2026 제41회
제어로봇시스템학회 학술대회
(ICROS 2026)

Visual Language Multi-robot SLAM

- What, When, and Why Visual-Language Navigation?

김기섭, gsk@dgist.ac.kr

DGIST 로봇및기계전자공학과 조교수

2026.07.03

DGIST

Giseop Kim

Team Leader

Giseop Kim

Assistant Professor

APRL | DGIST

gsk@dgist.ac.kr

Google Scholar

Biography

I am an Assistant Professor in the Department of Robotics and Mechatronics Engineering at DGIST, where I lead the Autonomy and Perceptual Robotics Lab (APRL). I also hold joint appointments in the Department of Artificial Intelligence and the Mechanical Engineering Track at DGIST.

Before joining DGIST, I was a Research Scientist at NAVER LABS (2021-2024), working on autonomous driving and 3D vision. I received my PhD, MS, and BS from KAIST, all in Civil and Environmental Engineering, under the guidance of Prof. Ayoung Kim.

My research focuses on enabling robots to perceive and navigate the real world robustly and efficiently through SLAM, 3D reconstruction, sensor fusion, and spatial AI.

Research Statement | Teaching Statement

Research Interests

- Simultaneous Localization and Mapping (SLAM)
- Mobile Robot Navigation
- Spatial AI & Physical AI

Education

- PhD in Civil & Env. Eng., 2022, KAIST
- MS in Civil & Env. Eng., 2019, KAIST
- BS in Civil & Env. Eng., 2017, KAIST

APRL

- Autonomy and Perceptual Robotics Lab (APRL)

Current Lab Members

Full-time Researchers

Bokeon Suh

Integrated MS/PhD student (2025F-)

Jiseon Kim

MS student (2025F-)

Yumin Lee

MS student (2025F-)

Hyoseok Ju

MS student (2025F-)

Nayak Bibhutibhusan

Postdoc (2025F-)

Doyeon Kim

PhD student (2026S-)

Beomsu Kim

MS student (2026S-)

Hoyun Kim

MS student (2026S-)

Incoming Student

MS student (2026F-)

Incoming Student

MS student (2027S-)

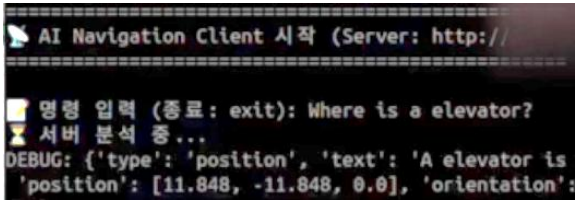
Visual Language Action (VLA) model

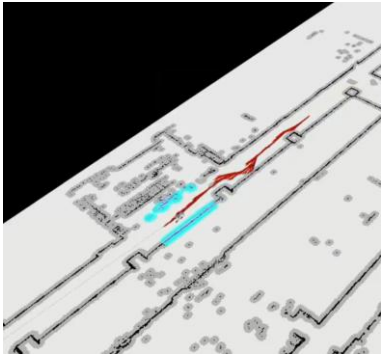
Visual Language Action (VLA) model for Navigation

Visual Language Action (VLA) model for Navigation

What is Visual-Language Navigation?

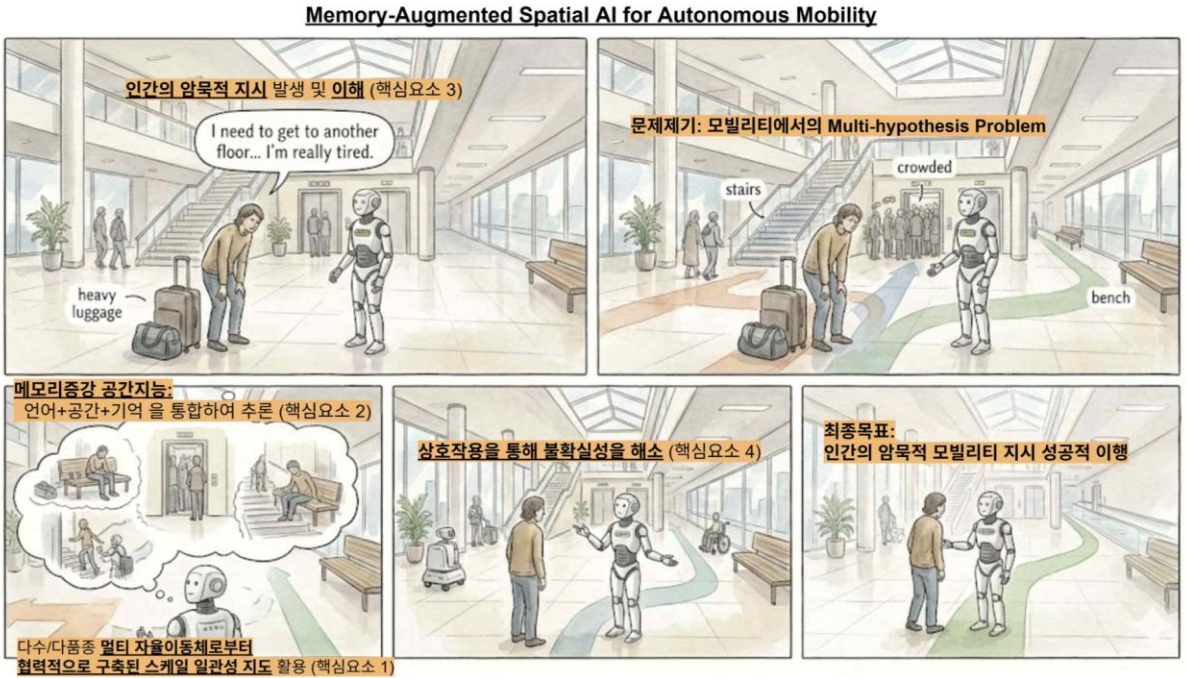
- “Where is the elevator?”





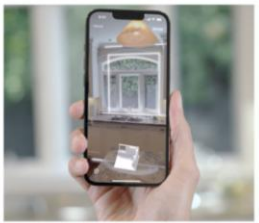
When Visual-Language Navigation?

- Human-robot Interactive Navigation (HRIN)



Why Visual-Language Navigation?

- **Robot Web Era**



Apple iPhone LiDAR RoomPlan

Project Aria from Meta

Unitree Go1 Legged Robot

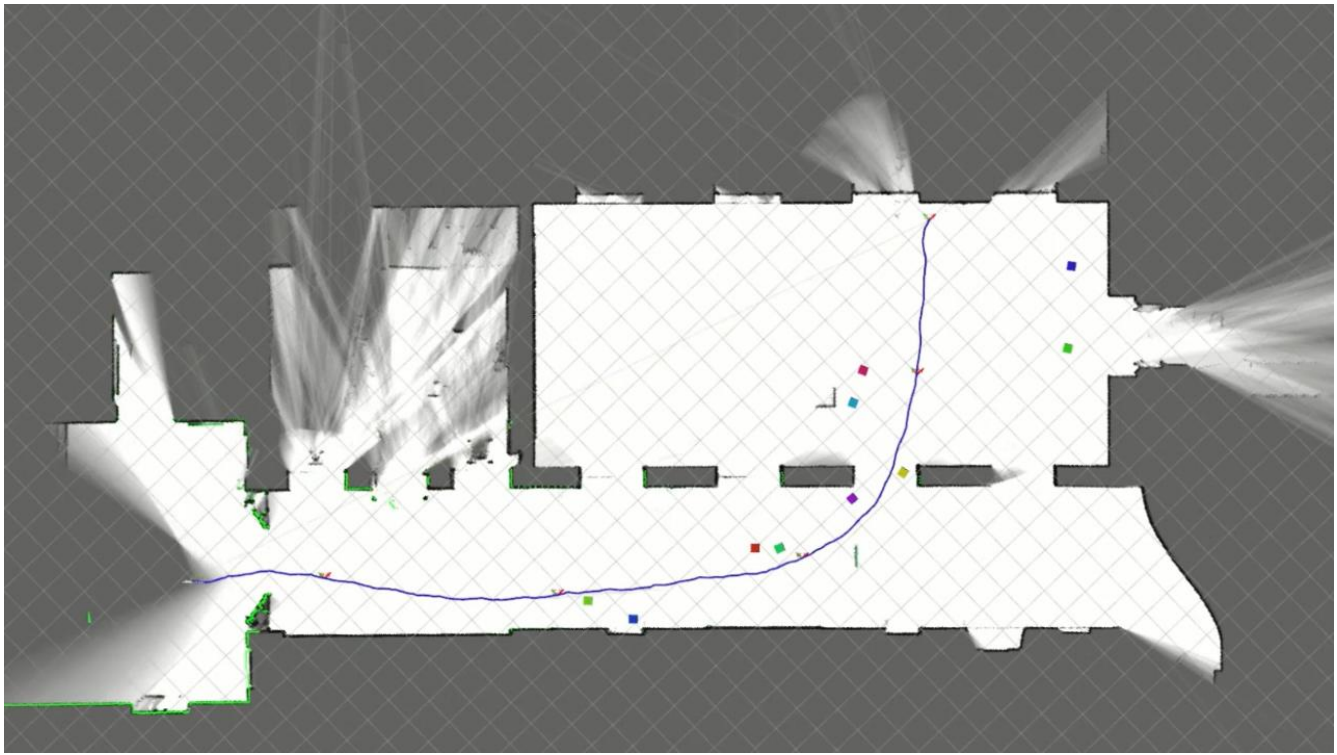
NAVER LABS Unmanned Shuttle (ALT-B)

- DGIST 컨실리언스홀 전체를 이종로봇 5대가 자율적으로 돌아다니며 하루 안에 공간적으로 이해하려면 어떻게 해야할까요?
- 휴머노이드 로봇이 대중교통을 이용해 강남에서 DGIST까지 안전하게 오려면 어떤 기술이 필요할까요?
- 네비게이션의 플래닝 및 의사결정에서 사람의 언어적 및 비언어적 표현은 어떻게 통합해야 할까요?

SLAM (Simultaneous Localization and Mapping)

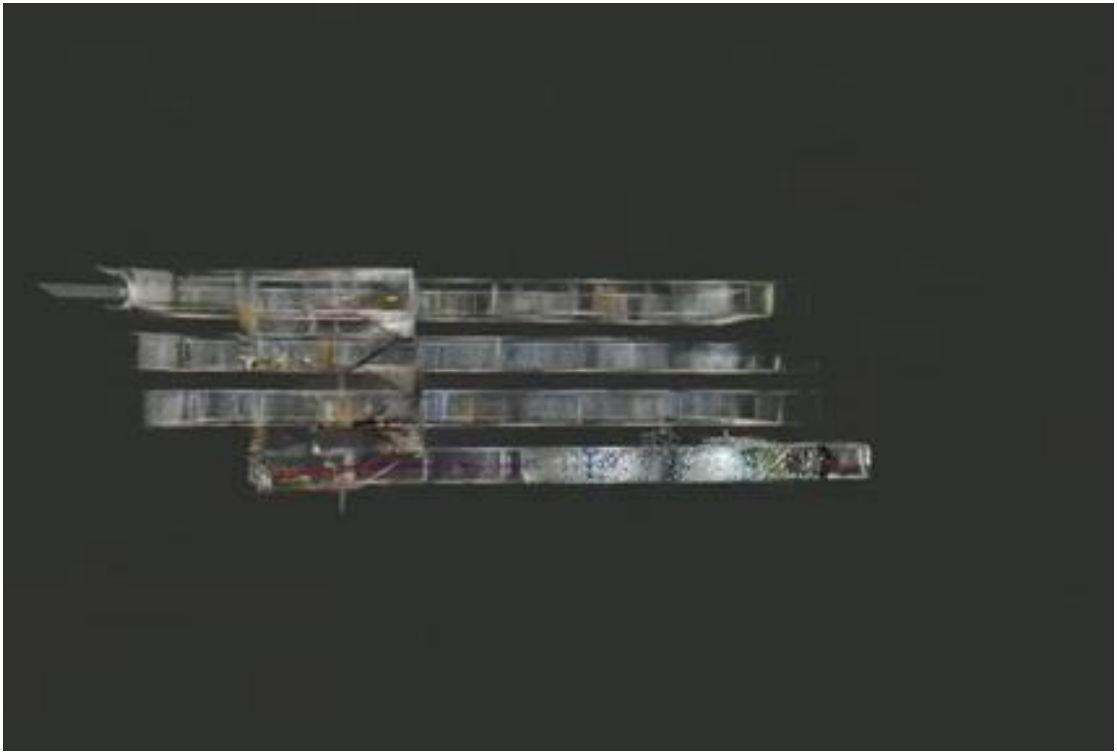
COLMAP (CVPR 2016, Structure-from-Motion Revisited)

SLAM (Simultaneous Localization and Mapping)

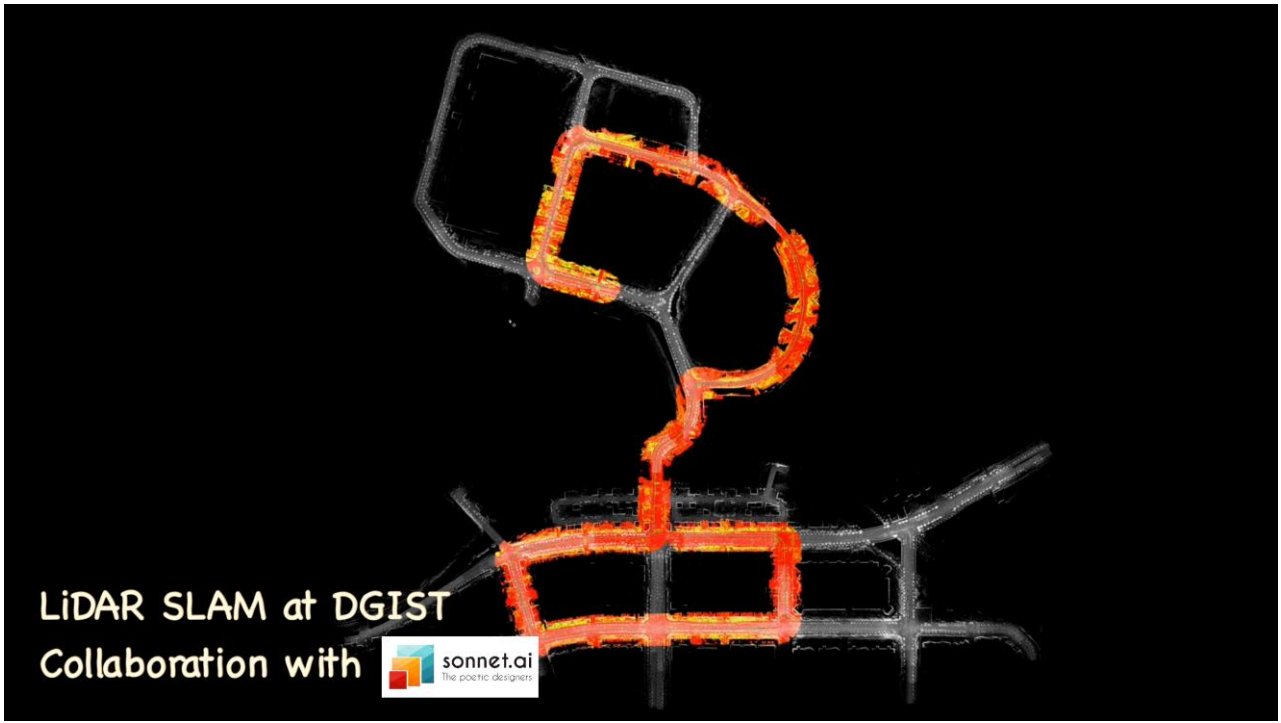


Google Cartographer, 2016

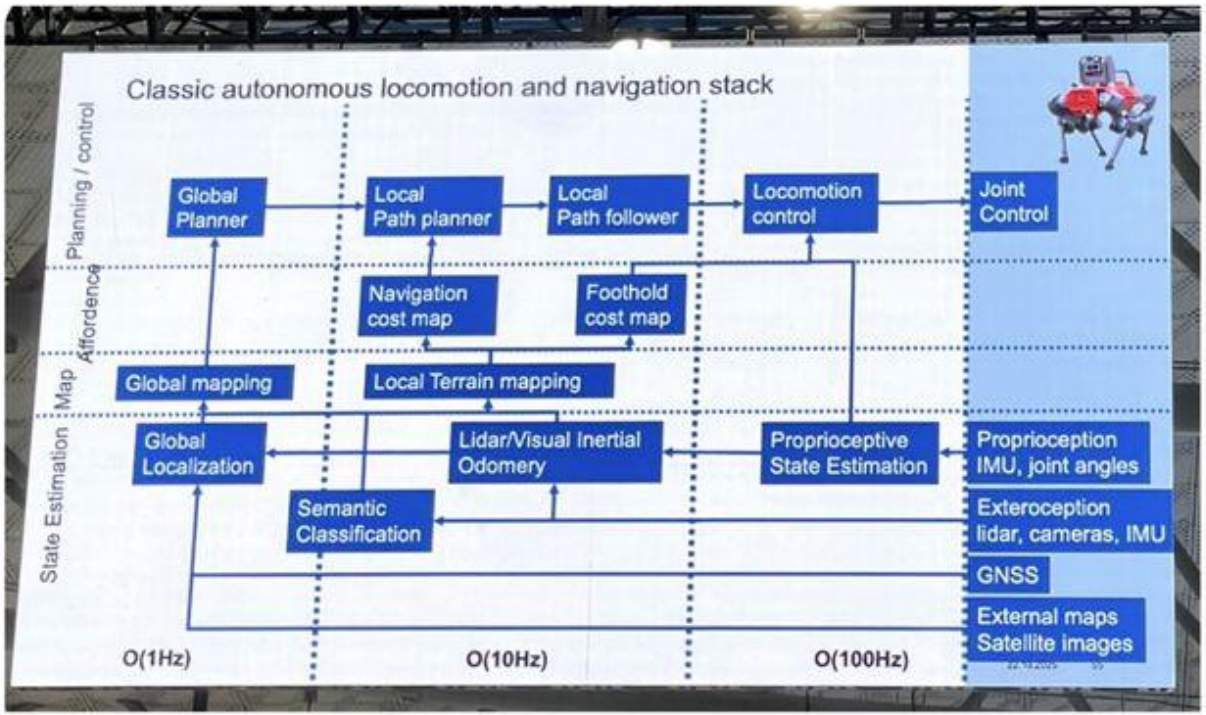
SLAM (Simultaneous Localization and Mapping)



Autonomous Robot Navigation

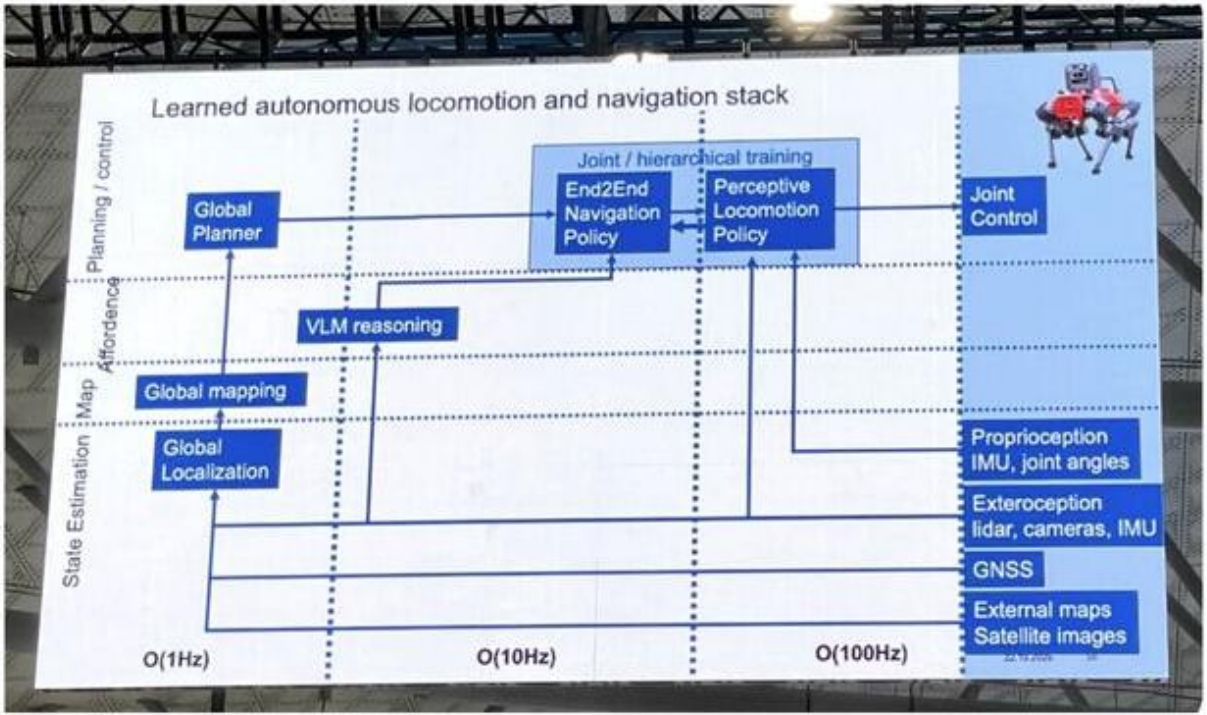


What is Visual-Language Navigation?



Source: Prof. Marco Hutter, from IROS2025

What is Visual-Language Navigation?

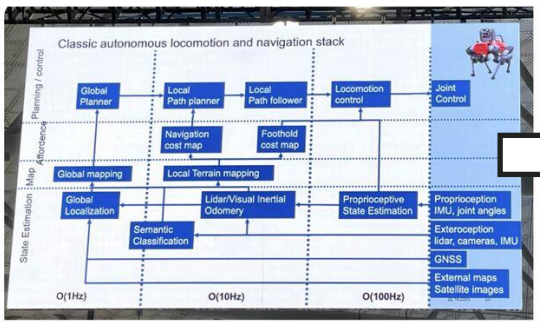


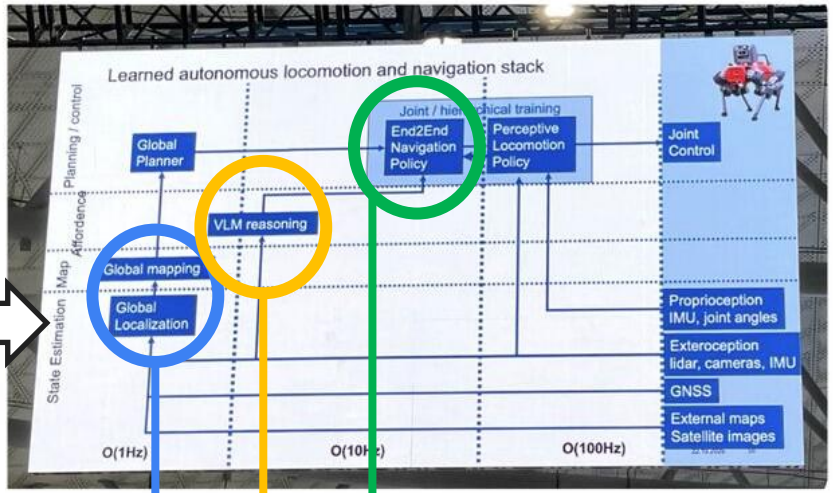
Source: Prof. Marco Hutter, from IROS2025

Autonomous Robot Navigation

– Historical View (a.k.a PPP)

Source: Prof. Marco Hutter,
from IROS2025





P1: Perception

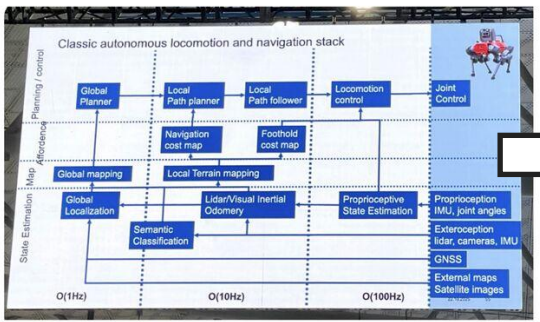
P2: (P1+) Prediction

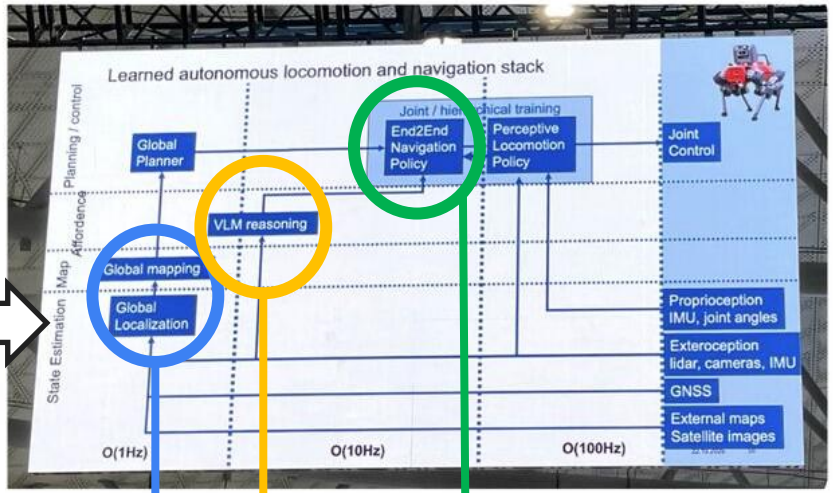
P3: (P1+P2+) Planning

Autonomous Robot Navigation

– APRL Perspective (Task-centric)

Source: Prof. Marco Hutter,
from IROS2025





Human-Robot Collaborative Mapping

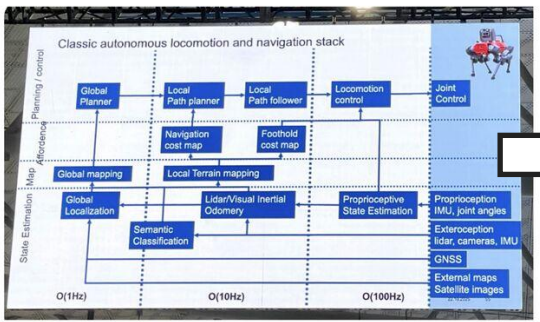
Continual Human-Environment Understanding

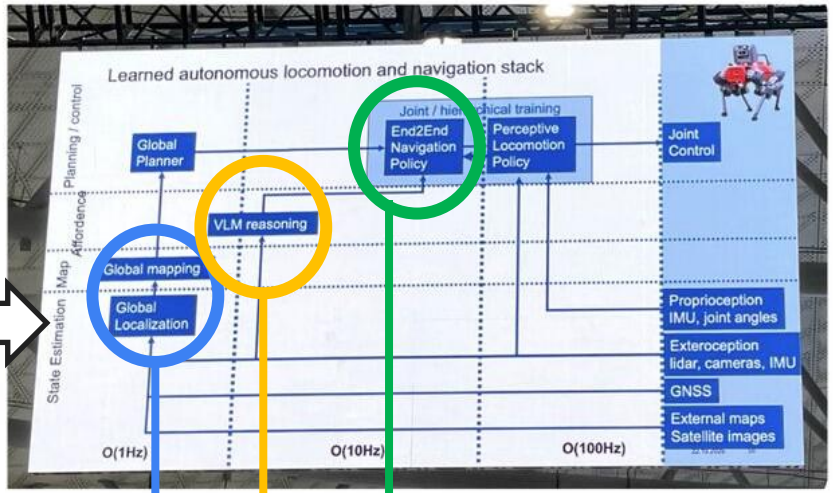
Human-Robot Interactive Navigation

Autonomous Robot Navigation

– APRL Perspective (Method-centric)

Source: Prof. Marco Hutter,
from IROS2025





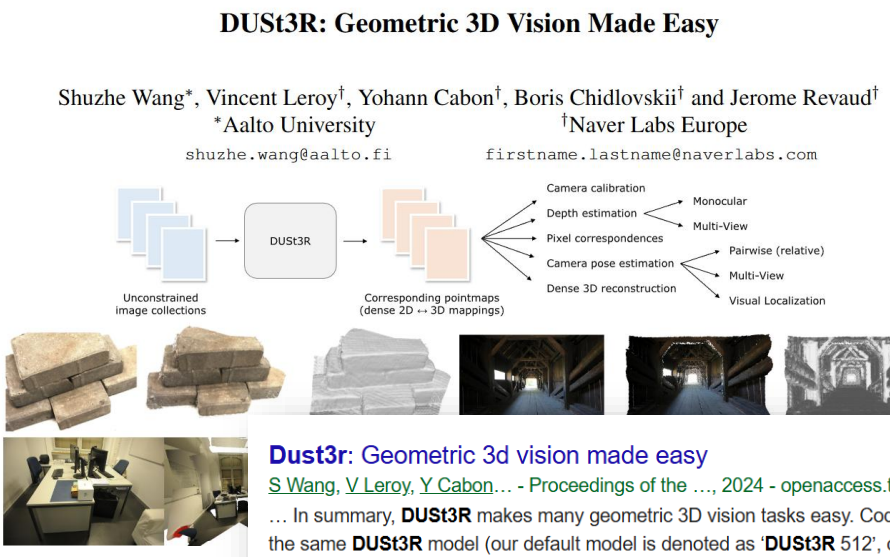
Geometric Foundation Models

Map and Memory Management

Visual-Language Navigation

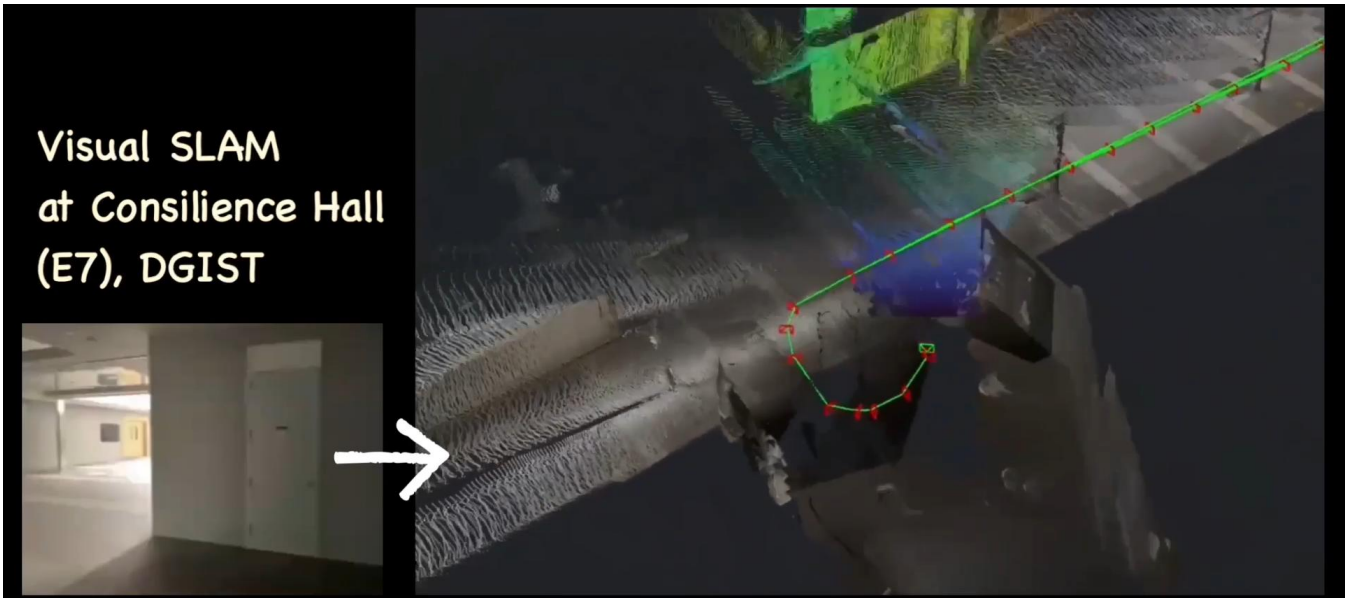
Part 1
Human–Robot Collaborative Mapping in Robot Web Era

Geometric Foundation Models



- Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." CVPR 2024
- Murai, Riku, Eric Dexheimer, and Andrew J. Davison. "Mast3r-slam: Real-time dense slam with 3d reconstruction priors.“, CVPR 2025

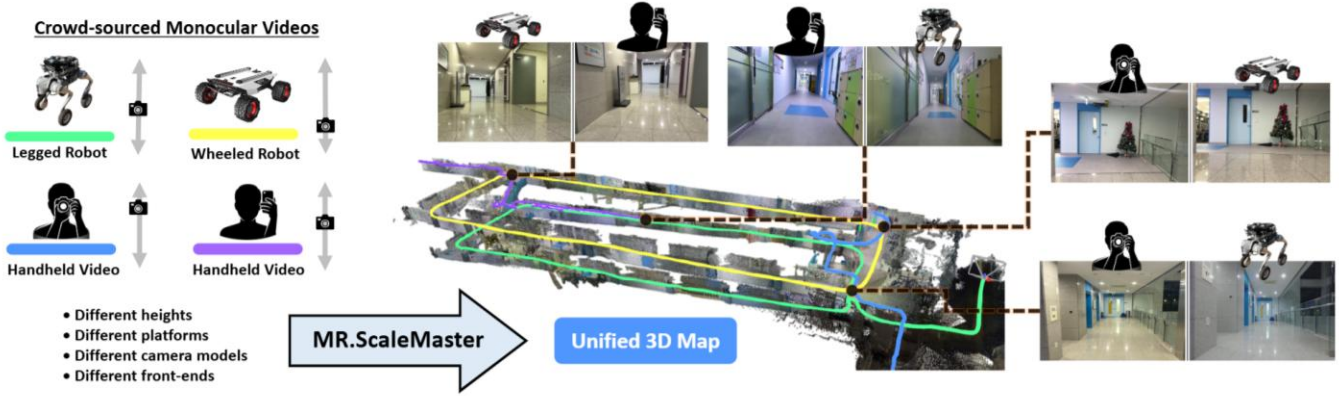
Geometric Foundation Models



Hyoseok Ju, Bokeon Suh, and Giseop Kim. "Have We Mastered Scale in Deep Monocular Visual SLAM? The ScaleMaster Dataset and Benchmark.“, ICRA 2026

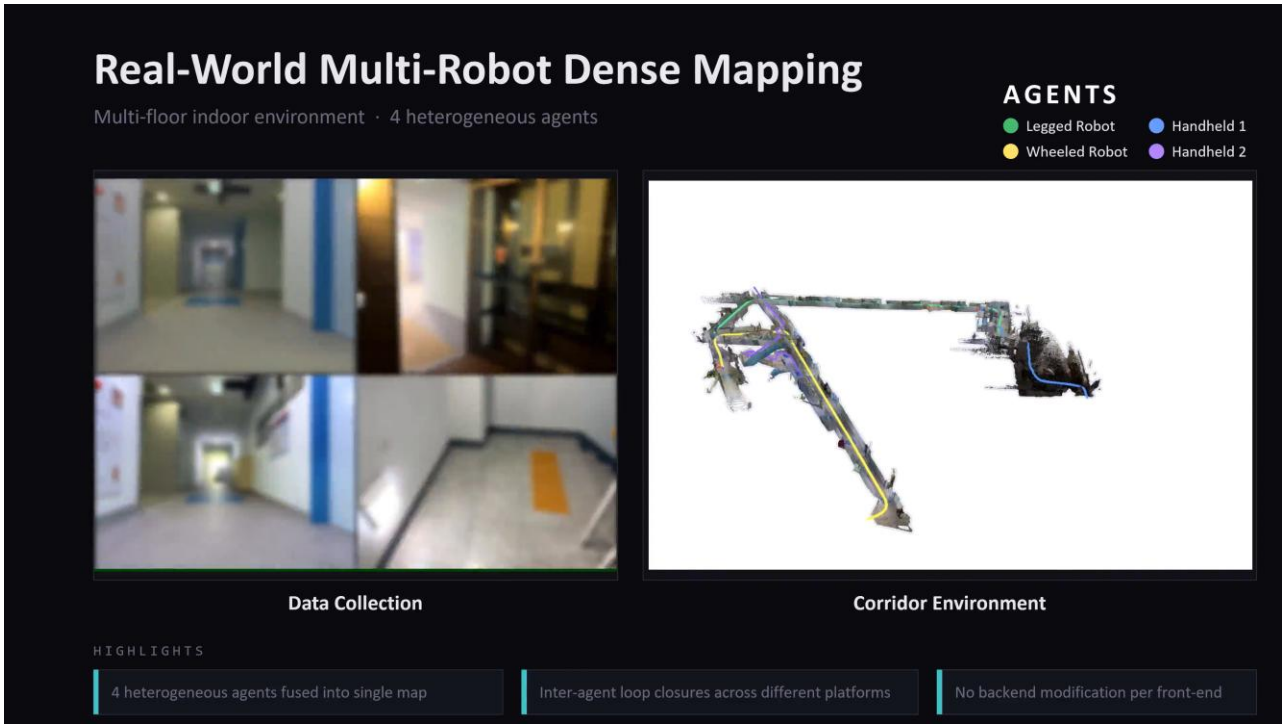
Human-Robot Collaborative Mapping

- **In the GFM era.**



Hyoseok Ju, and Giseop Kim. "MR.ScaleMaster: Scale-Consistent Collaborative Mapping from Crowd-Sourced Monocular Videos." arXiv preprint arXiv:2604.11372 (2026).

Human–Robot Collaborative Mapping



Hyoseok Ju, and Giseop Kim. "MR.ScaleMaster: Scale-Consistent Collaborative Mapping from Crowd-Sourced Monocular Videos." arXiv preprint arXiv:2604.11372 (2026).

KITTI 00,

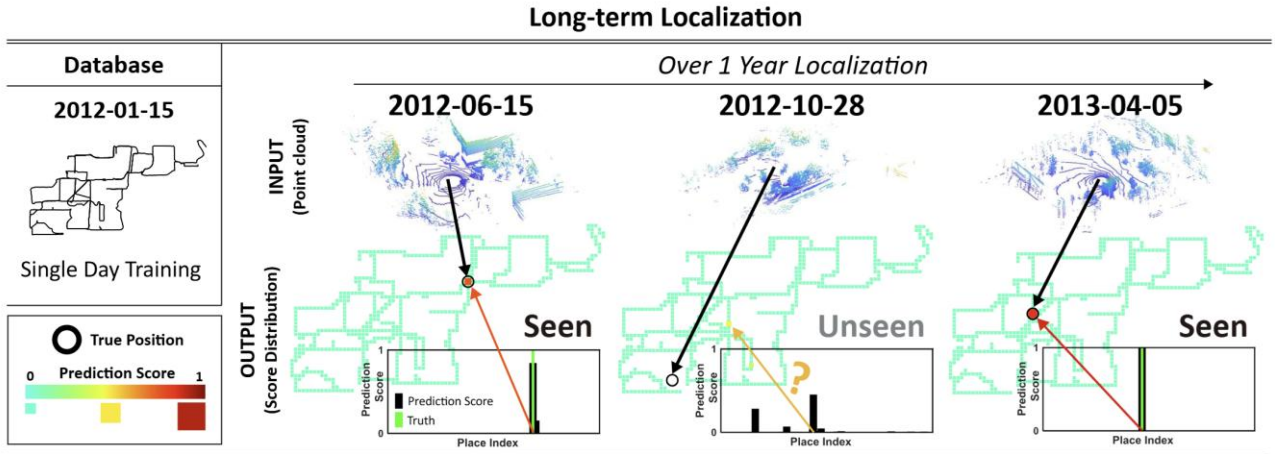
MR.ScaleMaster
with MASt3R-SLAM

Part 2

Continual Human–Environment Understanding

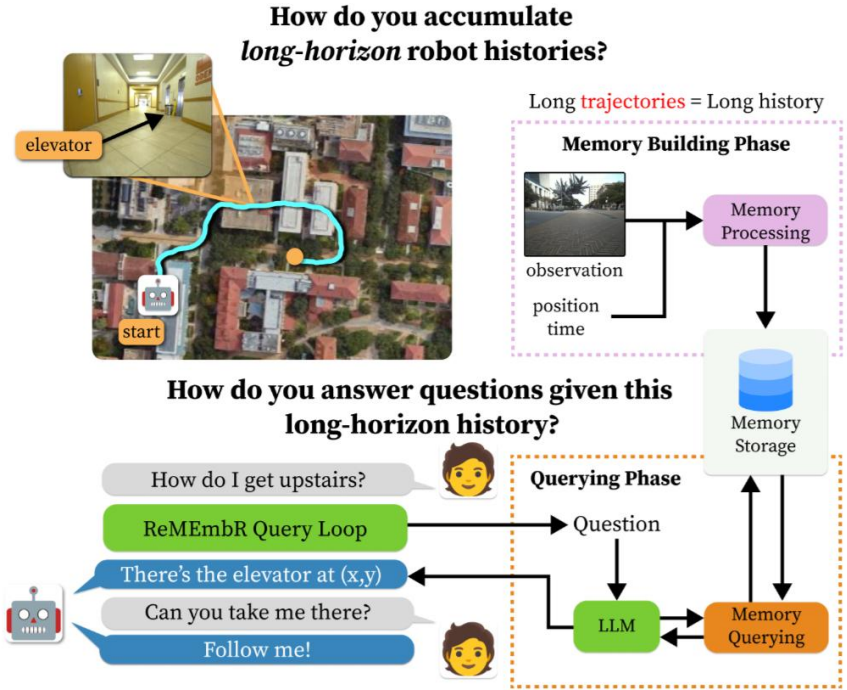
Traditional Lifelong Navigation

- **Map for Long-term Navigation**



Giseop Kim, Byungjae Park, and Ayoung Kim. "1-day learning, 1-year localization: Long-term lidar localization using scan context image." RA-L (2019)

Memory for Robot Navigation



Anwar, Abrar, et al. "Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation." ICRA 2025

Memory for Lifelong Robot Navigation

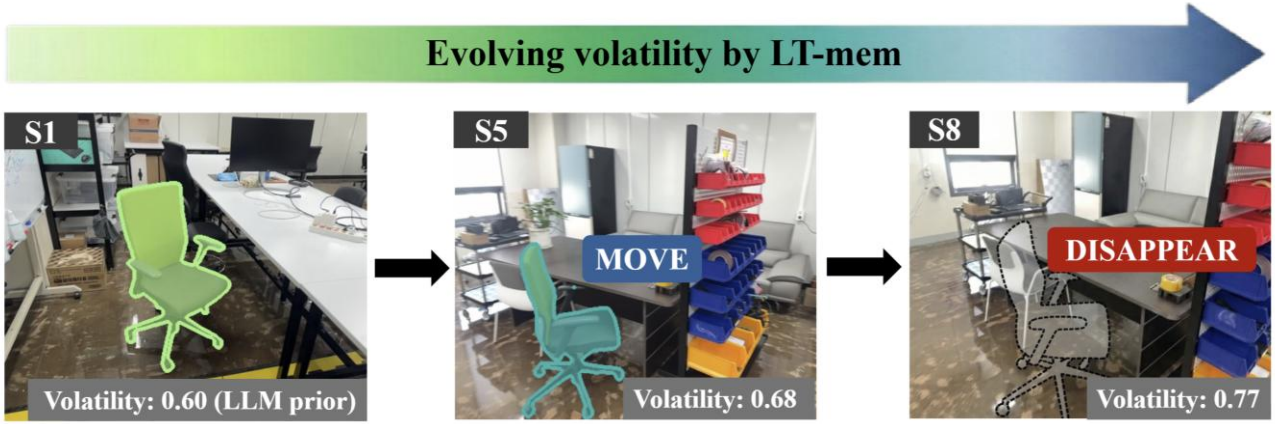
- **Long-term Navigation Memory**

Example Query
"When did the scissor first appear?"
"Was the vacuum in the room at Session 9?"
"Did the brown basket move at Session 6?"
"Did the fridge stay in place at Session 4?"
"Did the green chair come back at Session 9?"
"Where has the blue totebag been across all sessions?"
"Which object moved most frequently?"
"When was the last time the white board moved?"
"Was the robot dog in the same location in S_1 and S_{10}?"

Yumin Lee, Hyoseok Ju, Giseop Kim, LT-Mem: Volatility-Aware Spatio-Temporal Memory for Lifelong Scene Understanding, IROS 2026 (Accepted, To appear)

Memory for Lifelong Robot Navigation

- **Long-term Navigation Memory**



Yumin Lee, Hyoseok Ju, Giseop Kim, LT-Mem: Volatility-Aware Spatio-Temporal Memory for Lifelong Scene Understanding, IROS 2026 (Accepted, To appear)

Memory for Lifelong Robot Navigation

- **VQA for Lifelong Scene Understanding**

Q: When should I go if I want to find a parking spot?
A: Come around 8:00AM to find a parking spot (Session3, 0 cars at that time).

Q: When is it hardest to find a place to park?
A: It's hardest to find a spot around 1:03PM (Session8, 34 cars).
Avoid that time if you want an empty spot.

Yumin Lee, Hyoseok Ju, Giseop Kim, LT-Mem: Volatility-Aware Spatio-Temporal Memory for Lifelong Scene Understanding, IROS 2026 (Accepted, To appear)

Part 3

Human–Robot Interactive Navigation

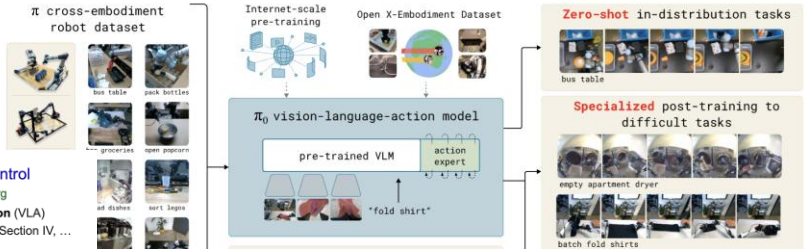
VLA

- π0

π_0: A Vision-Language-Action Flow Model for General Robot Control

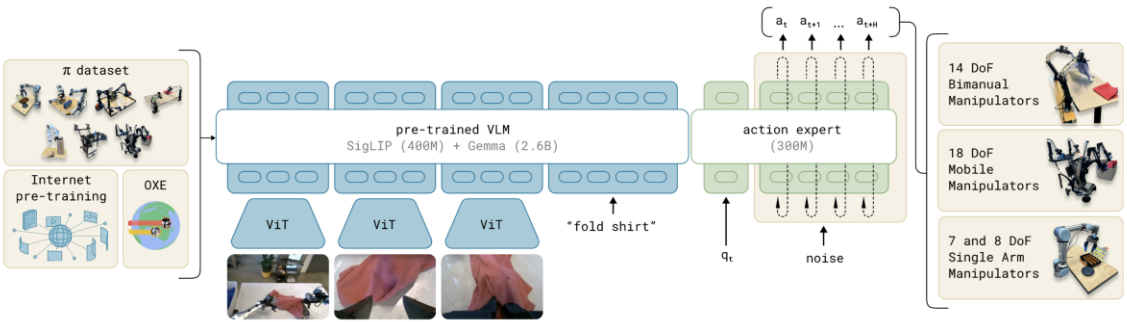
Physical Intelligence
Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, Ury Zhilinsky
https://physicalintelligence.company/blog/pi0

31 Oct 2024



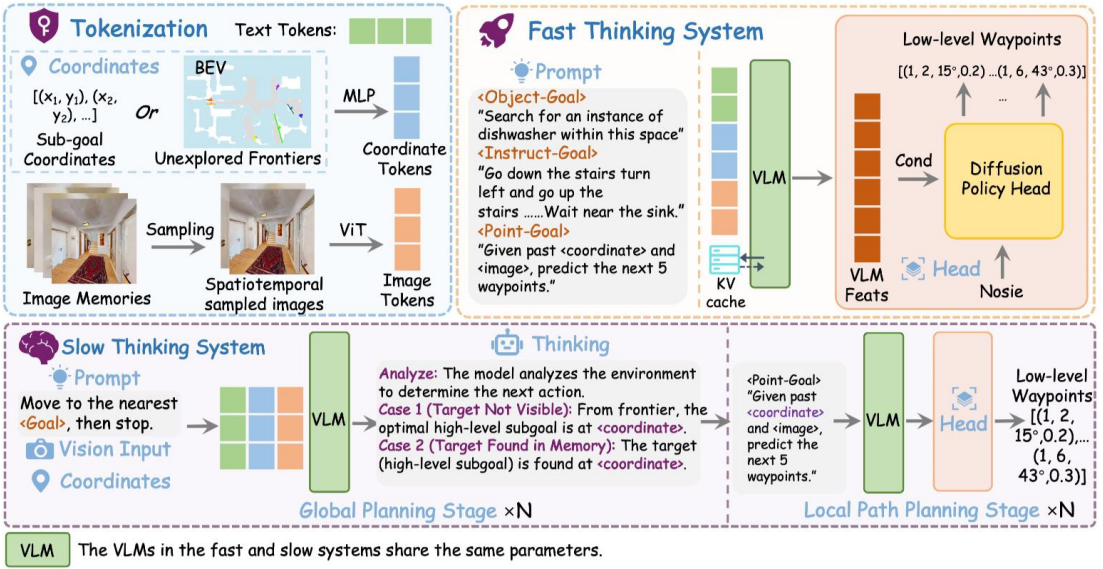
π_0: A Vision-Language-Action Flow Model for General Robot Control
K Black, N Brown, D Driess, A Esmail, M Equi… - arXiv preprint arXiv …, 2024 - arxiv.org
… **models**. Our work is most closely related to recently proposed **visionlanguage action** (VLA) **models**… We compare this fine-tuned **π0 model** with the **π0**small **model** described in Section IV, …
☆ 저장 99 인용 1898회 인용 관련 학술자료 전체 4개의 버전 »

$\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization
K Black, N Brown, J Darpin
… We describe **π0**.5, a ne
VLAs in environments that
☆ 저장 99 인용 174회 인



Visual-Language Navigation

- **VLN: VLA for Navigation**



Xue, Xinda, et al. "Omninav: A unified framework for prospective exploration and visual-language navigation." ICLR 2026

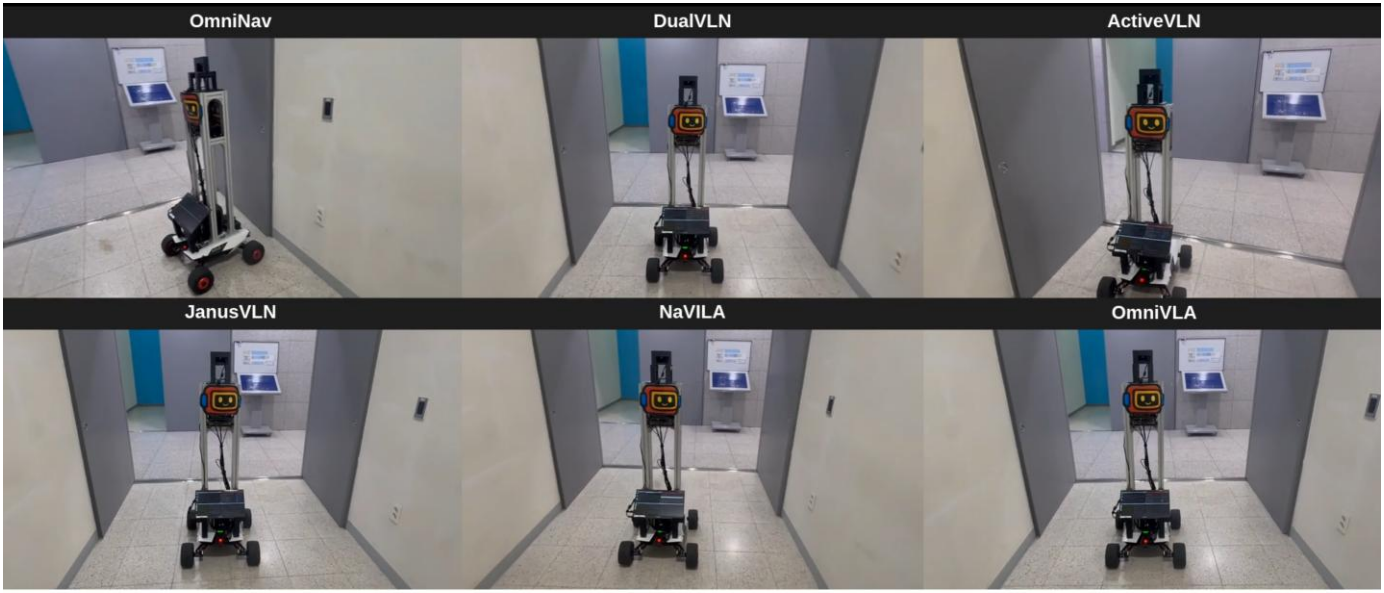
Visual-Language Navigation

for Human-Robot Interactive Navigation

Instruction:
Walk forward along the white wall .Once you are close to the passage on your right, make a right turn and enter the passage. Continue walking straight, following the black striped tiles on the floor. After passing the fire hydrant, turn slightly to the right and move forward until you see a black sofa. Walk towards the black sofa and stop in front of it to finalize the trajectory.

Instruction: Walk forward along the white wall .Once you are close to the passage on your right, make a right turn and enter the passage. Continue walking straight, following the black striped tiles on the floor. After passing the fire hydrant, turn slightly to the right and move forward until you see a black sofa. Walk towards the black sofa and stop in front of it to finalize the trajectory.

Are We Ready for VLN in the Real World?



Instruction: Go straight ahead and turn right. Then keep going straight, and when you turn right at the corner, you'll see a blue door labeled 105 on your left. Stop in front of it.

in preparation

Part 4

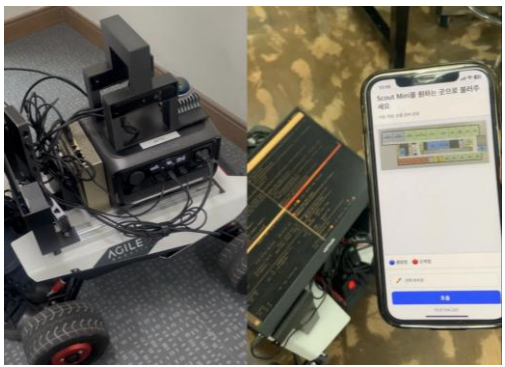
Multimodal Interaction

– what is the human-like way?

Visual-Language Nav meets HRI

- In the real world people don't describe every (low-level) detail.
  - They point.

- Point-to-Language Grounded Navigation via Diverse Map Interfaces



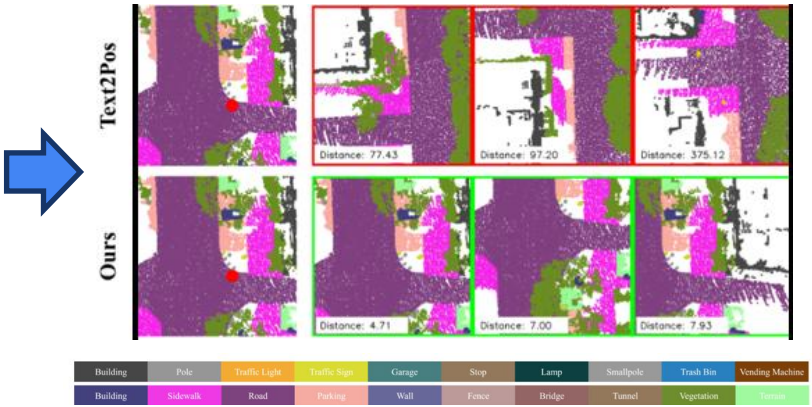
in preparation

Visual-Language Nav meets HRI

- Language meets robot sensing modality
- Humans think in language; robots think in sensor data.
  - Language (Hint, Human-side) to Pointcloud (LiDAR map, Robot-side)

Query

h_1: The pose is on-top of a gray road.
h_2: The pose is on-top of a bright-gray smallpole.
h_3: The pose is south of a bright-gray smallpole.
h_4: The pose is west of a beige sidewalk.
h_5: The pose is south of a beige smallpole.
h_6: The pose is south of a gray-green pole.



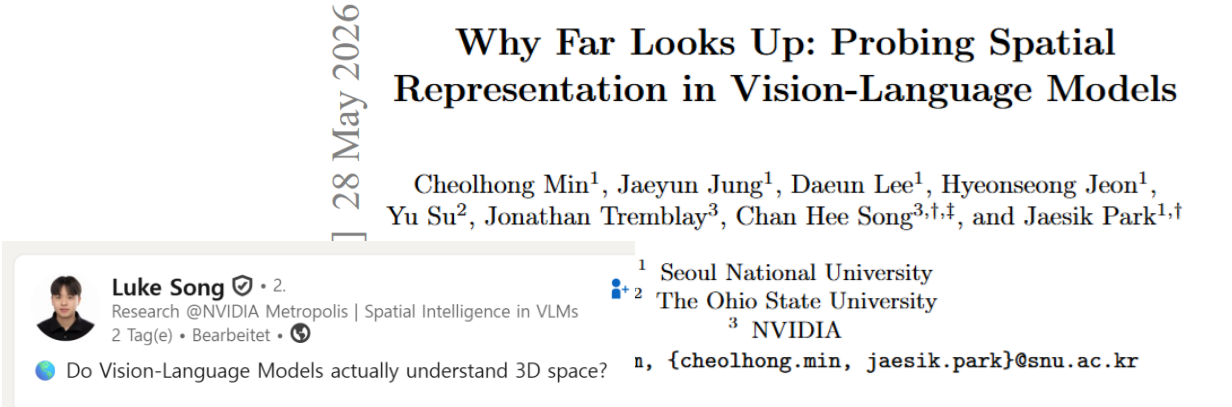
in preparation

Part 5

Reasoning or Geometry?

Visual-Language Nav, but why SLAM?

- Is Perception Foundation Model Still Necessary?



] 28 May 2026

Why Far Looks Up: Probing Spatial Representation in Vision-Language Models

Cheolhong Min[1], Jaeyun Jung[1], Daeun Lee[1], Hyeonseong Jeon[1], Yu Su[2], Jonathan Tremblay[3], Chan Hee Song[3,†,‡], and Jaesik Park[1,†]

[1] Seoul National University
[2] The Ohio State University
[3] NVIDIA

n, {cheolhong.min, jaesik.park}@snu.ac.kr

Luke Song • 2.
Research @NVIDIA Metropolis | Spatial Intelligence in VLMs
2 Tag(e) • Bearbeitet •

Do Vision-Language Models actually understand 3D space?

Do Vision–Language Models Understand 3D Scenes or Just Catalogue Objects?

[cs.CV] 19 May 2026

Visual-Language Nav, but why SLAM?

- **Is Perception Foundation Model Still Necessary?**

GPT 5.5 thinking

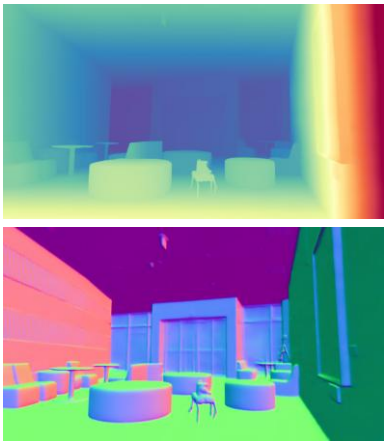
몇 초 동안 생각함 >

전방에 보이는 **건물은 1개**로 보여.

Wang, Ruicheng, et al. "Moge-2: Accurate monocular geometry with metric scale and sharp details." Advances in Neural Information Processing Systems 38 (2026): 35928-35959.

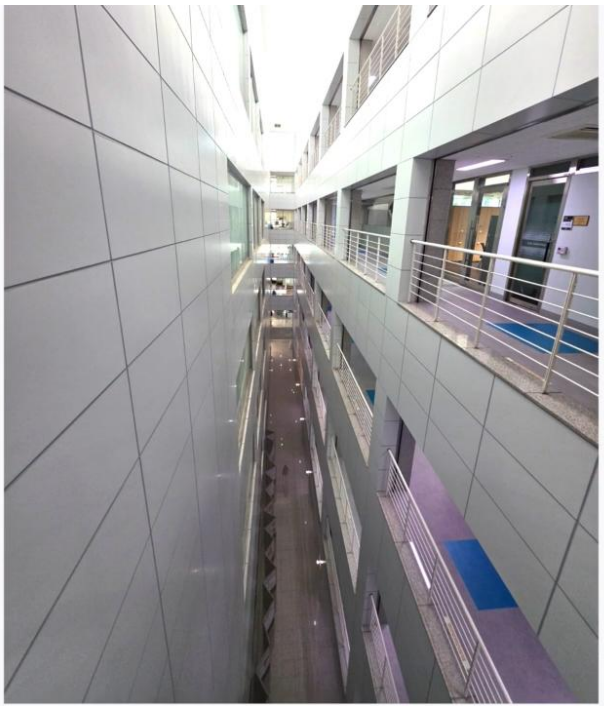
Visual-Language Nav, but why SLAM?

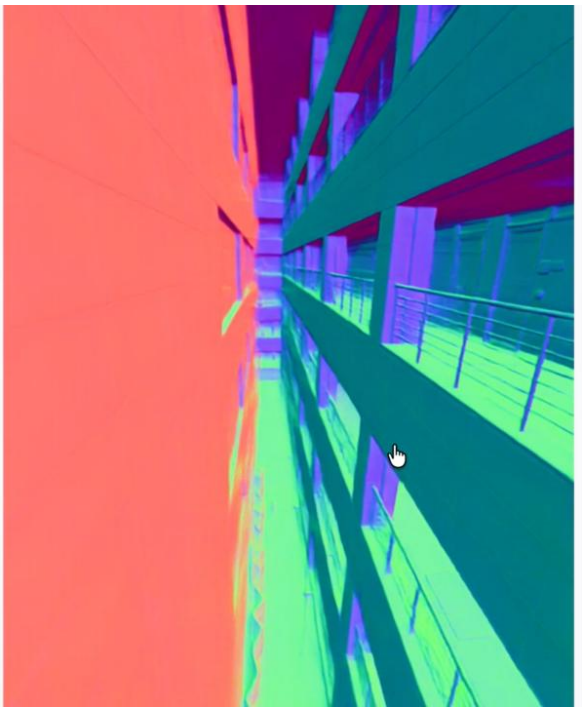
- Is Perception Foundation Model Still Necessary?



Visual-Language Nav, but why SLAM?

- Is Perception Foundation Model Still Necessary?



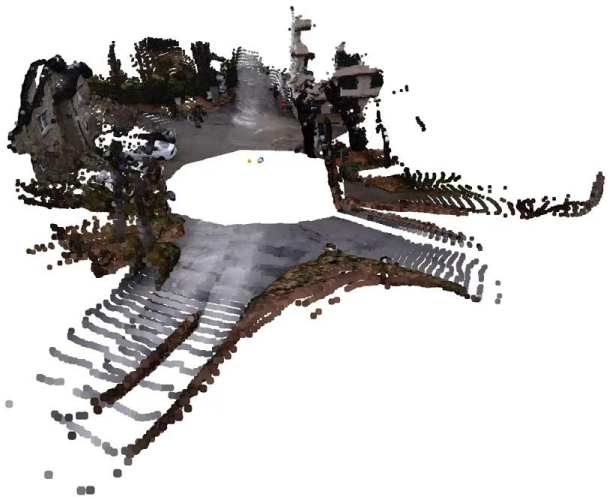


Visual-Language Nav, but why SLAM?

- Is Perception Foundation Model Still Necessary?
  - Multiple view consistency

Visual-Language Nav, but why SLAM?

- **Is Perception Foundation Model Still Necessary?**
  - **Multiple view consistency**

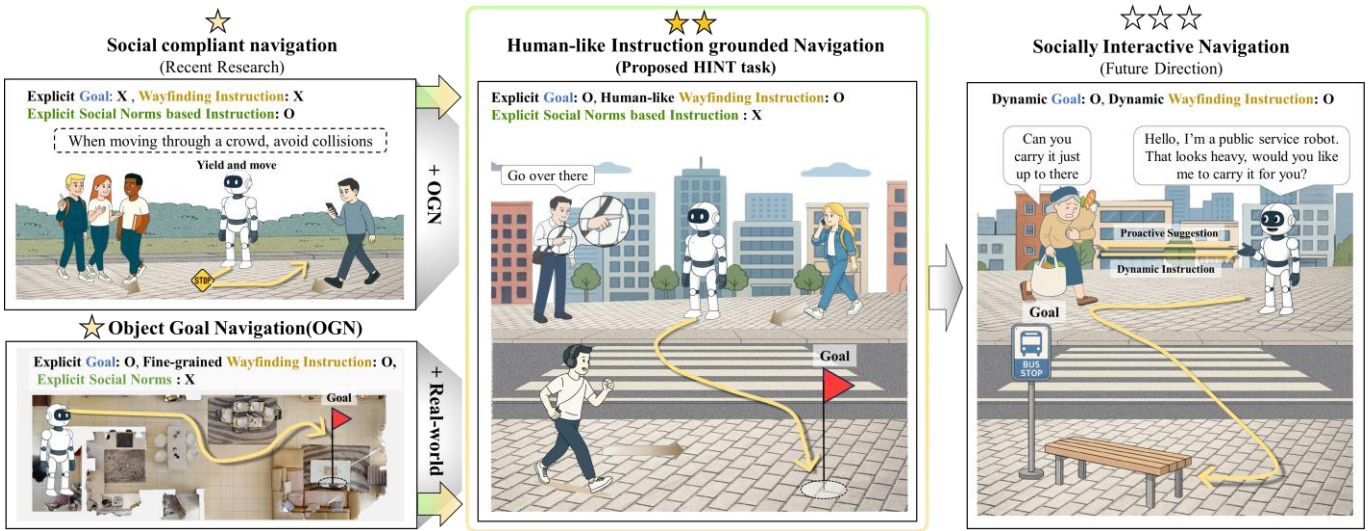


Zhang, Junyi, et al. "Loger: Long-context geometric reconstruction with hybrid memory." arXiv preprint arXiv:2603.03269 (2026).

Human-robot Interactive Navigation

Next 3 years Goal

- 암묵적 인간 지시를 이해하고 상호작용하는
  다중 자율 모빌리티를 위한 메모리 증강 공간지능 개발 (2026-2029)



Summary

Future Prediction
- Number of robots ↑
- Number of human–robot interactions ↑ ↑ ↑

Problem
- Diversity of robots, sensors, and configurations ↑
- Complexity, errors, and mission failures ↑ ↑ ↑

Solution
- Bridging Geometry and Reasoning gap
- **Visual Language Multi-robot SLAM (APRL's Method)**

Summary
- Help robots understand human language better.
- **Human-robot Interactive Navigation (APRL's Task)**

Closing: Language for Interaction

- How can this guy avoid getting fired? (이 친구가 짤리지 않으려면?)

Thank you!