CS586 (25 Spring) : Student Lecture

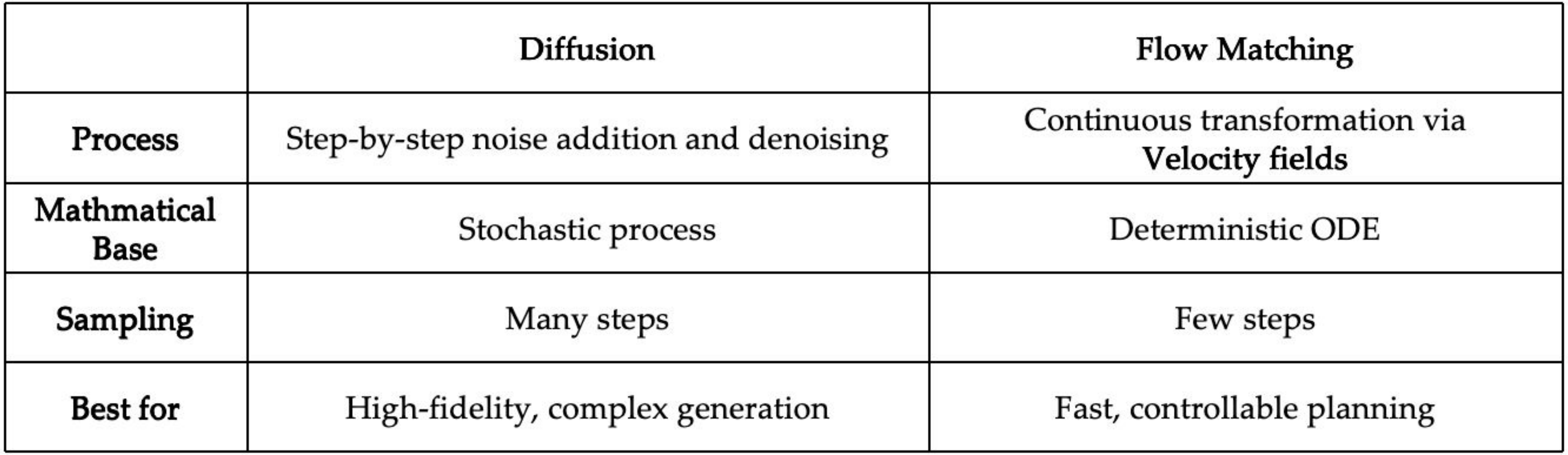
Vision-Language Model (VLM) for embodied navigation

20244076 Zhaoyan Wang

20244050 Xiangchen Liu

Review

Navigation with Probabilistic Generative Models: About Efficient Conditioning and its Architectures



	Diffusion	**Flow Matching**
Process	Step-by-step noise addition and denoising	Continuous transformation via **Velocity fields**
Mathmatical Base	Stochastic process	Deterministic ODE
Sampling	Many steps	Few steps
Best for	High-fidelity, complex generation	Fast, controllable planning

Embodied Navigation

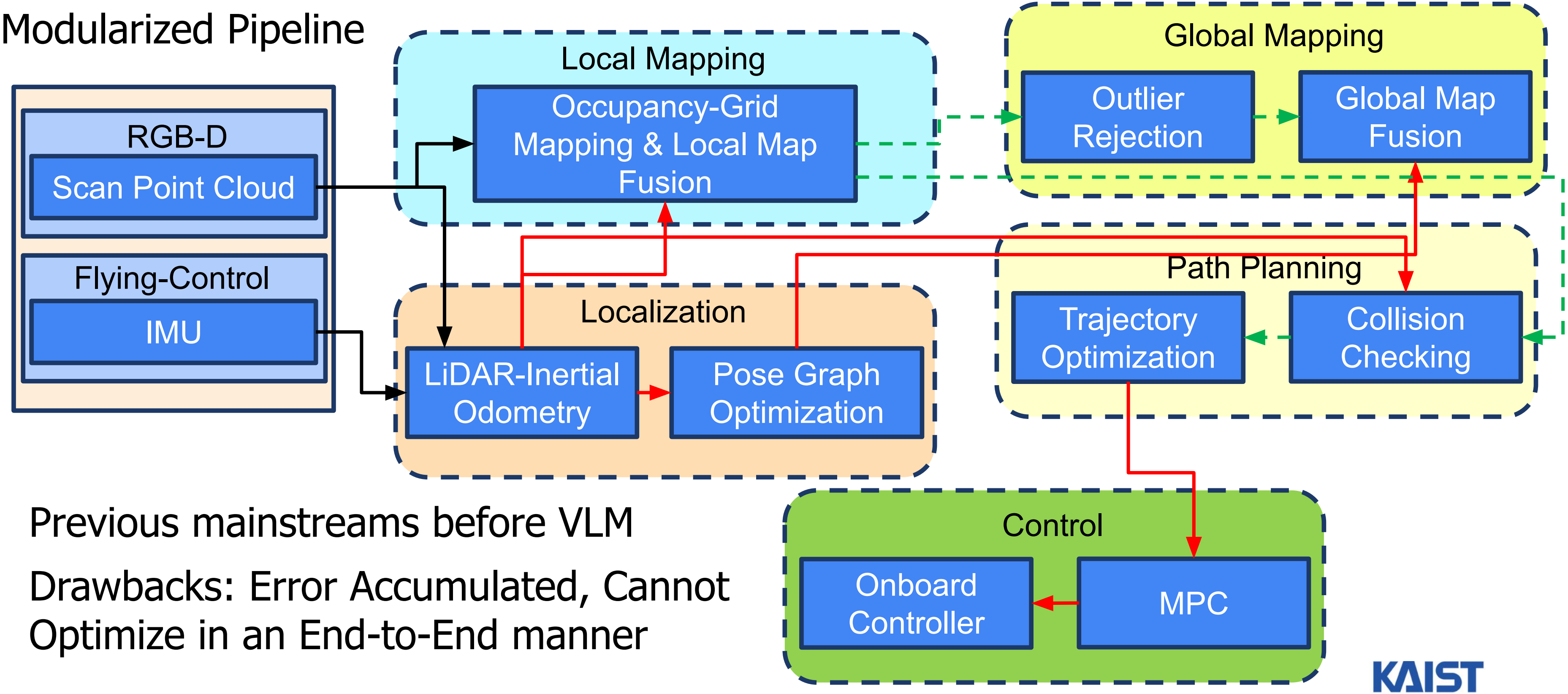
Task definition

Embodied Navigation: A physically agent autonomously accomplish a user-specified task in a 3-D environment — **without** access to a **privileged global map**.

The agent perceives the world only through its egocentric sensors (RGB-D, LiDAR, etc.), maintains internal memory, and decides a sequence of actions (e.g., turn ↔, move ↕, or Chain-of-Thought) that bring it finish the task.

Previous Embodied Navigation

Modularized Pipeline

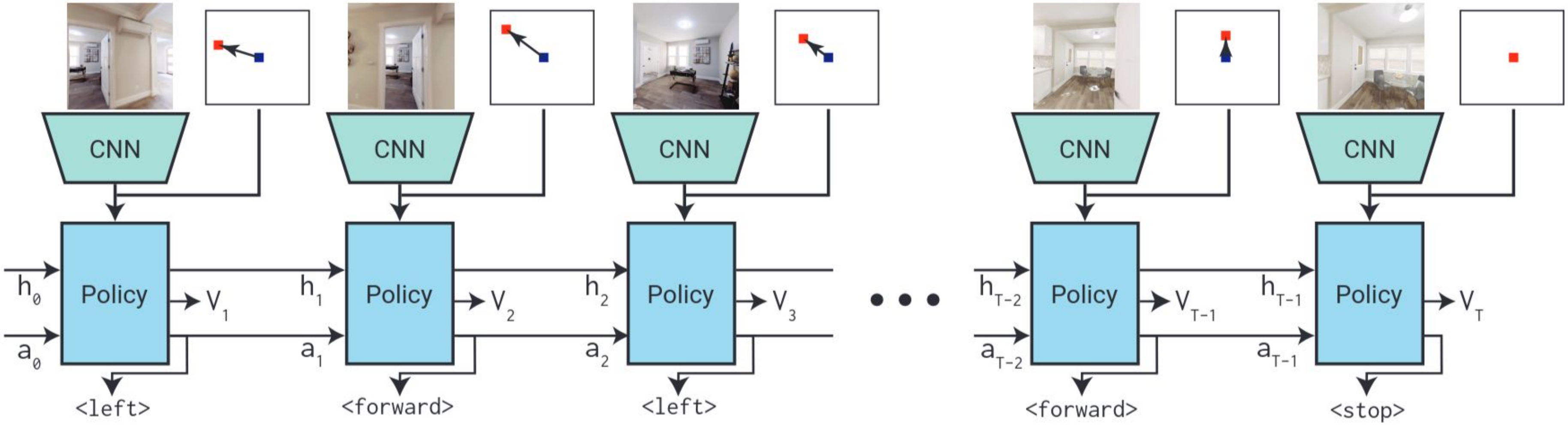


Previous mainstreams before VLM

Drawbacks: Error Accumulated, Cannot Optimize in an End-to-End manner

Previous Embodied Navigation

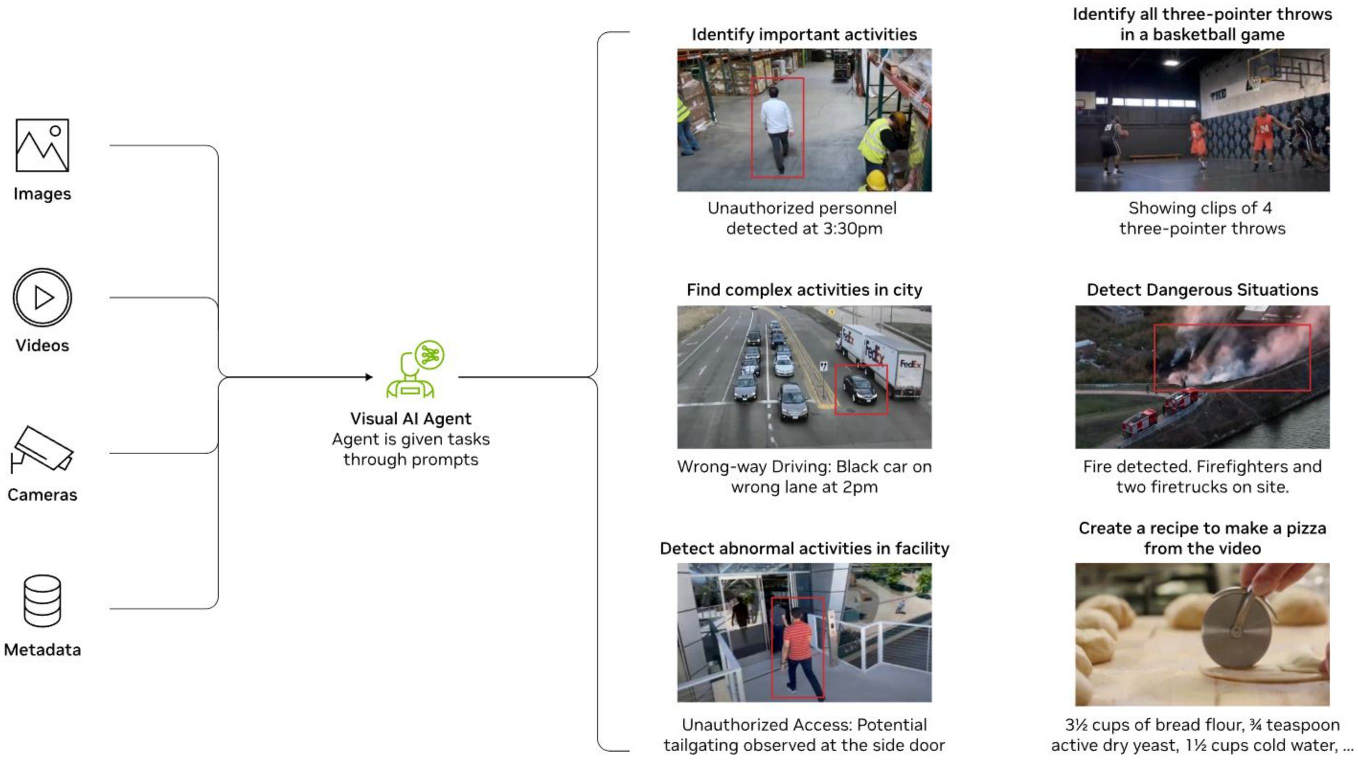
Reinforcement Learning-based Navigation



Drawbacks: Low Sample Efficiency, Poor Generalization Ability, Large Sim2Real Gap

Wijmans, Erik, et al. "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames." ICLR 2020.

Vision-Language Model (VLM)



Images

Videos

Cameras

Metadata

Visual AI Agent
Agent is given tasks
through prompts

Identify important activities
Unauthorized personnel
detected at 3:30pm

Find complex activities in city
Wrong-way Driving: Black car on
wrong lane at 2pm

Detect abnormal activities in facility
Unauthorized Access: Potential
tailgating observed at the side door

**Identify all three-pointer throws
in a basketball game**
Showing clips of 4
three-pointer throws

Detect Dangerous Situations
Fire detected. Firefighters and
two firetrucks on site.

**Create a recipe to make a pizza
from the video**
3½ cups of bread flour, ¾ teaspoon
active dry yeast, 1½ cups cold water, ...

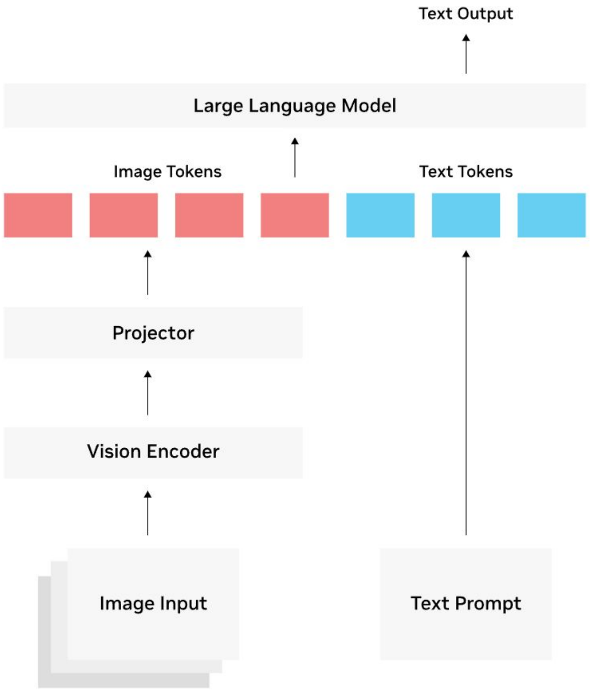
VLMs:

Multimodal AI systems built by combining a LLM with a vision encoder, giving the LLM the ability to “**see.**”

With this ability, VLMs can process and provide advanced understanding of **video**, **image**, and **text** inputs supplied in the prompt to generate text responses.

https://www.nvidia.com/en-us/glossary/vision-language-models/

Vision-Language Model (VLM)



Text Output

Large Language Model

Image Tokens

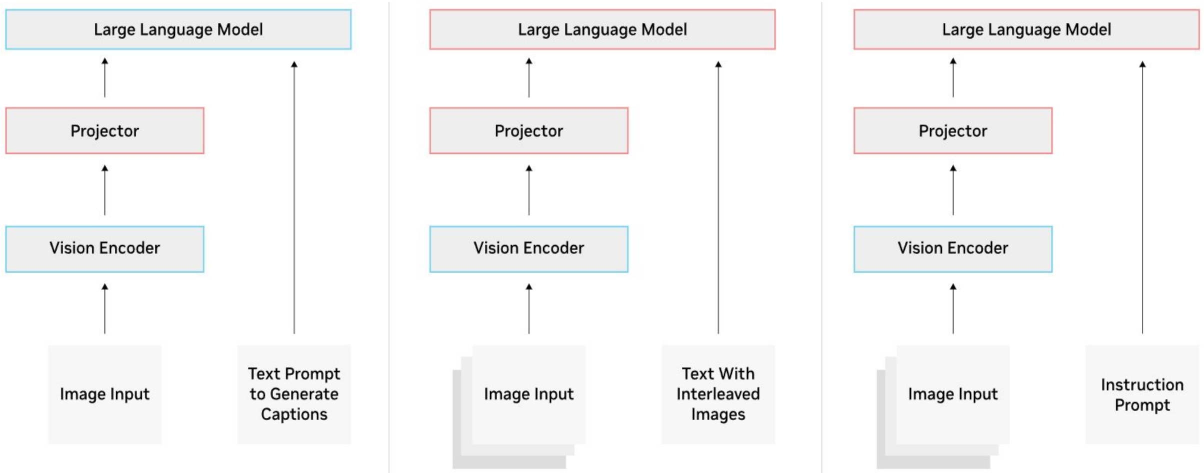
Text Tokens

Projector

Vision Encoder

Image Input

Text Prompt



Large Language Model

Projector

Vision Encoder

Image Input

Text Prompt to Generate Captions

Stage 1: Projector Initialization
Trained on image caption pairs

Large Language Model

Projector

Vision Encoder

Image Input

Text With Interleaved Images

Stage 2: Visual-Language Pre-Training
Trained on interleaved image-text and image-text pairs

Large Language Model

Projector

Vision Encoder

Image Input

Instruction Prompt

Stage 3: Visual Instruction Tuning
Trained on visual language instructions

🟥 The layers of the VLM that are updated during the training stage

🟦 The layers of the VLM that are not updated during the training stage

https://www.nvidia.com/en-us/glossary/vision-language-models/

VLM4Navigation

Navigation before VLM

1. Reinforcement Learning or Imitation Learning
2. No Language Grounding ability
3. Lack of open-world generalization ability
4. Large Sim2Real gap

What VLM brings

1. Vision-Language Efficient Fusion
2. Improved Semantic Understanding ability
3. Significantly improved zero-shot ability
4. Cross-Task transfer become available

VLM4Navigation

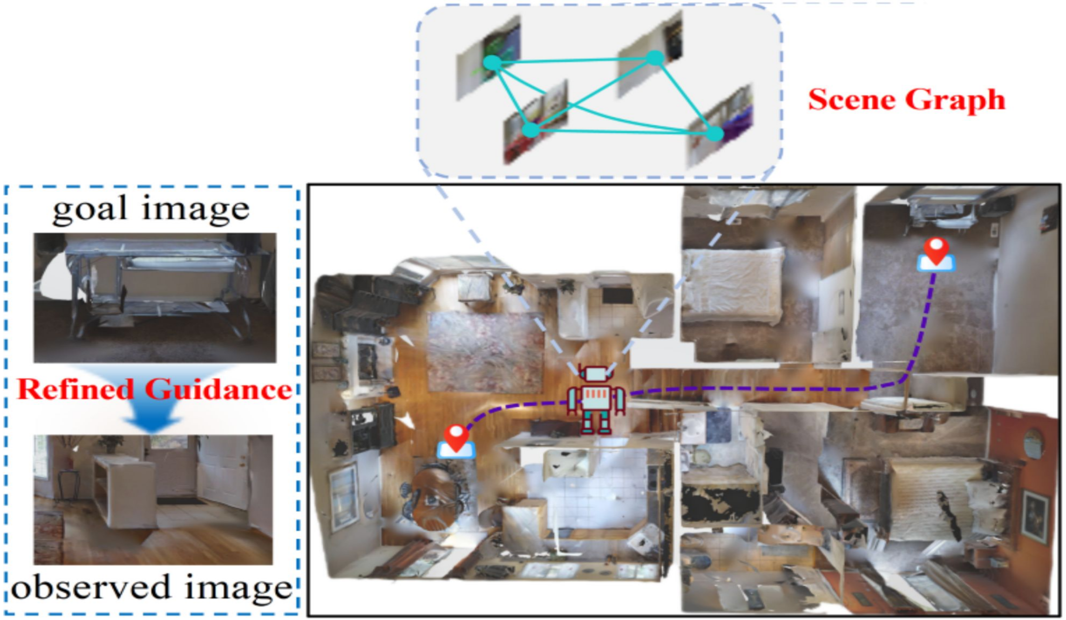
four core task categories:

- Image-goal navigation
- Embodied Question Answering (EQA)
- Vision-Language navigation (VLN)
- Object-goal navigation

Categorized by Goal Specification Modality

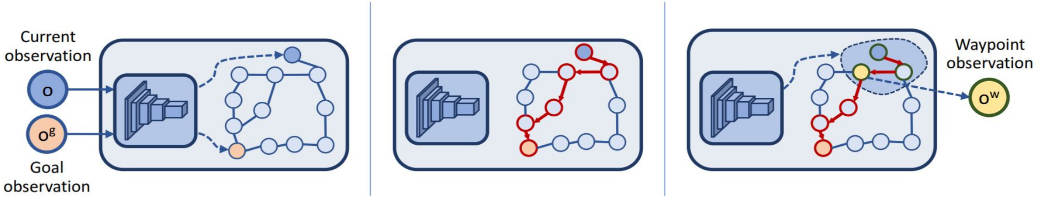
Image-goal Navigation

"Given an image of a target, an embodied agent must navigate in a 3D environment to reach that location."



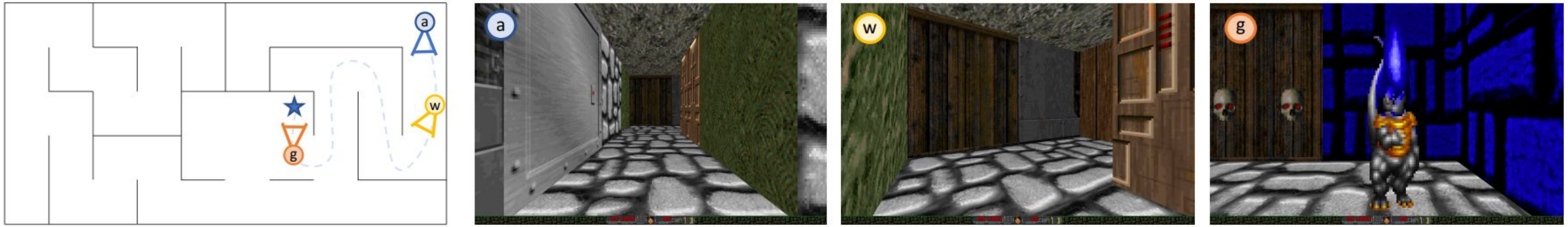
A paradigm shift in navigation, aiming to search targets without precise positional information, but relying solely on a goal image.

Image-goal Navigation

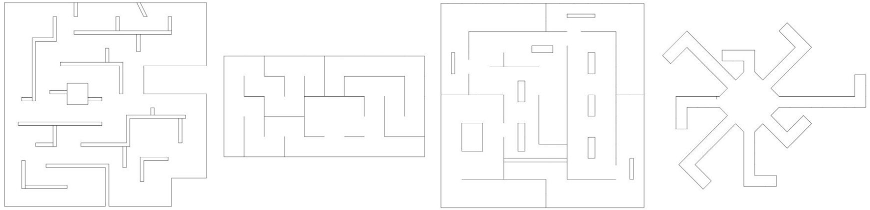


(a) Localization (b) Planning (c) Waypoint selection

TOPOLOGICAL MEMORY



(a) Maze (b) Agent's observation (c) Waypoint observation (d) Goal observation

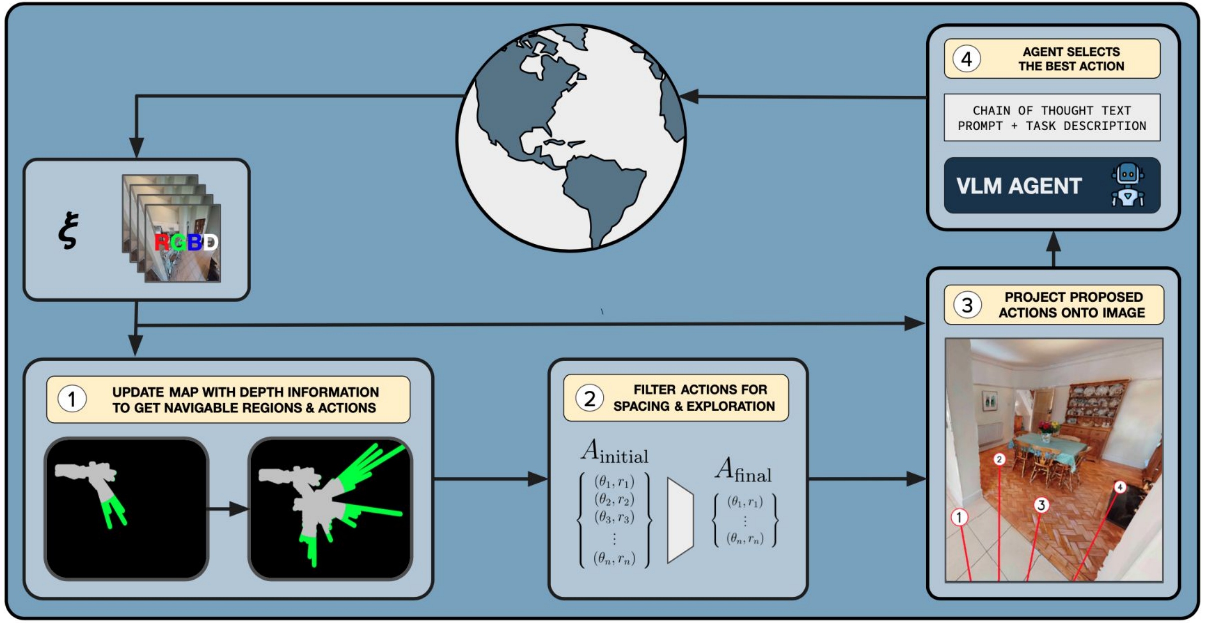


Train Test-1 Test-2 Test-3

Savinov, Nikolay, Alexey Dosovitskiy, and Vladlen Koltun. "Semi-parametric topological memory for navigation." ICLR 2018.

Image-goal Navigation

Apply **VLM** directly to action decision-making in embodied navigation.
Reformulate the navigation problem into image-based question answering.



ξ

RGBD

① UPDATE MAP WITH DEPTH INFORMATION TO GET NAVIGABLE REGIONS & ACTIONS

② FILTER ACTIONS FOR SPACING & EXPLORATION

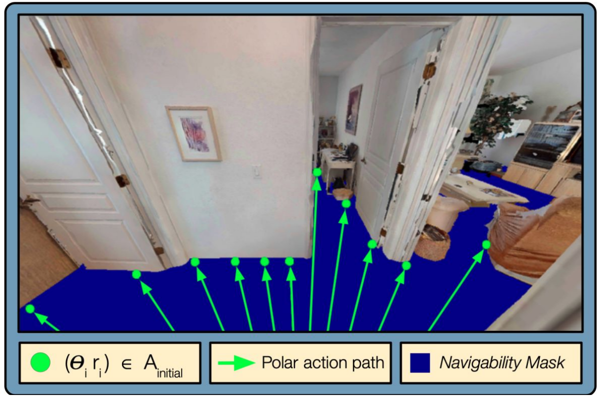
A_{initial} $\begin{Bmatrix} (\theta_1, r_1) \\ (\theta_2, r_2) \\ (\theta_3, r_3) \\ \vdots \\ (\theta_n, r_n) \end{Bmatrix}$ A_{final} $\begin{Bmatrix} (\theta_1, r_1) \\ \vdots \\ (\theta_n, r_n) \end{Bmatrix}$

③ PROJECT PROPOSED ACTIONS ONTO IMAGE

④ AGENT SELECTS THE BEST ACTION

CHAIN OF THOUGHT TEXT PROMPT + TASK DESCRIPTION

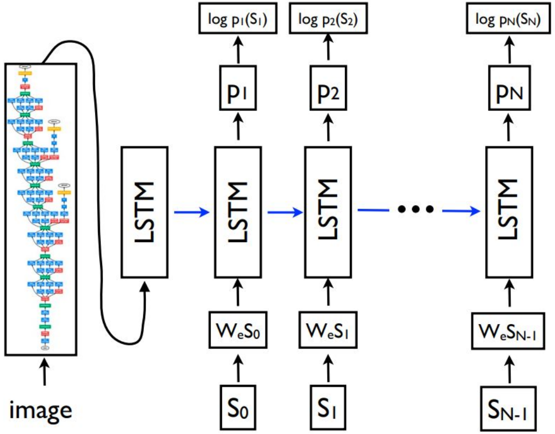
VLM AGENT

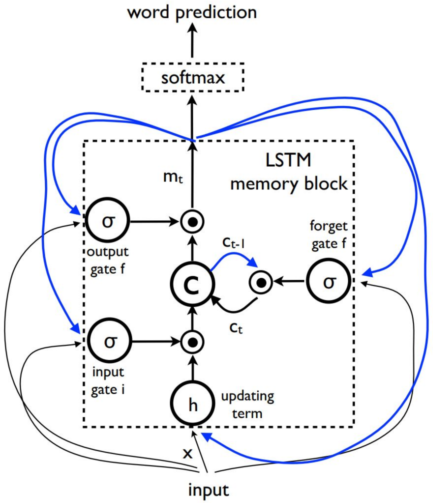


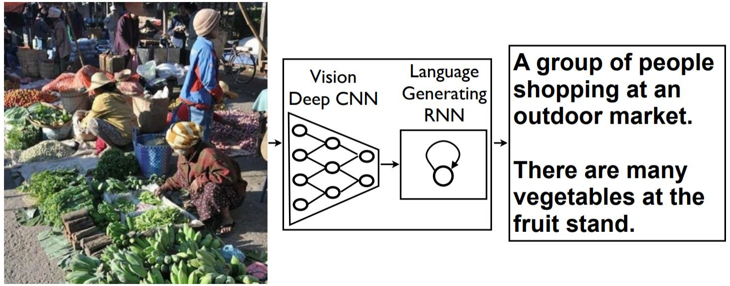
Goetting, Dylan, Himanshu Gaurav Singh, and Antonio Loquercio. "End-to-End Navigation with Vision Language Models: Transforming Spatial Reasoning into Question-Answering." *arXiv preprint arXiv:2411.05755* (2024).

Embodied Question Answering

"What color is the car parked next to the tree?."





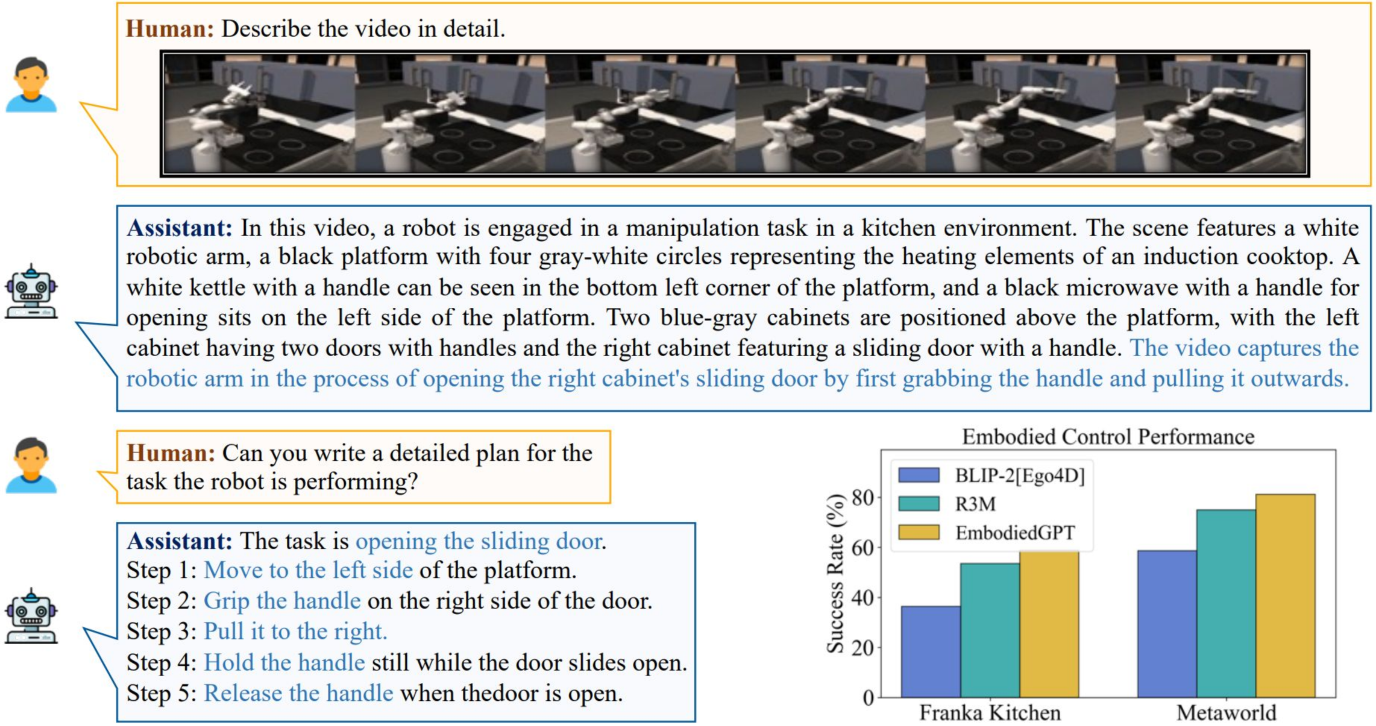


Vision Deep CNN
Language Generating RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

It requires an embodied agent (usually a virtual or physical robot) to navigate and perceive within an environment in order to answer a question.

Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Embodied Question Answering



Human: Describe the video in detail.

Assistant: In this video, a robot is engaged in a manipulation task in a kitchen environment. The scene features a white robotic arm, a black platform with four gray-white circles representing the heating elements of an induction cooktop. A white kettle with a handle can be seen in the bottom left corner of the platform, and a black microwave with a handle for opening sits on the left side of the platform. Two blue-gray cabinets are positioned above the platform, with the left cabinet having two doors with handles and the right cabinet featuring a sliding door with a handle. The video captures the robotic arm in the process of opening the right cabinet's sliding door by first grabbing the handle and pulling it outwards.

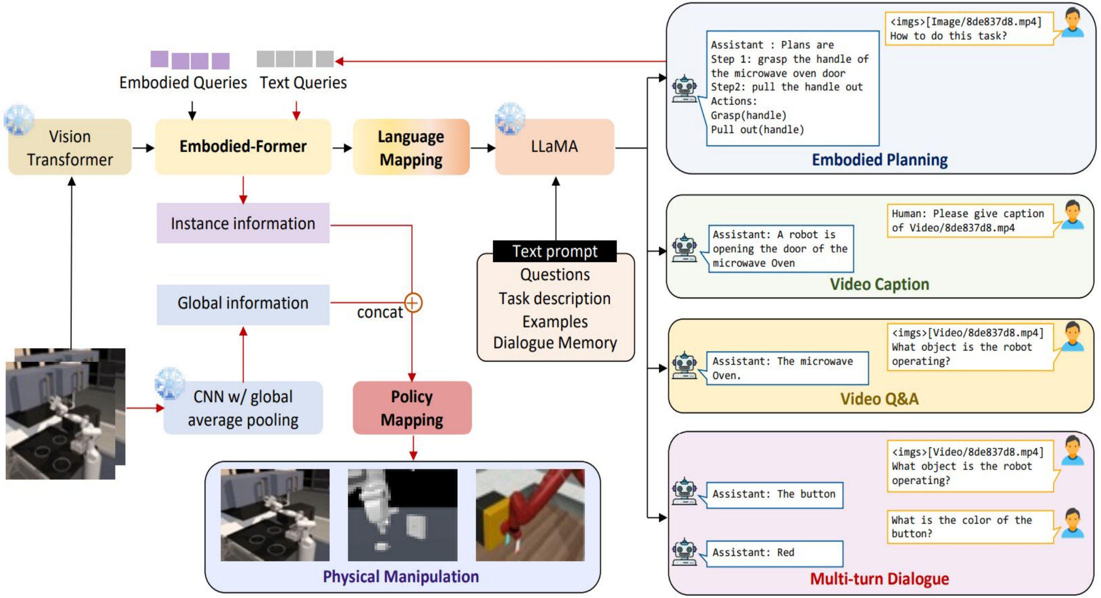
Human: Can you write a detailed plan for the task the robot is performing?

Assistant: The task is opening the sliding door.
Step 1: Move to the left side of the platform.
Step 2: Grip the handle on the right side of the door.
Step 3: Pull it to the right.
Step 4: Hold the handle still while the door slides open.
Step 5: Release the handle when thedoor is open.

EmbodiedGPT *:*
*Through the training paradigm of "**Embodied CoT** ", VLM learns to generate structured task plans from visual inputs and establishes a closed-loop system from high-level planning to low-level action execution.*

Mu, Yao, et al. "Embodiedgpt: Vision-language pre-training via embodied chain of thought." Advances in Neural Information Processing Systems 36 (2023): 25081-25094.

Embodied Question Answering



<imgs>[Image/8de837d8.mp4] How to do this task?

Assistant : Plans are
Step 1: grasp the handle of the microwave oven door
Step2: pull the handle out
Actions:
Grasp(handle)
Pull out(handle)

Human: Please give caption of Video/8de837d8.mp4

Assistant: A robot is opening the door of the microwave Oven

<imgs>[Video/8de837d8.mp4] What object is the robot operating?

Assistant: The microwave Oven.

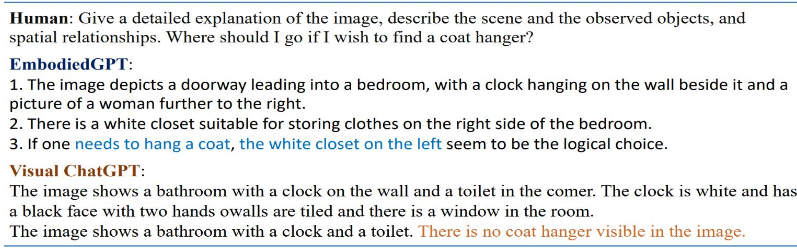
<imgs>[Video/8de837d8.mp4] What object is the robot operating?

Assistant: The button

What is the color of the button?

Assistant: Red

Overall framework of EmbodiedGPT



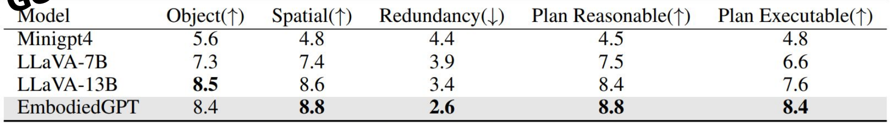
Human: Give a detailed explanation of the image, describe the scene and the observed objects, and spatial relationships. Where should I go if I wish to find a coat hanger?

EmbodiedGPT:
1. The image depicts a doorway leading into a bedroom, with a clock hanging on the wall beside it and a picture of a woman further to the right.
2. There is a white closet suitable for storing clothes on the right side of the bedroom.
3. If one needs to hang a coat, the white closet on the left seem to be the logical choice.

Visual ChatGPT:
The image shows a bathroom with a clock on the wall and a toilet in the comer. The clock is white and has a black face with two hands owalls are tiled and there is a window in the room.
The image shows a bathroom with a clock and a toilet. There is no coat hanger visible in the image.

Comparison between EmbodiedGPT and VisualGPT in the question-answering task

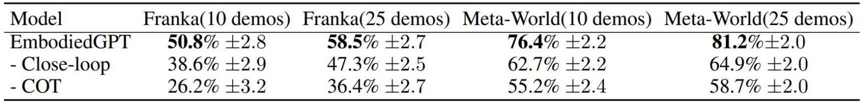
Caption Generation



Model	Object(↑)	Spatial(↑)	Redundancy(↓)	Plan Reasonable(↑)	Plan Executable(↑)
Minigpt4	5.6	4.8	4.4	4.5	4.8
LLaVA-7B	7.3	7.4	3.9	7.5	6.6
LLaVA-13B	**8.5**	8.6	3.4	8.4	7.6
EmbodiedGPT	8.4	**8.8**	**2.6**	**8.8**	**8.4**

Table 1: Generate Quality Evaluation on image input tasks.

Evaluation on embodied control tasks

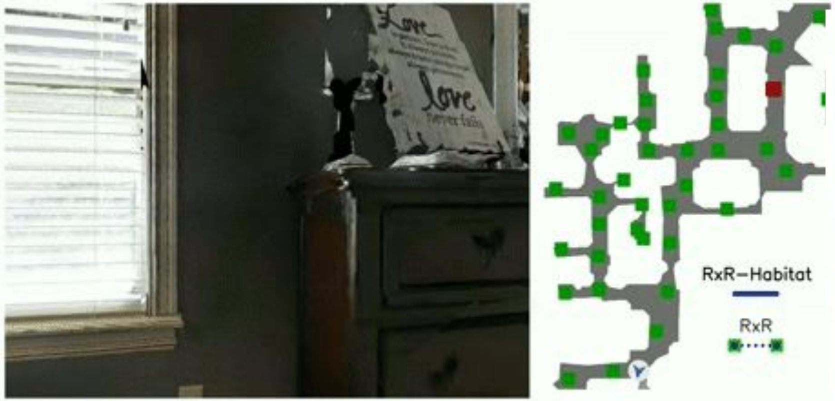


Model	Franka(10 demos)	Franka(25 demos)	Meta-World(10 demos)	Meta-World(25 demos)
EmbodiedGPT	**50.8**% ±2.8	**58.5**% ±2.7	**76.4**% ±2.2	**81.2**%±2.0
- Close-loop	38.6% ±2.9	47.3% ±2.5	62.7% ±2.2	64.9% ±2.0
- COT	26.2% ±3.2	36.4% ±2.7	55.2% ±2.4	58.7% ±2.0

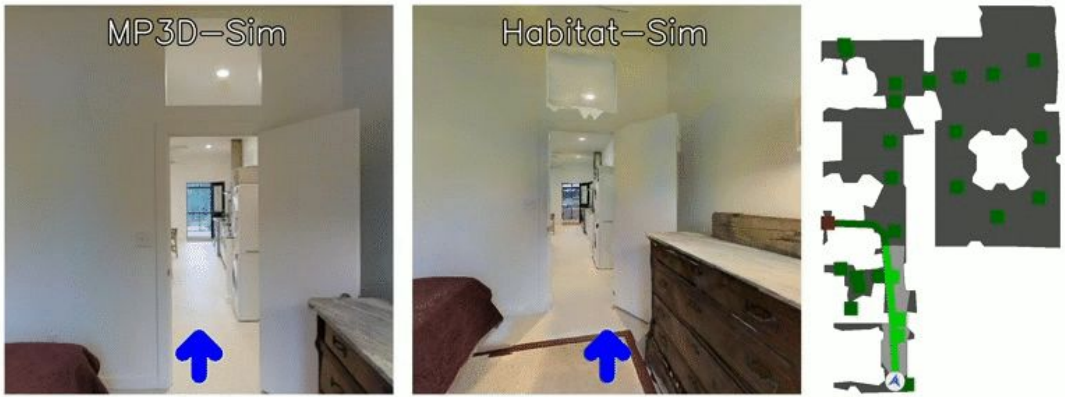
Table 2: Ablation on the closed-loop spans from planning to low-level control, and "chain-of-thought"

Vision-Language Navigation

"Given the egocentric image observation sequence with corresponding language instruction as input, following the text instruction and reach out to the target area."



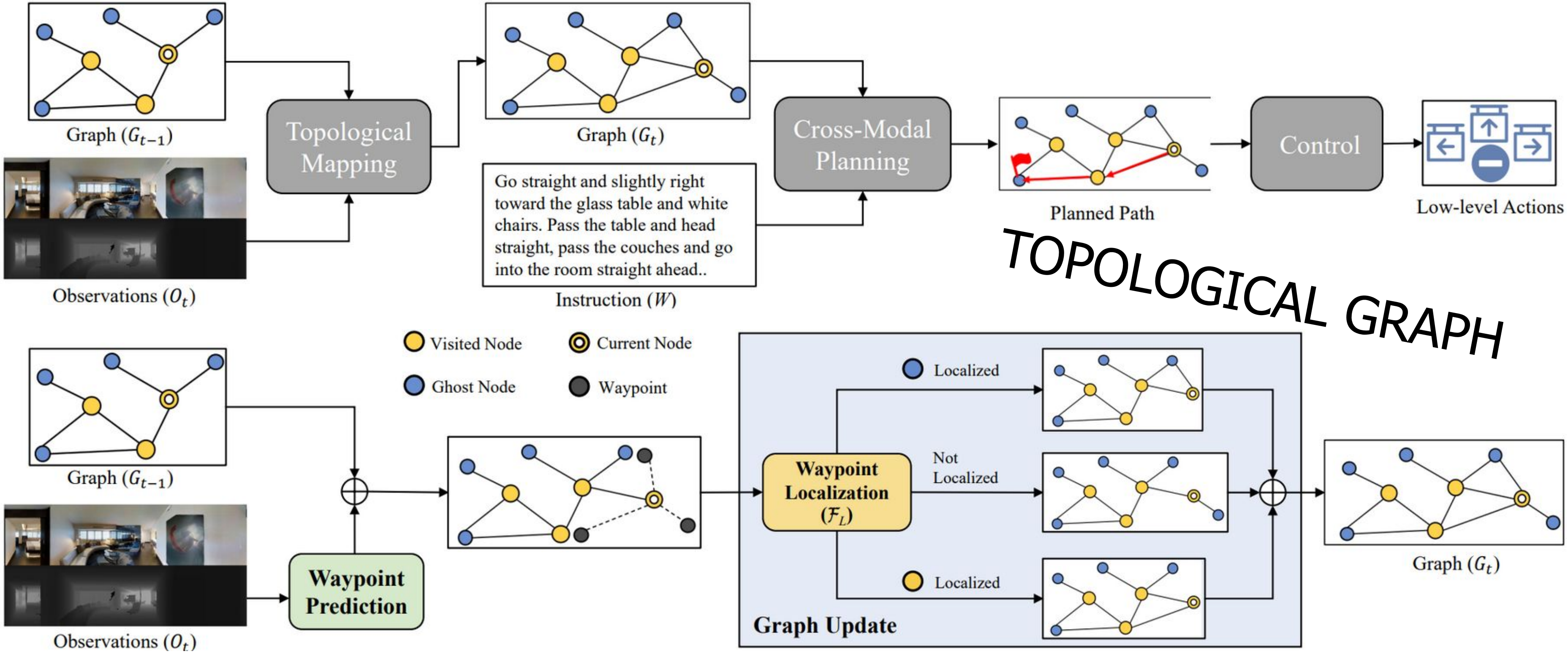
You are in a bedroom. Turn around to the left until you see a door leading out into a hallway, go through it. Hang a right and walk between the island and the couch on your left. When you are between the second and third chairs for the island stop.



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

https://jacobkrantz.github.io/vlnce/

Vision-Language Navigation



Graph (G_{t-1})

Observations (O_t)

Topological Mapping

Graph (G_t)

Go straight and slightly right toward the glass table and white chairs. Pass the table and head straight, pass the couches and go into the room straight ahead..

Instruction (W)

Cross-Modal Planning

Planned Path

Control

Low-level Actions

TOPOLOGICAL GRAPH

Visited Node

Current Node

Ghost Node

Waypoint

Graph (G_{t-1})

Observations (O_t)

Waypoint Prediction

Waypoint Localization ($\mathcal{F}_L$)

Localized

Not Localized

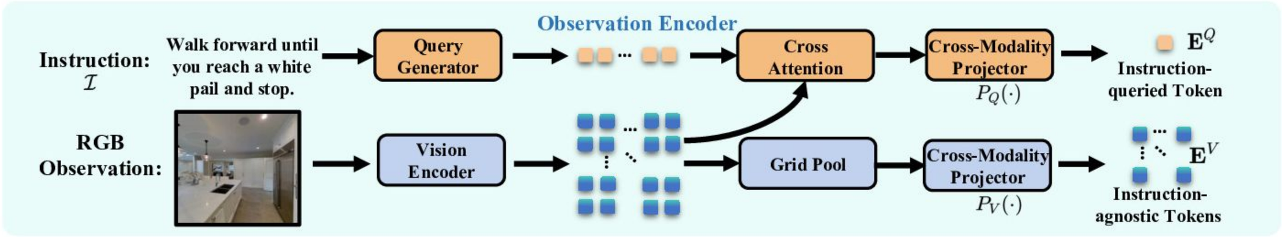
Localized

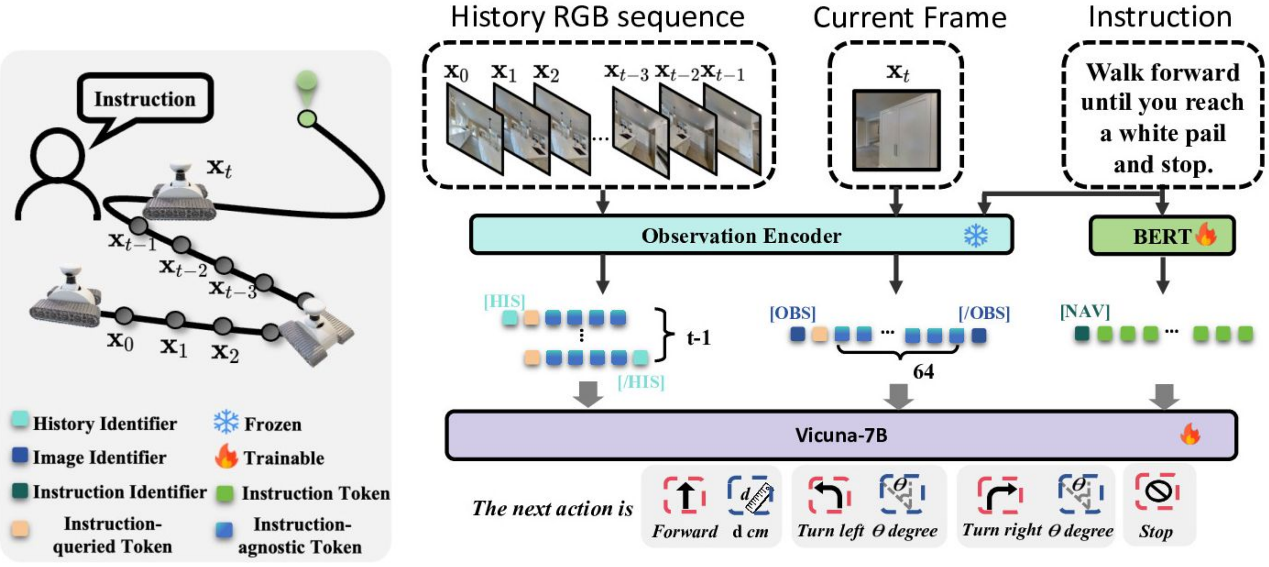
Graph Update

Graph (G_t)

An, et. al. “Evolving Topological Planning for Vision-Language Navigation in Continuous Environment”, TPAMI 2024

Vision-Language Navigation





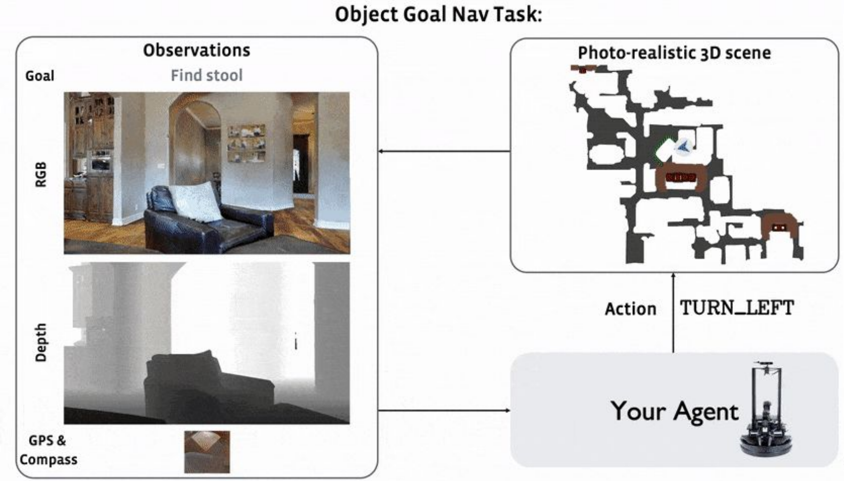
Using only RGB video as input and directly output low-level actions from a video-based vision-language-model

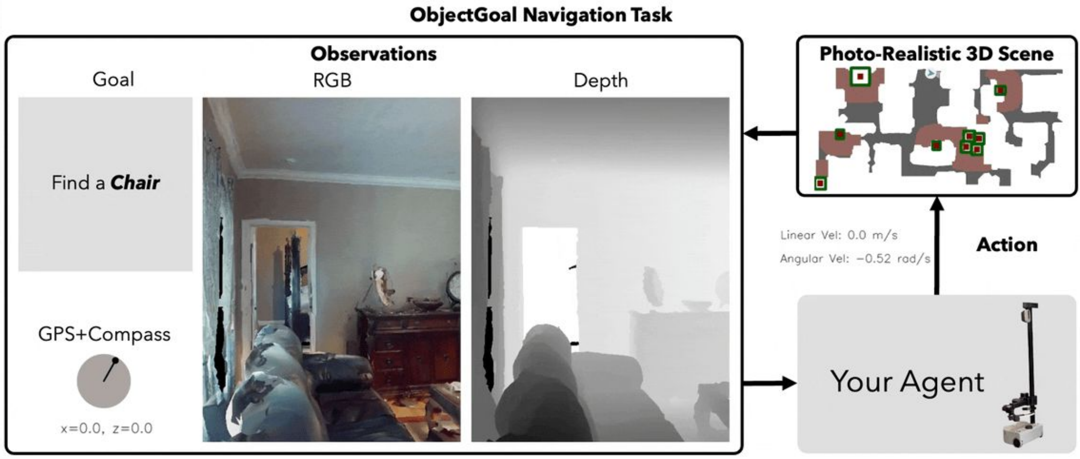
No need for translation, rotation, and depth information!

Zhang, et al. "NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation." RSS 2024.

Object-goal Navigation

"Given an object category, the embodied agent must navigate to the specified object in unknown 3D environment."

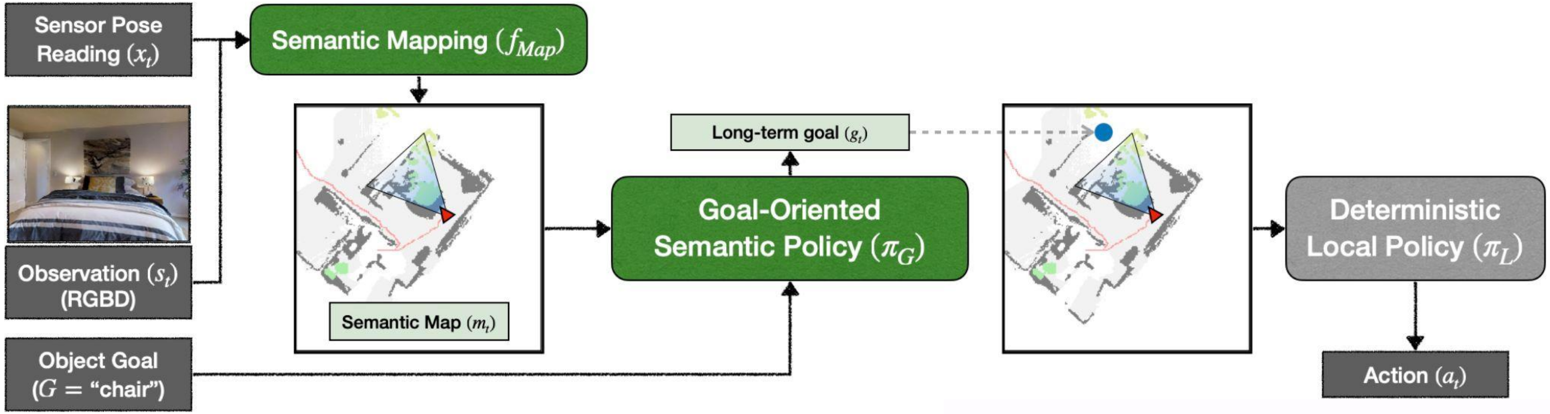




Find an instance of an object category ('find a chair') by navigating to it, no prior map information, only use its onboard sensory input to explore and navigate.

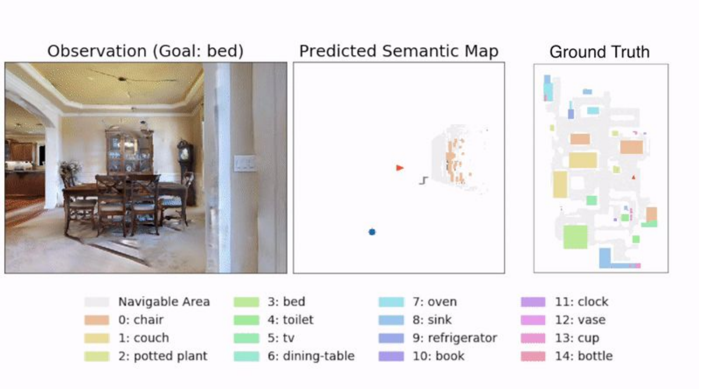
https://aihabitat.org/challenge/2023/

Object-goal Navigation


Sensor Pose Reading (x_t)
Semantic Mapping (f_{Map})
Observation (s_t) (RGBD)
Semantic Map (m_t)
Long-term goal (g_t)
Goal-Oriented Semantic Policy (π_G)
Deterministic Local Policy (π_L)
Object Goal $(G =$ "chair"$)$
Action (a_t)

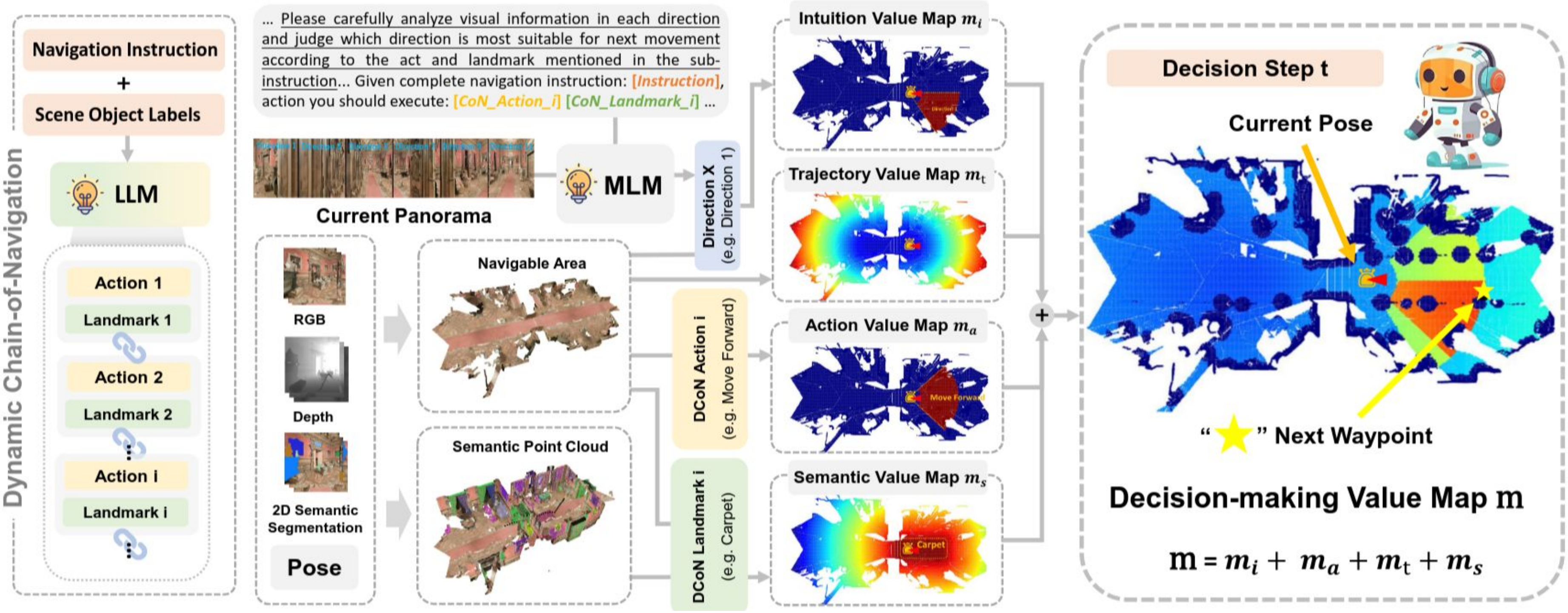
Semantic map-based Exploration

1. Lack of Zero-shot transfer ability
2. Limited categories of goal objects
3. Time-consuming training in Simulator
4. Sim-2-Real Gap



Chaplot, D.S.et. al. "Object Goal Navigation using Goal-oriented Semantic Exploration", NeuIPS 2020

Object-goal Navigation



... Please carefully analyze visual information in each direction and judge which direction is most suitable for next movement according to the act and landmark mentioned in the sub-instruction... Given complete navigation instruction: [*Instruction*], action you should execute: [*CoN_Action_i*] [*CoN_Landmark_i*] ...

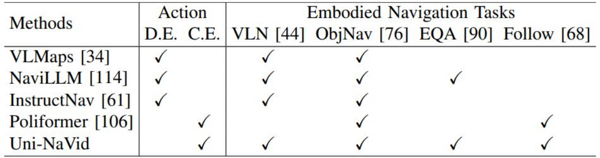
$$\mathrm{m} = m_i + m_a + m_t + m_s$$

Using VLM to judge navigation directions, use LLM to plan Dynamic Chain-of-Navigation

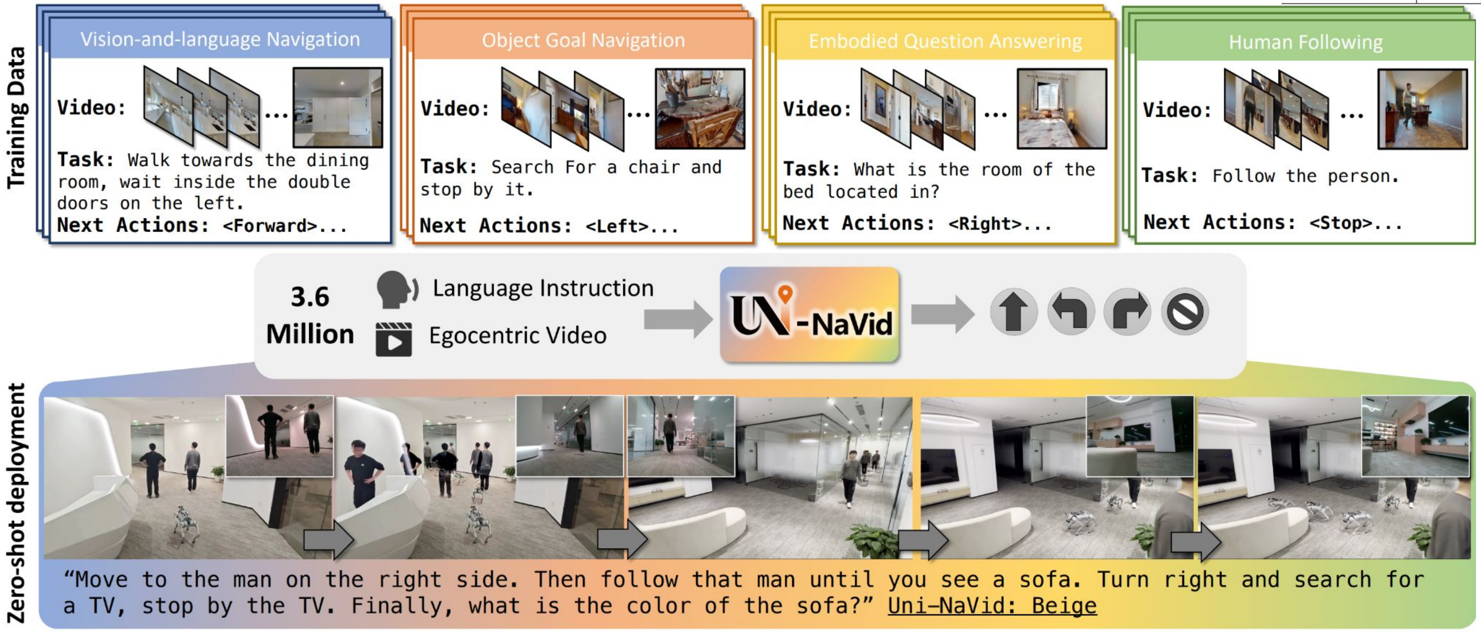
Long, et. al. “InstructNav: Zero-shot System for Generic Instruction Navigation in Unexplored Environment”, CoRL 2024

Unified Frameworks for Embodied AI

PKU Uni-Navid : Unifying Embodied Navigation Tasks



Methods	Action		Embodied Navigation Tasks			
	D.E.	C.E.	VLN [44]	ObjNav [76]	EQA [90]	Follow [68]
VLMaps [34]	✓		✓	✓		
NaviLLM [114]	✓		✓	✓	✓	
InstructNav [61]	✓		✓	✓		
Poliformer [106]		✓		✓		✓
Uni-NaVid		✓	✓	✓	✓	✓



Task: Walk towards the dining room, wait inside the double doors on the left.
Next Actions: <Forward>...

Task: Search For a chair and stop by it.
Next Actions: <Left>...

Task: What is the room of the bed located in?
Next Actions: <Right>...

Task: Follow the person.
Next Actions: <Stop>...

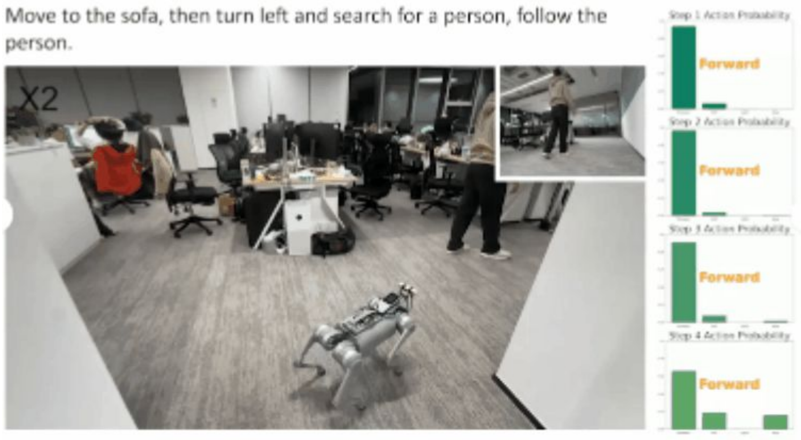
“Move to the man on the right side. Then follow that man until you see a sofa. Turn right and search for a TV, stop by the TV. Finally, what is the color of the sofa?” Uni-NaVid: Beige

Uni-NaVid learns general navigation skills across **four** embodied navigation tasks.

Zhang, Jiazhao, et al. "Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks." RSS 2025.

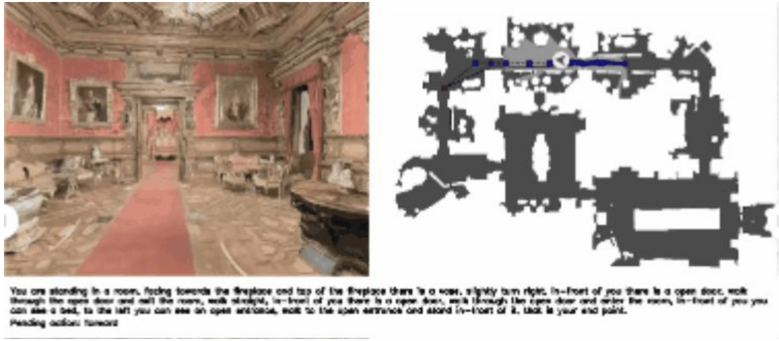
Unified Frameworks for Embodied AI

Compositional Navigation Task



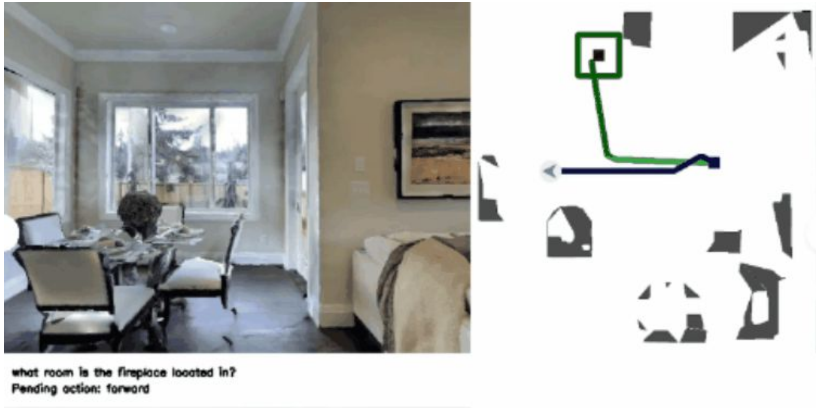
Move to the sofa, then turn left and search for a person, follow the person.

X2
Step 1 Action Probability
Forward
Step 2 Action Probability
Forward
Step 3 Action Probability
Forward
Step 4 Action Probability
Forward



You are standing in a room, facing towards the fireplace and top of the fireplace there is a vase, slightly turn right. In-front of you there is a open door, walk through the open door and exit the room, walk straight, in-front of you there is a open door, walk through the open door and enter the room, in-front of you you can see a bed, to the left you can see an open entrance, walk to the open entrance and stand in-front of it, that is your end point.
Pending action: forward

VLN Task



what room is the fireplace located in?
Pending action: forward

Embodied Question Answering Task

Zhang, Jiazhao, et al. "Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks." RSS 2025.

Conclusions

VLM enable the agent with super powerful temporal-spatial understanding ability. *"One ring to rule them all"* become possible. Using one VLM to unify all embodied navigation tasks will be the future research direction

Some good research directions for future research

1. Agent-Human-Interaction
2. Flying-wheel high-quality data augmentation
3. Cross-Floor and Long-Horizon inference

CS586 (25 Spring) : Student Lecture

Vision-Language Model for embodied navigation

Thank you.

20244076 Zhaoyan Wang
20244050 Xiangchen Liu

KAIST

Quiz

Quiz 1.
Visual-Language Models (VLMs) enable robots to jointly understand natural language instructions and visual scenes, thereby enhancing navigation tasks (True or False).

Quiz 2.
In VLMs-based Vision-Language-Navigation, the agent needs pose and depth information for training and inference (True or False)