapter; Fine-tuned with LoRA on 8×A100.
3. 3D Lifting
    - 2D predictions → 3D point cloud.



Overall Pipeline

State-of-the-Art Result

- In-distribution: +4.7% (Nr3D)
- Cross-dataset generalization: +17.4% (ARKitScenes)



(a) it is a black sofa. it is located to the wall behind the fan.

(b) the door is tan with a metal grating. the door is in the left corner.

(c) there is a white squared shaped sink right next to the door. below the sink there is small white trash can.

Qualitative Results

My Contribution

- Proposed the core idea & architecture.
- Developed and fine-tuned the VLM pipeline.
- Conducted extensive ablations and benchmarking.

Solving Rubik's Cube

Seoul National University

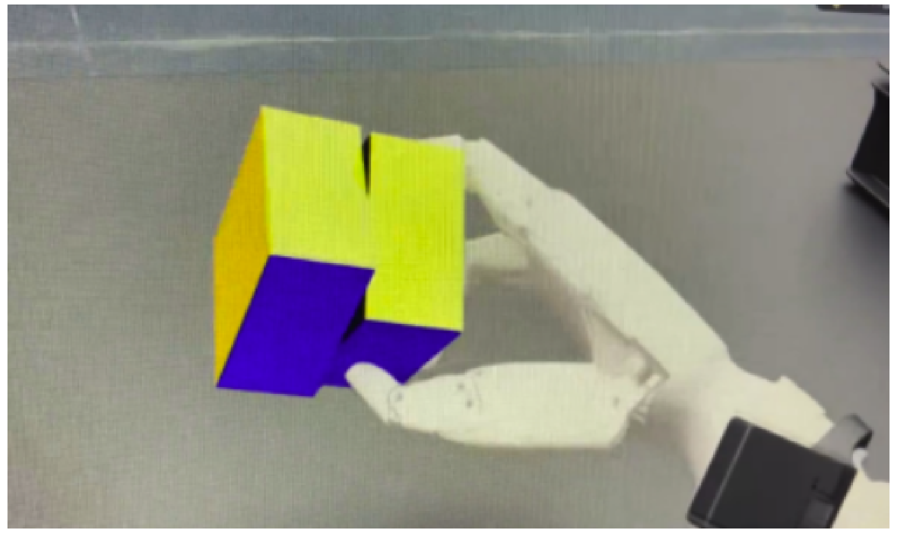
03

2025.09 - Present

The Challenge

Bimanual manipulation for precise tasks.

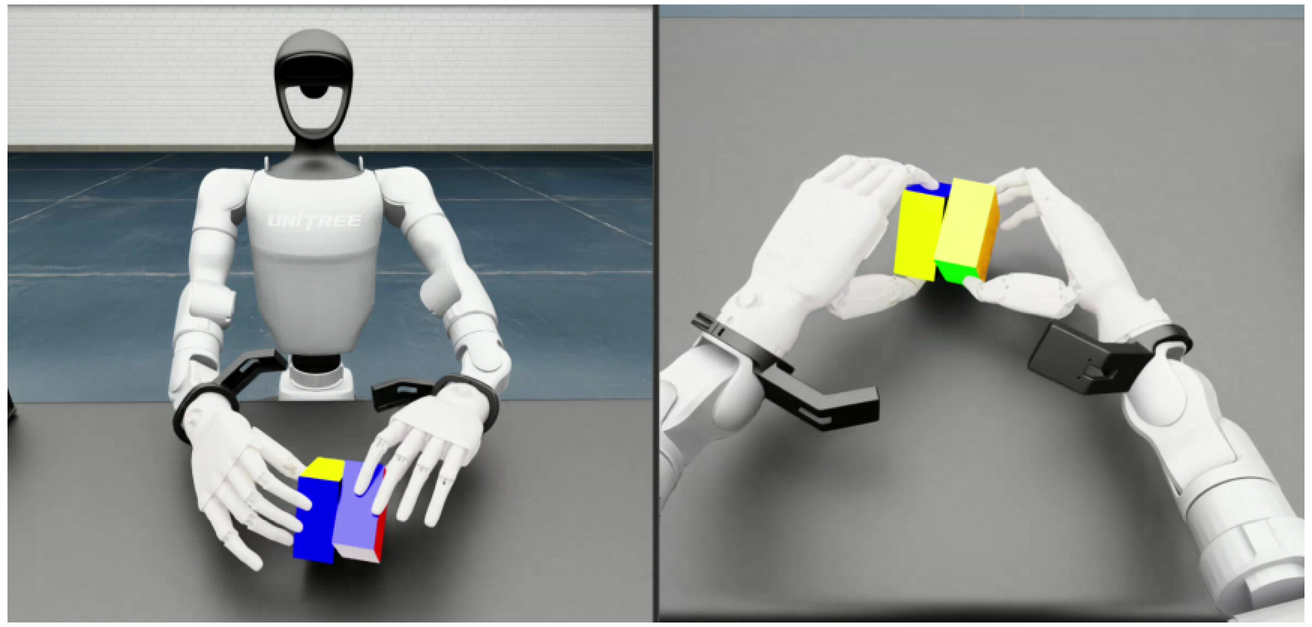
- 2 hands coordination
- Stable grasping & force control
- Dynamic geometry: Cube shape changes with rotation



Our Approach

VR Teleoperation → VLA Fine-tuning

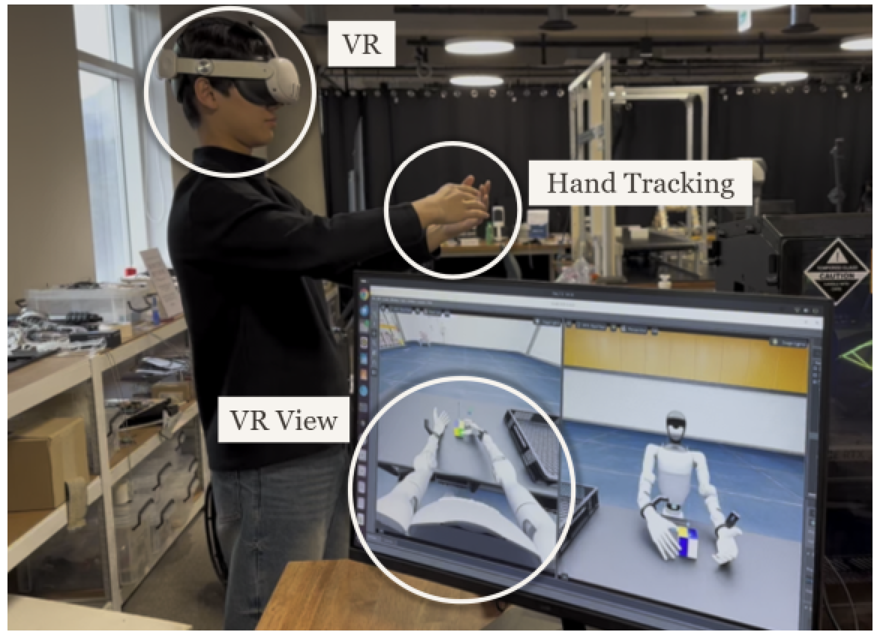
Simulation-only for simplification



Success case example [Video link]

Data Collection [Video link]

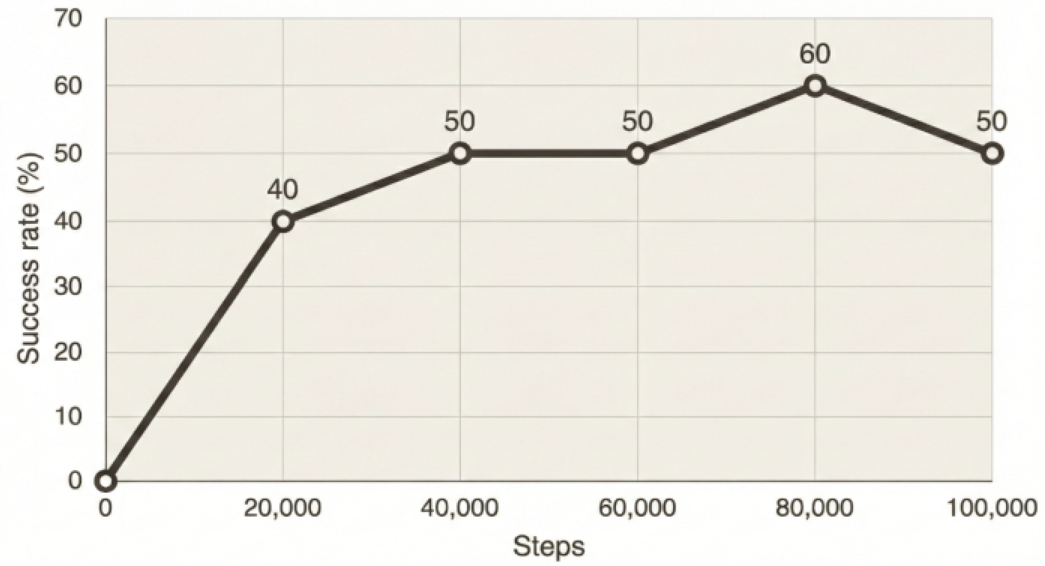
- VR teleop using Meta Quest 3.
- Collected ~120 episodes.
- Mapped human to robot hand motions.



Data collection setup

Fine-Tuning Results [Video link]

- Followed GR00T official fine-tuning pipeline.
- Achieved 60% success rate.
- Key insight: Data diversity critical for generalization.



Fine-tuning result

Future Work (Course Project → Research Extension)

- Exploring video hand tracking for scalable data collection (eliminates VR hardware dependency).
- Early-stage pipeline development, targeting paper submission.

Neural SLAM

Seoul National University

2024.09 - 2024.12

The Challenge

NeRF-based SLAM too slow for real-time deployment.
Camera tracking: 0.5s/frame (iterative rendering).

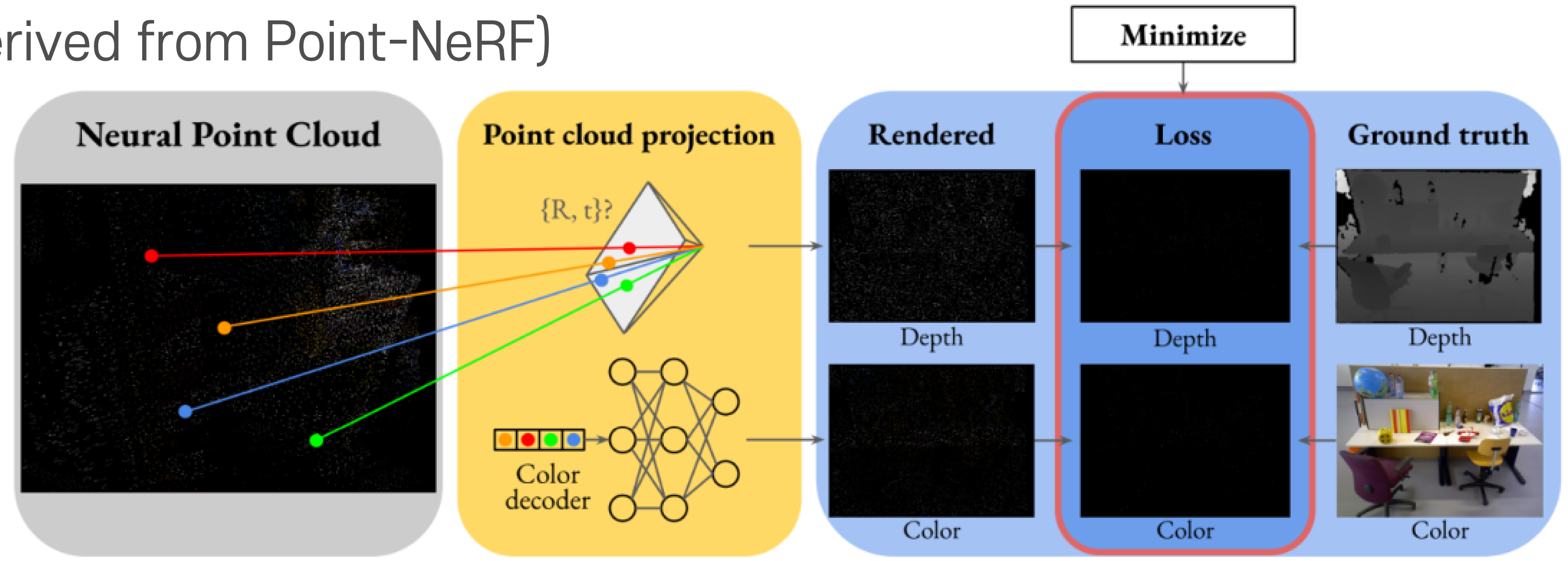
Our Approach

Based on Point-SLAM.
Key: Replace NeRF rendering → Point Projection.
Keep other parts of the pipeline.

Result

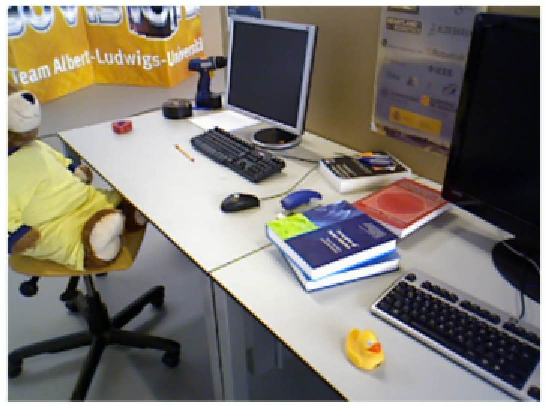
Achieved over 3× faster tracking.
Similar or better trajectory accuracy than Point-SLAM.

(Derived from Point-NeRF)

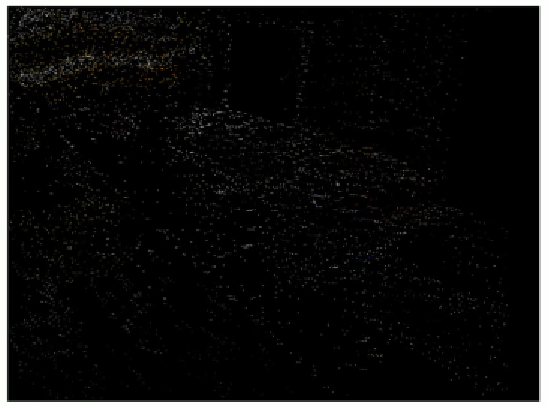


Overall Architecture

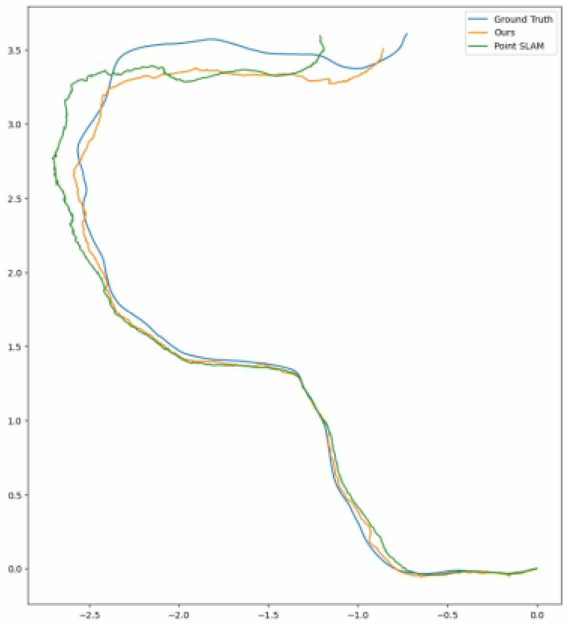
[Video link]



GT RGB frames



Projected NPC



Trajectory

High-Speed Autonomous Navigation

05

Yonsei University

2023.07 - 2023.12

Goal: Reach goal and return to start at high speed.

Hardware

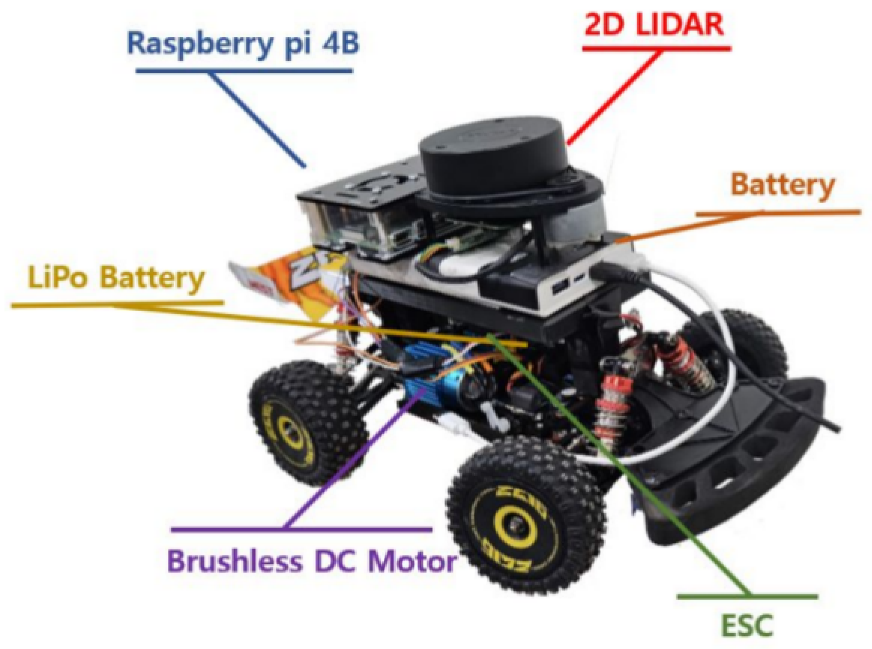
- Embodied system with 2D LiDAR and BLDC motor.

Software

- ROS & C++ pipeline with PWM motor control.
- Dead reckoning for localization and control (throttle & steering).
- LiDAR point cloud clustering for obstacle detection & avoidance.

Result

- Max 5m/s navigation.
- Real-time obstacle avoidance.



Hardware Setup

[Video link] Goal pose is behind the obstacle

Thank You.

Let's revolutionize *human-robot interaction*
with unmatched dexterity and intelligence.

Seongmin Jung

sm18570@snu.ac.kr | +82-10-8073-6228

[Homepage] [LinkedIn] [GitHub]