

---

# Evaluating GPT-4o’s Embodied Intelligence: A Comprehensive Empirical Study

---

Yujie Wu<sup>1,†</sup>, Huaihai Lyu<sup>2,†</sup>, Yingbo Tang<sup>2,†</sup>, Lingfeng Zhang<sup>3,4,†</sup>

Zhihui Zhang<sup>5</sup>, Wei Zhou<sup>6</sup>, Siqi Hao<sup>3,♦</sup>

† Core Contributors ♦ Corresponding Author

<sup>1</sup> Nankai University

<sup>2</sup> Institute of Automation, Chinese Academy of Sciences

<sup>3</sup> Tsinghua University

<sup>4</sup> The Hong Kong University of Science and Technology (Guangzhou)

<sup>5</sup> Beijing Institute of Technology

<sup>6</sup> Cardiff University

## Abstract

This paper explores the embodied intelligence capabilities of GPT-4o, emphasizing its performance in key areas such as embodied reasoning, embodied manipulation, and embodied navigation. We are the first to systematically evaluate GPT-4o on embodied intelligence tasks. As an advanced language model, GPT-4o demonstrates significant potential in complex tasks by integrating visual and linguistic understanding, with particularly noteworthy abilities in image generation. In the **embodied reasoning** section, the study evaluates GPT-4o’s performance in perception, spatial reasoning, temporal reasoning, planning, and causal reasoning. The findings reveal how the model extracts information from images to make informed inferences. The **embodied manipulation** section analyzes its applications in object understanding, environmental perception, and task planning, showcasing the model’s adaptability in dynamic environments. Additionally, the **embodied navigation** section focuses on GPT-4o’s ability to process navigation instructions, perform map reasoning, infer trajectories, and predict actions, indicating its effectiveness in navigation tasks. In summary, GPT-4o shows remarkable advancements in the field of embodied intelligence. Future development plans aim to further enhance its capabilities in multimodal interactions and complex environmental adaptability, laying a foundation for achieving higher levels of artificial intelligence. Through ongoing research and optimization, GPT-4o is poised to play a greater role in the practical applications of embodied intelligence.

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Motivation and Overview . . . . .	3
1.2	Guidance . . . . .	4
<b>2</b>	<b>Embodied Reasoning</b>	<b>5</b>
2.1	Perceptual Reasoning . . . . .	6
2.2	Spatial Reasoning . . . . .	9
2.3	Temporal Reasoning . . . . .	14
2.4	Planning Reasoning . . . . .	17
2.5	Causal Reasoning . . . . .	20
<b>3</b>	<b>Embodied Manipulation</b>	<b>22</b>
3.1	Compositional Object Understanding . . . . .	22
3.2	Affordance Grounding . . . . .	26
3.3	Task Planning . . . . .	33
3.4	Trajectory Prediction . . . . .	38
<b>4</b>	<b>Embodied Navigation</b>	<b>43</b>
4.1	Navigation Instruction Reasoning . . . . .	43
4.2	Navigation Map Reasoning . . . . .	49
4.3	Navigation Trajectory Inference . . . . .	54
4.4	Navigation Action Prediction . . . . .	59
<b>5</b>	<b>Conclusions</b>	<b>64</b>
5.1	Capabilities and Limitations of GPT-4o in Embodied Intelligence . . . . .	64
5.2	Future Work . . . . .	65

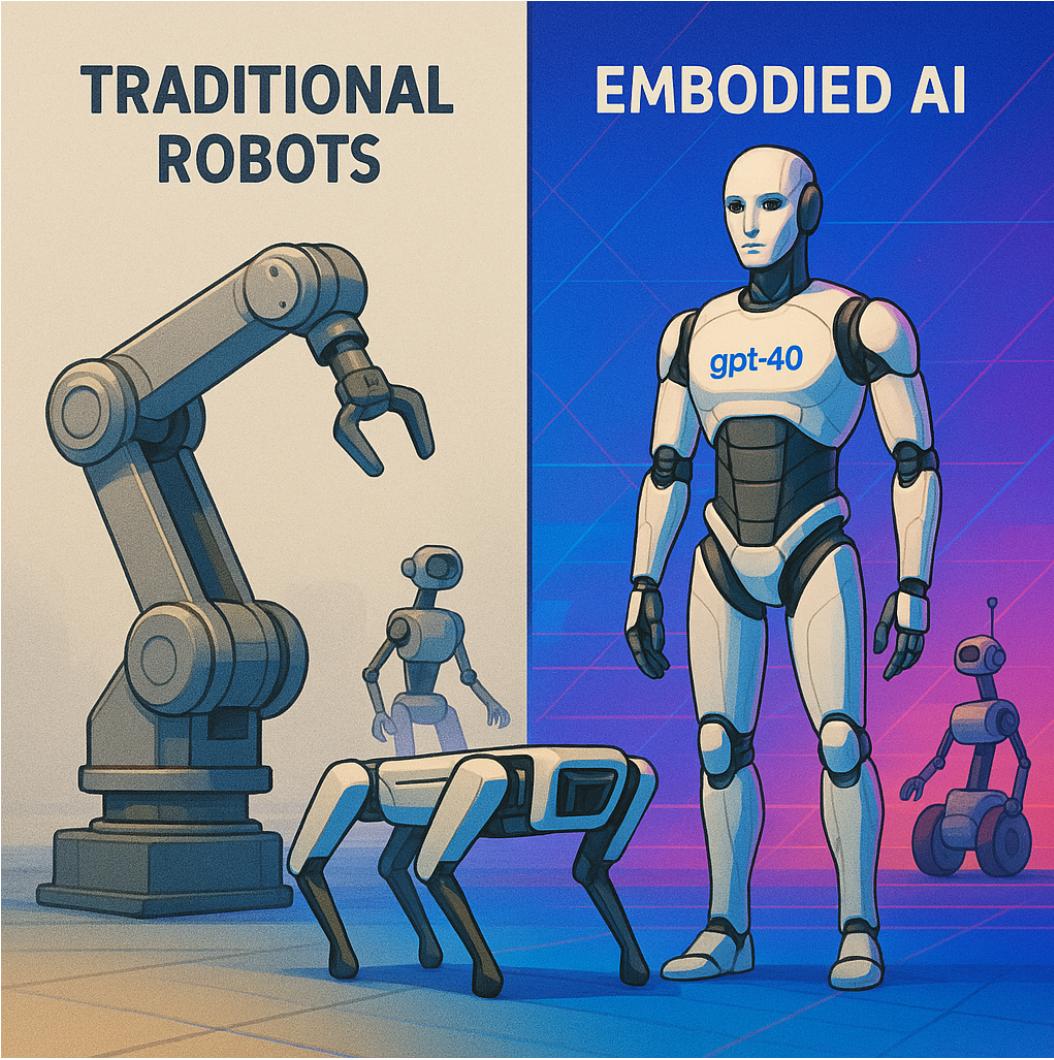


Figure 1: An illustration showing the transition from the traditional robots to the integration of vision-language models such as GPT-4o. This picture is generated by GPT-4o.

## 1 Introduction

### 1.1 Motivation and Overview

The emergence of GPT-4o [11] as a state-of-the-art Vision-Language Model (VLM) has unlocked new research opportunities in Embodied Intelligence. While recent breakthroughs like GPT-4o demonstrate high-fidelity multimodal generation, their architectures remain undisclosed. This raises a critical question: Can GPT-4o serve as a foundation for advancing embodied reasoning, manipulation, and navigation?

In this paper, we address this pivotal question by conducting a comprehensive evaluation of GPT-4o's capabilities. Our research explores the model's performance across the intricate dimensions of embodied reasoning, embodied manipulation, and embodied navigation within the realm of Embodied Intelligence. Through thorough testing and detailed analysis, we clarify both the strengths and limitations of GPT-4o, providing valuable insights for researchers as they explore potential future applications in Embodied Intelligence.

Our exploration of GPT-4o in the field of Embodied Intelligence mainly focuses on the following aspects:

1. **Embodied Reasoning:** Embodied reasoning is a core capability in embodied artificial intelligence, enabling agents to understand and act in physical environments based on perceptual input. It integrates multi-modal signals—such as vision, language, and motion—with cognitive processes like spatial inference, temporal sequencing, and causal modeling. This integration allows agents to interpret scenes, anticipate changes, make informed decisions, and execute meaningful actions. As the cognitive foundation of embodied intelligence, embodied reasoning connects perception with action, facilitating tasks from object recognition to execution. It translates visual information into symbolic understanding, reasons through spatial-temporal relationships, and formulates strategies, which are essential for navigating dynamic and unstructured environments.
2. **Embodied Manipulation:** Embodied manipulation represents a cornerstone yet challenging capability within the domain of embodied artificial intelligence, enabling agents to interact with their environments through physical actions. As a concrete application of embodied AI, it focuses on how an agent perceives object properties, understands potential physical interactions, plans action sequences, and executes precise motions to alter environmental states. Building upon embodied reasoning, embodied manipulation addresses complex physical constraints and fine-grained motion control—a critical competency for intelligent agents to accomplish meaningful tasks in real-world environments. It emphasizes learning and decision-making through active environmental interaction rather than passive perception. Effective embodied manipulation requires a comprehensive set of perceptual and cognitive capabilities, including robust scene understanding for interpreting environmental contexts, sophisticated long-horizon task planning for sequencing multi-step operations, and precise motion planning for executing physically feasible actions. These competencies are fundamental to enabling agents to perform complex, real-world interactions in dynamic environments such as domestic service, industrial automation, or assistive robotics.
3. **Embodied Navigation:** Embodied navigation represents a fundamental task within the broader domain of embodied artificial intelligence. As a specific application of embodied AI, navigation focuses on an agent’s ability to perceive its environment, reason about spatial relationships, and execute movements to reach desired destinations. While embodied AI encompasses a wide range of capabilities including manipulation, interaction, and multi-modal perception, embodied navigation specifically addresses the challenge of purposeful movement through space—a capability critical for agents to function effectively in physical or simulated environments.

In conclusion, we offer initial insights as a foundation for inspiring future research endeavors in the realm of Embodied Intelligence with GPT-4o. While this methodology may be somewhat less stringent, it affords the opportunity for a comprehensive analysis.

## 1.2 Guidance

This article focuses on testing within the field of Embodied Intelligence, using a curated selection of images that represent diverse embodied scenarios. The test samples are sourced from various outlets, including open-source datasets such as MapNav [20], R2R-CE [8, 13], SpatialQA-E [1], GraspNet-1Billion [3], AGD20k [10], Where2Place [18], BridgeData V2 [16], Ego4D [4] and EgoThink [2]. All experiments detailed in this paper were conducted before April 8, 2025, utilizing the web-hosted GPT-4o Pro (version from March 28, 2025).

## 2 Embodied Reasoning

Embodied reasoning [14, 7, 6, 15, 9] serves as a foundational yet complex capability within the field of embodied artificial intelligence, enabling agents to understand, infer, and act within physical environments based on perceptual input. Unlike passive perception, embodied reasoning emphasizes the integration of multi-modal signals (e.g., vision, language, motion cues) with structured cognitive processes such as spatial inference, temporal sequencing, causal modeling, and goal-oriented planning. It underpins an agent’s ability to not only interpret the current scene but also to anticipate changes, make informed decisions, and execute actions that meaningfully alter the environment.

As the cognitive substrate of embodied intelligence, embodied reasoning supports a wide range of tasks [24, 2, 4], from object recognition to action execution, by linking what the agent sees, knows, and intends to do. Effective embodied reasoning requires a fusion of perception and abstraction—translating visual information into symbolic understanding, reasoning over spatial-temporal relationships, and formulating action strategies. These capabilities are critical for agents operating in dynamic, unstructured environments such as homes, warehouses, and outdoor navigation.

With the rise of large multimodal models like GPT-4o, there is growing interest in exploring how these models perform across key dimensions of embodied reasoning. By systematically evaluating GPT-4o across carefully designed scenarios, we aim to reveal its strengths, limitations, and potential for future deployment in physical AI systems.

In this paper, we assess GPT-4o’s embodied reasoning capabilities across five essential dimensions:

Sec. 2.1 Perceptual Reasoning evaluates GPT-4o’s ability to interpret raw visual inputs and extract semantically meaningful insights. This includes identifying object categories, detecting visual anomalies, recognizing functional attributes, and linking perceptual cues to linguistic concepts. Tasks such as object selection in cluttered scenes or answering visual questions require not only vision understanding but also grounded language reasoning.

Sec. 2.2 Spatial Reasoning assesses how GPT-4o understands the geometric layout of scenes, including object positions, orientations, and affordance constraints. It tests whether GPT-4o can infer spatial relationships (e.g., left/right, above/below), evaluate the feasibility of physical actions (e.g., whether an object fits through a space), and reason about movement or accessibility in complex environments.

Sec. 2.3 Temporal Reasoning focuses on GPT-4o’s capacity to process events across time. This includes understanding temporal sequences, identifying causal event chains, and predicting future states or reconstructing past ones from visual observations. By analyzing multi-frame scenes or temporally evolving situations, we assess whether GPT-4o can perform time-sensitive inference and maintain coherent temporal models.

Sec. 2.4 Planning Reasoning examines GPT-4o’s hierarchical reasoning for multi-step problem solving. This involves decomposing high-level goals into ordered subtasks, selecting appropriate tools or objects, and optimizing action sequences. GPT-4o is tested on its ability to generate structured plans that are physically executable and logically consistent in visual manipulation settings.

Sec. 2.5 Causal Reasoning evaluates whether GPT-4o can infer cause-and-effect relations in physical contexts. This includes predicting the consequences of actions (e.g., what happens if an object is pushed or dropped), explaining observed outcomes, and distinguishing between correlation and causation. Strong causal reasoning is essential for safe and adaptive behavior in real-world interactions.

These five dimensions form the cognitive core of embodied intelligence, spanning from low-level perception to high-level reasoning and planning. Together, they provide a comprehensive framework for evaluating how GPT-4o interprets and acts upon the world, offering critical insights into its readiness for real-world physical intelligence applications.

## 2.1 Perceptual Reasoning

Perceptual reasoning is the cognitive process of interpreting sensory input—particularly visual information—to extract meaningful patterns, recognize objects, understand attributes, and make semantically grounded decisions. Within embodied intelligence, it plays a foundational role by bridging raw perception and high-level reasoning, enabling agents to comprehend their environments and act accordingly.

In the context of multimodal models like GPT-4o, perceptual reasoning requires aligning visual understanding with language-based queries or instructions. This includes recognizing object categories and attributes (such as shape, color, or material), detecting task-relevant affordances, and selecting or describing objects based on semantic criteria. Strong perceptual reasoning allows an agent to answer visual questions, perform object-centric reasoning, and make informed decisions grounded in perceptual evidence.

Unlike low-level visual processing, perceptual reasoning emphasizes conceptual understanding over pure recognition. It involves interpreting visual input in context, often under linguistic guidance or task constraints. For example, choosing a tool not just based on appearance but based on its intended use in a given scenario.

In this section, we assess GPT-4o’s perceptual reasoning abilities through visual tasks that test its understanding of object-function mappings, visual decision-making, and multi-frame perception. These examples examine the model’s capacity to extract actionable meaning from images and to reason across modalities in grounded, task-specific settings. In the first example (Figure 2), we present two task scenarios to evaluate the model’s understanding of object affordances and contextual tool usage. The first task asks GPT-4o to select the appropriate tool to open and read an email, while the second involves choosing a suitable tool to draw a cloud as a decorative element. Both tasks require the model to link semantic language instructions with the correct visual affordance of objects. GPT-4o successfully selects the correct tool in both cases, demonstrating strong alignment between visual recognition and task-oriented interpretation.

In the second example (Figure 3), we further evaluate GPT-4o on two fronts. The first subtask tests its ability to plan and execute decisions based on the perceived scene and task instruction—selecting and using an appropriate tool to complete a visually grounded goal. GPT-4o demonstrates correct understanding and generates reasonable action sequences. The second subtask focuses on object perception across a temporal visual sequence, where GPT-4o is asked to count the number of chairs appearing in a video. Here, the model fails to produce the correct answer, indicating a current limitation in maintaining temporal consistency and cross-frame object tracking across visual inputs.

These results show that while GPT-4o exhibits strong perceptual reasoning in single-image, task-grounded settings, its ability to integrate information across multi-frame visual sequences still presents challenges—highlighting a key direction for future improvements in temporal and spatial attention within multimodal reasoning models.

## Perceptual Reasoning

### Prompt:

Scene Description: You are an embodied intelligent agent with visual perception, language understanding, and the ability to interact with your environment.

Language Instruction: "Please find an item that can be used to read emails."

Reasoning Question: First, analyze the objects in your current environment and consider their possible functions. Then, select the most suitable item for the task of reading emails, and explain the reasoning behind your choice.



### GPT-4o:

Scene Description:

In your visual field, you can perceive the following objects: A desktop computer with monitor, keyboard, and mouse placed on a long desk. A coffee mug beside the computer. A photo frame, a decorative white figure, and a potted plant on the desk. Shelving with a lamp, wooden mannequin, and small decorative sculptures. A cabinet, striped wallpaper, and an armchair in the corner

Expected Response:

The most suitable item for reading emails is the desktop computer on the desk.

Final Answer:

Reading emails requires digital access to the internet and a screen to view messages. The desktop computer is the only object in the scene with that capability—it includes a monitor, keyboard, and mouse.

Other visible items, such as the mug, plant, or decorative objects, do not have the functionality to display or access electronic communication.

Therefore, **the desktop computer** is the correct and most functional choice for completing this task.

### Prompt:

Language Instruction: "The child has just completed a drawing. Now, the teacher has asked her to add a decoration of clouds, using only the materials visible on the table."

Reasoning Question: First, analyze the items available on the table and consider their possible uses. Then, **identify the item that is most suitable for completing the task of adding clouds to the drawing**. Explain your reasoning based solely on what is available in the visual scene and the teacher's instruction.



### GPT-4o:

Cloud decoration typically requires white or light material that can be shaped or drawn. **The child already drew part of the clouds with blue marker**, but the teacher wants an added decorative element. The rolled fabric is decorative but not cloud-like and harder to apply on paper. Stickers or cutouts do not appear to include cloud shapes. Markers or crayons offer the best control for drawing.

Most Suitable Item:

White crayon or white marker (if available among the supplies in the cup).

If white is not visible in the cup, the light blue marker or crayon already used on the right side of the paper could continue the theme in a decorative way.

Final Answer:

The **light blue marker or crayon** is the most suitable visible material for decorating the clouds, as it closely resembles cloud color and texture and is already part of the drawing. It allows the child to follow the teacher's instruction using materials clearly visible on the table.

Figure 2: Illustration of GPT-4o's ability to perform perceptual reasoning. **Green** highlights the right answer in understanding. Check Section 2.1 for detailed discussions.

## Perceptual Reasoning

### Prompt:

Language Instruction: "The goal is to prepare a garnish using the white radish to serve with the patties."  
Reasoning Question: First, analyze the available tools and ingredients in your environment and consider their functions. Then, determine the most appropriate method to prepare the radish for garnishing the patties, and explain your reasoning based on the scene and the cooking task.



[图片]

### GPT-4o:

Since the instruction is to prepare a garnish using the radish: Garnishes should be fine, light, and visually appealing. The grater held in the right hand is ideal for making finely shredded or grated daikon, which is a common garnish in Japanese and Korean cuisines, often served with fried or grilled items like patties or tempura. This would complement the patties well, adding a refreshing, slightly spicy contrast in texture and taste. Slicing or dicing with a knife would result in chunks that are too large or clunky for a garnish, whereas grating provides the right texture and appearance.

Final Answer:

The most appropriate method to prepare the radish for garnishing the patties is to **grate it using the box grater held in hand**.

This tool allows the radish to be transformed into a fine, moist, and refreshing topping that visually and texturally complements the patties, aligning perfectly with the goal of creating a garnish.

### Prompt:

Language Instruction: "Please analyze all the provided images and determine how many individual chairs are present in total."

? Reasoning Task:

Carefully inspect each image to locate all visible chairs. Identify whether any chairs appear in multiple images (avoid double-counting).

Provide the total count of unique chairs.



### GPT-4o:

Image 1 (Front-Facing View): 6 on the far side and sides of the table (backs facing the camera). 2 on the near side (backs partially visible, facing away from the camera).

Image 2 (Angled Left View): The same 8 chairs around the table are visible (some from the side or back now). Final Chair Count: No duplicates added, as both images show the same 9 chairs from different angles.

Chair Locations: Around the large table: 8 chairs.

Near the bookshelf (right side): 1 chair

Total Unique Chairs: 9

Figure 3: Illustration of GPT-4o's ability to perform perceptual reasoning. **Green** highlights the right answer in understanding. Check Section 2.1 for detailed discussions.

## 2.2 Spatial Reasoning

Spatial reasoning refers to the cognitive ability to perceive, understand, and manipulate spatial relationships between objects within a physical or imagined environment. In the context of embodied intelligence, spatial reasoning is essential for interpreting geometric structures, estimating object locations, navigating through space, and planning physically feasible actions. It underpins a wide range of embodied tasks, from basic orientation and object alignment to complex motion planning and obstacle avoidance.

For multimodal models like GPT-4o, effective spatial reasoning involves not only recognizing the spatial layout of a scene from visual input, but also aligning this understanding with linguistic instructions and goal constraints. This includes interpreting relative positions (e.g., “to the left of”), estimating scales and distances, mentally simulating movement trajectories, and understanding affordances constrained by geometry (e.g., whether an object can pass through a narrow opening).

In this section, we evaluate GPT-4o’s spatial reasoning capabilities through a set of four diverse tasks that range from static perception and direction estimation to dynamic path planning and decision making. These examples are designed to probe different aspects of spatial cognition, including geometric abstraction, spatial imagination, and multi-modal integration.

In Example 1 (Figure 4), the model is given two sub-tasks: determining the floor level of a scene and estimating the height of a chair from a given image. Both tasks require spatial abstraction from visual cues, such as horizon alignment, object scale, and surrounding reference structures. GPT-4o accurately completes both sub-tasks, demonstrating strong static spatial inference and object-scale estimation from limited visual context.

In Example 2 (Figure 5), we test GPT-4o’s mental spatial rotation and direction estimation. The task involves interpreting a virtual agent’s facing direction and inferring the relative position of a target object. GPT-4o incorrectly answers this task, revealing its current limitations in allocentric spatial reasoning and internal spatial visualization, especially when handling imagined 3D rotations or egocentric-to-allocentric frame transitions.

Example 3 (Figure 6) presents a path planning task where GPT-4o is required to reason from a given start location to a goal by inferring a valid trajectory through a structured environment. The model successfully infers the correct spatial path, showcasing its capacity for topological reasoning, relative positioning, and visual route planning—key competencies for navigation and motion control in embodied agents.

In Example 4 (Figure 7), we evaluate GPT-4o in a compound reasoning task. The first sub-task requires integrating textual instructions with spatial layouts to make a context-aware decision, combining its perceptual grounding and positional inference. The second sub-task involves identifying the relative spatial relationship between objects. GPT-4o performs well on both, indicating strong abilities in multi-modal spatial integration, language-conditioned spatial decision-making, and deductive reasoning based on visual layouts.

These results indicate that GPT-4o demonstrates robust capabilities in direct spatial perception, visual-motor mapping, and task-conditioned spatial inference. However, it still faces challenges in abstract spatial imagination, especially when required to mentally transform perspectives or reason beyond the directly observable scene. Improving this aspect remains essential for advancing general-purpose embodied intelligence.

## Spatial Reasoning

### Prompt:

⌚ Scene Context: You are an embodied robot agent located in a kitchen.

🗣 Language Instruction: "You are currently inside a kitchen."

❓ Reasoning Task: Analyze the kitchen environment using all available perceptual input. **Infer which floor level you are currently on.** Provide a step-by-step explanation of how you arrived at this conclusion using observable evidence or environmental cues.



### GPT-4o:

⌚ Step-by-Step Analysis:

1. Observation of Outside Environment:

Through the window, you can see the tops of cars and rooftops of small structures, parts of the neighboring houses' roofs and a clear view over the brick garage labeled "20 PARKER", suggesting a higher vantage point.

2. Relative Elevation Comparison:

The orange-tiled roof of the garage is below the line of sight from the window. That implies you're situated above the garage roof level, which is typically at single-story height. The view looks down onto cars, which suggests you're not on ground level.

3. Visible Downpipe and Brick Wall:

There is a large drainpipe attached to the side of the building with visible clamps, running downward along a vertical brick wall.

4. No Stairs or Basement Indicators:

No stairs going downward or signs of being underground, so basement is unlikely.

📄 Conclusion:

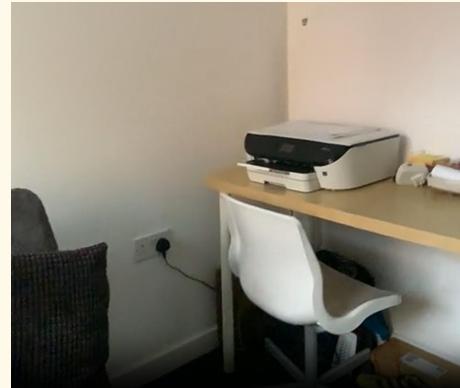
You are most likely on the second floor.

### Prompt:

⌚ Scene Context: You are an embodied robot agent located in an indoor room.

🗣 Language Instruction: "**Estimate the length of the longest dimension (length, width, or height) of the chair in centimeters.**"

❓ Reasoning Task: Analyze the visual scene and identify the chair in question. Use visible reference objects to estimate the chair's dimensions. Determine which of the chair's dimensions—length (depth), width, or height—is the longest. Provide your final estimated measurement in centimeters, and explain your reasoning process.



### GPT-4o:

⌚ Step 1: Identify Reference Objects

Several familiar objects help establish scale: Standard UK sockets are approximately 8.6 cm high and 14.6 cm wide. Typical desks are around 72–76 cm tall. The HP-style inkjet printer is about 45 cm wide.

⌚ Step 2: Analyze Chair Geometry

The height of the chair seems slightly less than the desk's height, so let's say around 70 cm. The width (side-to-side) seems smaller than the printer's width, perhaps ~40–45 cm. The depth (front to back) appears slightly longer than the width.

⌚ Step 3: Determine the Longest Dimension

Comparing: Height: 70 cm; Width: 45 cm; Depth: likely slightly longer than width → estimated around 50–55 cm.

Conclusion: **The height of the chair(70 cm) is the longest dimension.**

Figure 4: Illustration of GPT-4o's ability to perform spatial reasoning. **Green** highlights the right answer in understanding. Check Section 2.2 for detailed discussions.

## Spatial Reasoning



### Prompt:

⌚ Scene Context: You are an embodied robot agent situated in a domestic living room.

🗣 Language Instruction: "You are currently sitting on a sofa, facing a coffee table."

❓ Reasoning Task:

Based on your current body orientation and surrounding environment: Identify the location of the power strip within the room. **Determine whether the power strip is on your left side or right side relative to your seated, forward-facing position.** Provide a step-by-step explanation of how you derived that spatial relationship — referencing nearby landmarks (e.g., walls, furniture, lighting, objects) that support your conclusion.

### GPT-4o:

💡 Power Strip Location

In the first image, a close-up of the rightmost end of the sofa shows: A beige armrest. A power strip on the floor, near the corner of the wall and the window. A white vertical blind, which matches the right side of the scene from the second image. This confirms: The power strip is located on the floor near the window and wall, next to the right armrest of the sofa.

⌚ Agent-Centered Spatial Reasoning

Since the agent is sitting on the middle seat of the sofa. Facing forward toward the coffee table. The power strip is near the rightmost seat (from the viewer's perspective), which translates to the agent's right side.

☑ Final Answer

The power strip is **on your right side** relative to your seated, forward-facing position.

Figure 5: Illustration of GPT-4o’s ability to perform spatial reasoning. **Green** highlights the right answer in understanding. Check Section 2.2 for detailed discussions.

## Spatial Reasoning



### Prompt:

⌚ Scene Description: You are an embodied robot agent currently standing with your back facing a window. You are located in an indoor environment where several objects may be visible in your field of view.

🗣 Language Instruction: "You now need to go to the location where the bicycle is."

❓ Reasoning Task:

Analyze the environmental layout and the positions of surrounding objects based on your current view. Determine your relative position to the bicycle. **Plan a movement route from your current position to the bicycle.** Explain the reasoning behind the chosen path, considering object positions, obstacles, and the shortest or most efficient route.

### GPT-4o:

⌚ Current Location and Orientation

You are facing away from the window. Positioned near a wooden dining table with mixed wood chairs. The bookshelf is slightly to your front-right. Behind that, the space transitions into a living room area with a couch, TV, and lamps.

🚴 Goal Location: The bicycles are mounted on the far-right wall, beyond the archway in the back wall. This space appears to be a side hallway or a bike storage nook.

🧭 Reasoning and Route Planning

- **Turn Slightly Right:** Your goal is beyond the archway seen past the living room, slightly to your right-front. This avoids the dining table directly in front of you.
- **Move Forward, Navigating Past the Table:** Carefully move forward and slightly to the right to pass the edge of the dining table and avoid chairs.
- **Navigate Along the Couch:** Proceed along the left side of the brown sectional couch. Stay clear of the coffee table and floor furniture like the brown ottoman.
- **Pass Through the Archway:** Head through the arched doorway visible just behind the couch area.
- **Arrive at the Bicycle Zone:** Upon crossing through the archway, you will find the bicycles mounted on the right wall, near some boxes and a small plant.

Figure 6: Illustration of GPT-4o's ability to perform spatial reasoning. **Green** highlights the right answer in understanding. Check Section 2.2 for detailed discussions.

## Spatial Reasoning

### Prompt:

⌚ Scene Description: You are an embodied robot agent located in a parking lot.

🗣 Language Instruction: "You need to close the driver-side car door."

❓ Reasoning Task: Analyze your current position relative to the car and the open door. **Determine which direction you should move in order to reach and close the driver-side door.** Explain the reasoning behind your movement decision, based on spatial layout and task



### GPT-4o:

⌚ Environmental Analysis: I am facing the red car with the driver-side door open. The door is located to the front-right of my current position. The car is on my right side, and the door is hinged on the front edge, meaning it opens outward and slightly forward.

⌚ Movement Plan:

Step to the right and slightly forward, toward the car door. I need to move around the vacuum hose in my hands and avoid stepping on the car mats laid on the ground.

❓ Reasoning:

The door opens away from the car, so to close it, I must be positioned outside the car, near the edge of the door.

My current position is too far back and left; **I need to adjust right and forward to get within reach.** The goal is to minimize movement while maintaining balance and avoiding obstacles, like the vacuum, cables, and mats.

### Prompt:

⌚ Scene Context: You are an embodied robot agent.

🗣 Language Instruction: "Estimate the distance, in meters, between the closest points of the piano and the table/bench."

❓ Reasoning Task: Identify the piano and the bench, and estimate their spatial positions, focusing on the gap between their nearest edges. Provide an estimated distance in meters, and explain your reasoning.



### GPT-4o:

⌚ Estimation Process:

- Object Identification:

The piano is placed against the wall. A bench is positioned in front of it, aligned along the same axis.

- Real-World Size Priors:

Standard upright piano depth: ~60 cm. Typical piano bench width: ~30–35 cm.

- Gap Estimation (Visually):

The gap between the front edge of the piano and the back edge of the bench appears to be relatively small. Using the bench width for scale comparison, the gap seems to be slightly less than the bench depth.

⌚ Estimated Distance:

The estimated distance is approximately **0.25 to 0.35 meters (25–35 cm)**.

Figure 7: Illustration of GPT-4o’s ability to perform spatial reasoning. **Green** highlights the right answer in understanding. Check Section 2.2 for detailed discussions.

### 2.3 Temporal Reasoning

Temporal reasoning is the cognitive ability to understand, represent, and infer relationships between events across time. In the realm of embodied intelligence, it enables agents to track how states evolve, predict future outcomes, and reconstruct past sequences based on current or sequential observations. This capability is crucial for interacting with dynamic environments, where success often depends on anticipating what will happen next or understanding what has already occurred.

For multimodal models like GPT-4o, temporal reasoning requires integrating visual frames, identifying event progressions, and aligning them with linguistic or task-specific goals. This involves understanding temporal order, continuity, and causality—such as inferring whether one event precedes another or predicting how a scene will change over time.

Unlike static perception, temporal reasoning focuses on change, sequence, and predictive inference. It forms the backbone of real-world tasks like video understanding, robotic control, and multi-step interaction, where events unfold incrementally and require time-aware decision-making.

In this section, we evaluate GPT-4o’s temporal reasoning capabilities through two representative examples:

In Example 1 (Figure 8), GPT-4o is presented with two predictive tasks, where it must forecast the likely future outcome of a given scene. Each scenario requires the model to observe the current physical state and infer what is most likely to happen next. GPT-4o successfully completes both tasks, demonstrating a strong capacity for future state prediction and understanding implicit causal progressions in visual inputs.

In Example 2 (Figure 9), GPT-4o is tasked with reconstructing the temporal order of object appearances based on a sequence of images. This evaluation probes the model’s ability to track entities across multiple frames and reason about their appearance sequence or temporal transitions. GPT-4o correctly infers the event order, suggesting a promising ability to bind multi-frame observations and derive consistent temporal relationships across scenes.

Together, these examples reveal GPT-4o’s solid performance in both predictive and sequential temporal reasoning. While further analysis is needed on more complex, long-horizon dynamics, these results indicate a strong foundational capacity for time-sensitive reasoning within embodied contexts.

## Temporal Reasoning

### Prompt:

⌚Scene Context: You are an embodied robot agent.  
🗣Language Instruction: "Based on the current scene and object positions, what action am I about to perform?"  
❓Reasoning Task: Analyze the spatial layout and object interactions. Consider hand positions, object types, and typical kitchen workflows. **Predict whether the noodles are about to be poured into the plate or into the pot.** Provide your prediction and reasoning based on action intention and task context.



### GPT-4o:

⌚Spatial and Intent Analysis:  
 • Hand Positions:  
 Left hand is holding a bowl of noodles above the sink area. Right hand is touching or steadying the empty square plate, suggesting preparation or alignment.  
 • Object Proximity and Orientation:  
 The square plate is directly in front of the right hand, positioned conveniently on the counter. The pot on the stove is further away and already has a lid partially open, possibly used for boiling (now finished).  
 • Task Context:  
 Given that the noodles are already cooked and being held in a bowl, and a sauce bowl is also prepared nearby, the next logical step is plating — not cooking further. The presence of chopsticks nearby also hints toward meal serving rather than additional cooking.  
 ☑Conclusion:  
 You are likely in the serving/plating phase, and **about to pour the noodles from the bowl into the square plate** for final presentation or mixing with the sauce.

### Prompt:

⌚Scene Context : You are an embodied robot agent.  
🗣Language Instruction: "Based on the current visual context and your hand-object interaction, what are you going to do next?"  
❓Reasoning Task: Observe the hand placement and object orientation. Consider the nature of the object (container with lid), and the environment (worksit). **Predict the most likely next action you are about to perform.** Explain your reasoning using clues from hand position, object type, and task setting.



### GPT-4o:

⌚Reasoning:  
 • Hand Position:  
 Your left hand is gripping the top edge of the container lid. Your right hand is stabilizing or securing the bottom of the container. This two-handed grip is typical when preparing to twist or pry open a lid.  
 • Object Type:  
 The container resembles one commonly used for paint, joint compound, or spackle—materials that need to be accessed during construction work.  
 • Environmental Clues:  
 The presence of drywall pieces, tools, pliers, and wires indicates an ongoing construction or renovation task. Such tasks frequently involve using materials stored in containers like the one shown.  
 • Posture and Focus:  
 Your gaze and posture are directed toward the container, indicating intention and attention. This suggests an immediate interaction, such as opening it to access the contents. Thus, your most probable next action is to **remove the green lid from the container** to begin using its contents.

Figure 8: Illustration of GPT-4o’s ability to perform temporal reasoning. **Green** highlights the right answer in understanding. Check Section 2.3 for detailed discussions.

## Temporal Reasoning



### Prompt:

Scene Context: You are an embodied robot agent.

Language Instruction: "You will be given a list of object categories. Your task is to identify the first-time appearance order of each category as you move through the environment."

Task Objective: Analyze the visual stream and determine the order in which the following categories first appear: "Blanket", "Trash can", "Microwave", "Plant".

Reasoning Task: For each object category, detect the first visual frame where an instance of that category appears. Sort the categories based on their first appearance in the temporal sequence. Provide the final ordered list, and briefly justify the order by referencing key visual cues or locations.

### GPT-4o:

Ordered List of First-Time Appearances:

**Plant, Blanket, Trash can, Microwave**

Justification with Visual Cues:

- Plant

Frame: 1st image (Dining Area)

Cue: A potted plant is visible on top of the bookshelf on the right.

Conclusion: This is the earliest appearance of any target object.

- Blanket

Frame: 2nd image (Living Room)

Cue: A light gray blanket is draped across the corner of the L-shaped couch.

Conclusion: This appears after the dining area in the camera path.

- Trash can

Frame: 3rd image (Kitchen)

Cue: A white trash can is clearly visible under the sink, partially tucked into a cupboard.

Conclusion: Kitchen appears later in the sequence after the living room.

- Microwave

Frame: 4th image (Further into Kitchen)

Cue: A white microwave is seen on the right counter, near the dish rack.

Conclusion: It's introduced after the trash can within the deeper part of the kitchen.

Figure 9: Illustration of GPT-4o's ability to perform temporal reasoning. **Green** highlights the right answer in understanding. Check Section 2.3 for detailed discussions.

## 2.4 Planning Reasoning

Planning reasoning refers to the ability to formulate coherent sequences of actions that transform an initial state into a desired goal state. In embodied intelligence, this involves decomposing high-level goals into subgoals, making decisions under spatial and temporal constraints, and adapting plans based on contextual understanding. Effective planning reasoning enables agents to operate autonomously in complex environments by combining perception, inference, and decision-making.

For large multimodal models like GPT-4o, planning reasoning encompasses not only generating action sequences from textual prompts but also grounding those sequences in visual and task-specific contexts. This includes selecting appropriate tools or paths, resolving dependencies between subtasks, and adapting strategies when immediate solutions are not visible. A robust planning model should be able to infer missing steps, generalize from prior experience, and handle tasks that require exploration or multi-step reasoning.

In this section, we examine GPT-4o’s planning reasoning capabilities through two distinct examples:

In Example 1 (Figure 10), GPT-4o is tasked with solving two embodied planning problems in a simulated environment, where the goal object (e.g., a pot or an apple) is not immediately visible within the initial observable area. These tasks require the model to go beyond visual saliency, use common sense priors, and plan exploratory actions to reach the goal. However, GPT-4o fails to construct valid action sequences in both cases, leading to unsuccessful outcomes.

This failure highlights two key limitations:

Lack of exposure to simulator-style environments during training results in a visual gap that affects the model’s ability to reason about interaction constraints in such domains.

The prevalence of direct-answer reasoning formats in its training data may bias the model toward tasks that can be solved in one shot, limiting its effectiveness on problems requiring multi-step trial-and-error or uncertain search-based planning.

In Example 2 (Figure 11), GPT-4o is given a real-world task that requires step-by-step planning to achieve a final objective based on an observable scene and textual instruction. The model successfully generates logically coherent, context-aware action sequences, demonstrating its strength in deterministic, goal-driven planning when the relevant objects and pathways are clearly present and interpretable.

These results suggest that GPT-4o is proficient at generating linear, goal-aligned plans in visually grounded contexts but faces challenges when tasks demand exploratory reasoning, environmental uncertainty, or out-of-sight goal modeling. Enhancing the model’s robustness in such settings remains a crucial direction for scaling embodied planning capabilities toward real-world deployment.

## Planning Reasoning

### Prompt:

⌚ Scene Context: You are an embodied AI agent in a kitchen environment. You are given a task instruction, and your perception system provides a visual observation of the current scene. Your goal is to reason and plan your next action based on what you observe.  
 📁 Task Instruction: "Put the Pot in the Fridge."  
 ? Prompt Question: What is your current plan to begin achieving the task based on what you see? Please respond with a <Observation> <Thinking> <Planning> <DecisionMaking> step.



### GPT-4o:

<Observation> I observe a kitchen environment with several objects placed on the counters. A black fridge is visible against the far wall. A silver pot is located on the left-side counter near the sink, next to a blue coffee machine.  
 <Thinking> The task is to put the pot in the fridge. **The pot is already in view and reachable.** The fridge is also clearly visible and accessible, with space below that likely corresponds to the fridge compartment. No obstacles currently block either the pot or the fridge.  
 <Planning>

1. Navigate to the counter near the sink to reach the pot.
2. Pick up the pot.
3. Navigate to the fridge.
4. Open the fridge door.
5. Place the pot inside the fridge.
6. Close the fridge door.

### Prompt:

⌚ Scene Context: You are an embodied AI agent in a kitchen environment. You are given a task instruction, and your perception system provides a visual observation of the current scene. Your goal is to reason and plan your next action based on what you observe. 📁 Task Instruction: "put the Apple in the Microwave"  
 ? Prompt Question: What is your current plan to begin achieving the task based on what you see? Please respond with a <Observation> <Thinking> <Planning> <DecisionMaking> step.



### GPT-4o:

<Observation> In the kitchen environment, I can see a red apple placed on a small round table near a chair, on the right side of the image. A microwave is located on the counter to the left of the sink and underneath the hanging cabinets, near the center-right of the kitchen layout.

<Thinking> The task is to put the apple in the microwave. **Both the target object (apple) and destination (microwave) are visible and appear to be accessible.** The table is not too far from the microwave, and there are no visible obstacles preventing the pickup or placement.

- <Planning>
1. Navigate to the round table.
  2. Pick up the apple from the table.
  3. Navigate to the counter with the microwave.
  4. Open the microwave door.
  5. Place the apple inside the microwave.
  6. Close the microwave door.

Figure 10: Illustration of GPT-4o’s ability to perform planning reasoning. **Green** highlights the right answer in understanding. Check Section 2.4 for detailed discussions.

## Planning Reasoning

### Prompt:

⌚Scene Context: You are an embodied robot agent.  
 🗂️Language Instruction: "Figure out how to unscrew the screw that's located inside the round object in your left hand."  
 ?Reasoning Task: Identify the type of screw and its orientation within the object. Determine whether there is sufficient space to insert a standard tool. Based on your surroundings, select an appropriate tool to reach and turn the screw. Describe how to perform the unscrewing action, including grip, angle, and motion.



### GPT-4o:

⌚ Step 1: Identify the Screw Type & Orientation  
 The screw inside the round black housing appears to be a Phillips-head screw (cross-shaped slot).  
 🗂️ Step 2: Assess Tool Clearance  
 The cavity looks deep and narrow. You likely need a long, narrow Phillips screwdriver or a magnetic bit screwdriver with an extension.  
 🗂️ Step 3: Choose and Use the Right Tool  
 Find a long, thin Phillips screwdriver or a bit driver with a long extension. Insert the tool straight down into the screw slot inside the housing. Make sure the tip seats snugly in the cross-shaped head. With your left hand stabilizing the housing, use your right hand to: Apply downward pressure. Rotate the screwdriver counterclockwise (lefty-loosey). If the screw is tight, increase torque carefully to avoid stripping it.  
 🗂️ Tips: If space is too tight, consider using a flex-shaft screwdriver extension or a ratchet with a Phillips bit and extension. If the screw is stuck, applying a drop of penetrating oil around its base and waiting a minute can help loosen it.

### Prompt:

⌚Scene Context: You are an embodied robot agent.  
 🗂️Language Instruction: "You want to read the magazines that are on the table in front of you."  
 ?Reasoning Task: Identify the location and orientation of the magazines. Plan how to access and read them, considering any obstructing items. Describe a step-by-step manipulation sequence that would allow you to read the magazines comfortably. Justify the reasoning behind the steps.



### GPT-4o:

⌚ Perception and Initial Assessment: The magazines are on the glass coffee table, partially covered by a remote control and a mobile phone.  
 🗂️Reasoning and Planning: To read the magazines, the goal is to clear visual obstructions and bring the magazines into optimal viewing range.  
 ☑Step-by-Step Manipulation Plan  

- Assess Reachability
- Remove Obstructions: Use right hand to pick up the remote control first. Move it to the sofa armrest or side table. Use left hand to pick up the mobile phone. Place it next to the remote or in a stable, non-obstructing spot on the sofa.
- Expose and Align Magazines: Slide the magazines forward slightly (closer to edge of the coffee table) using both hands or one hand if stable. Ensure they are fully visible and not overlapping.
- Adjust for Reading

Option 1: Lift the top magazine and place it on your lap for better reading angle.

Option 2: Leave it on the table and lean slightly forward, keeping good posture, to read it directly.

Figure 11: Illustration of GPT-4o's ability to perform planning reasoning. Green highlights the right answer in understanding. Check Section 2.4 for detailed discussions.

## 2.5 Causal Reasoning

Causal reasoning is the cognitive process of identifying and understanding cause-and-effect relationships between events, actions, and outcomes. In the context of embodied intelligence, it enables agents to not only perceive what is happening in a scene but also infer why it is happening and what might happen next if certain actions are taken. This ability is critical for predicting the consequences of physical interactions, avoiding hazards, and making decisions that lead to desirable outcomes.

For multimodal models like GPT-4o, causal reasoning involves integrating visual cues, world knowledge, and counterfactual thinking to evaluate the downstream effects of physical states or actions. It includes identifying potential risks (e.g., objects falling, collisions), recognizing implicit intentions or threats, and reasoning about how interventions may alter the outcome of a scene.

In this section, we test GPT-4o’s causal reasoning abilities through two visual scenarios centered on risk and danger detection:

In Example 1 (left Figure 12), GPT-4o is presented with a real-world scene where a potential danger is present. The model is asked to identify and explain what might go wrong based on the current visual context. GPT-4o successfully detects the hazard, demonstrating a strong understanding of physical affordances, gravity dynamics, and safety implications in a realistic scenario. This shows its ability to simulate causal outcomes from static images by applying physical common sense.

In Example 2 (right Figure 12), the task involves a cartoon or comic-style scene, requiring the model to reason in a stylized but semantically rich setting. Despite the abstract visual representation, GPT-4o is still able to infer the latent causal structure and correctly identify the potential danger in the scene. This suggests that the model possesses robust causal generalization that extends across different visual styles and abstraction levels.

## Causal Reasoning

### Prompt:

⌚ Scene Description: You are a service robot operating in a kitchen, assisting humans in preparing tea.

👤 Language Instruction: "Please move the hot water kettle to the table."

❓ Reasoning Question: Before executing this task, what potential consequences should you consider? Is it safe to carry out the action immediately? Why or why not?



### GPT-4o:

Before executing the task, several safety and consequence considerations must be made:

Recommended Action Plan:

Decision: **Do not move the kettle immediately.**

Instead, follow these safer steps:

- Wait for the boiling to stop and the kettle to switch off.
- Unplug or confirm the base is disconnected (if cordless).
- Check temperature of the outer surface or wait a short period for cooling.
- Use handles to lift and move carefully, keeping the kettle upright and stable.

⌚ Summary:

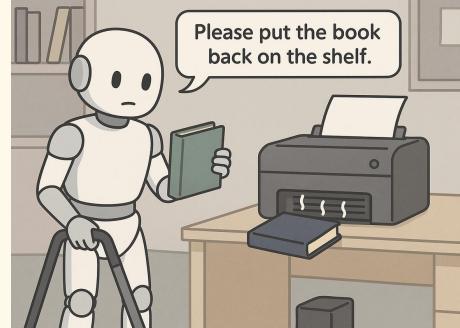
No, it is **not safe** to carry out the action immediately due to risks of burns, electrical hazards, and hot water spillage. Delaying the task until boiling stops and the kettle cools is the responsible choice.

### Prompt:

⌚ Scene Description: You are cleaning in an office environment.

👤 Language Instruction: "Please return the book to the bookshelf."

❓ Reasoning Question: While attempting to complete the task, you observe that the book is partially blocking the ventilation outlet. **How should you handle this situation? What might happen if you ignore it and simply follow the instruction?**



### GPT-4o:

⌚ Reasoning and Decision-Making

► Potential Consequences of Ignoring the Obstruction Blocking the ventilation could cause the printer to overheat, leading to: hardware damage, fire hazard, workflow disruption. Following the command without considering context may lead to unsafe outcomes.

Appropriate Action Plan

- **Assess Safety First:** Ensure the printer is not overheating already. If it is, consider alerting a human or taking protective measures.
- **Remove the Book from the Vent Outlet:** Safely lift the book away from the vent to restore airflow.
- **Inspect the Book for Heat Damage:** If the book is hot or shows damage, avoid returning it to the shelf immediately and notify a human.
- **Return the Book to the Shelf (if safe)**

Only proceed with the original instruction if the book is undamaged and safe to handle.

- **Optional: Notify Supervisor or Log the Event:** Report the incident or log it as an anomaly for further inspection.

Figure 12: Illustration of GPT-4o's ability to perform causal reasoning. **Green** highlights the right answer in understanding. Check Section 2.5 for detailed discussions.

### 3 Embodied Manipulation

Embodied manipulation represents a cornerstone yet challenging capability within the domain of embodied artificial intelligence, enabling agents to interact with their environments through physical actions. As a concrete application of embodied AI, it focuses on how an agent perceives object properties, understands potential physical interactions, plans action sequences, and executes precise motions to alter environmental states. Build upon on embodied reasoning, embodied manipulation deals with more complex physical constraints and fine-grained motion control—a critical competency for intelligent agents to accomplish meaningful tasks in real-world physical environments. It emphasizes learning and decision-making through active environmental interaction rather than passive perception. Effective embodied manipulation requires a comprehensive set of perceptual and cognitive capabilities, including robust scene understanding for interpreting environmental contexts, sophisticated long-horizon task planning for sequencing multi-step operations, and precise motion planning for executing physically feasible actions. These competencies are fundamental to enabling agents to perform complex, real-world interactions in dynamic environments such as domestic service, industrial automation, or assistive robotics.

Embodied manipulation has evolved with many tasks and benchmarks, including object detection, grasping, planning, and mobile manipulation challenges, covering a spectrum of difficulty levels from simple grasping to complex manipulation. Notably, recent studies have begun exploring how large language models (LLMs) and vision-language models (VLMs) can enhance embodied manipulation through their semantic understanding and reasoning capabilities. As a leading VLM, GPT-4o represents a prime candidate for evaluating the potential and limitations of VLMs in embodied manipulation. By systematically assessing its performance across key tasks in embodied manipulation, we can uncover both the strengths of current VLM-based approaches and their critical gaps. This analysis can provide valuable references for improving the integration of VLM in embodied manipulation.

In this paper, we evaluate GPT-4o’s capabilities across four critical dimensions of embodied manipulation:

Sec. 3.1 Compositional Object Understanding assesses GPT-4o’s ability to parse complex scenes into hierarchical object representations, including: object grounding and object attribute understanding such as shape, color, size, and object relation. This is tested through target object grounding and visual question answering in manipulation scenes.

Sec. 3.2 Affordance Grounding evaluates GPT-4o’s capacity to bridge visual perception with executable actions by: object affordance grounding that focus on object-centric interaction regions and spatial affordance grounding for environment-centric free-space prediction. We formulate the outputs of both tasks using the bounding box coordinates of the affordance area.

Sec. 3.3 Task Planning examines GPT-4o’s hierarchical reasoning for long-horizon manipulation, including: tool selection given task and environmental constraints, as well as subtask sequencing with object dependencies and logical reasoning. The planning output comprises a hierarchical task representation, where high-level goals are decomposed into sequentially executable subtasks and atomic actions.

Sec. 3.4 Trajectory Prediction tests GPT-4o’s physics-aware motion reasoning through predicting a trajectory with several waypoints indicating the optimal movement of the robot’s end effector. This process also involves the comprehensive understanding of object, their spatial relations, and obstacles to be avoid during movement.

These dimensions form a complete pipeline from perception to action, providing a comprehensive assessment of an agent’s multi-modal understanding and execution capabilities in physical interaction. Through this evaluation, we aim to reveal the true competency of GPT-4o in physical interaction and offer insights to guide the development of next-generation embodied intelligence.

#### 3.1 Compositional Object Understanding

In this section, we evaluate GPT-4o’s compositional object understanding capability. We input either third-person or first-person perspective images and assess the model’s ability to comprehend manipulable objects through visual question answering (VQA). In our VQA prompts, we test GPT-

4o’s object grounding ability by incorporating queries about object category, color, size and other attributes, while requiring it to output the bounding box coordinates of the referenced objects.

Figure 13 and Figure 14 demonstrate the results of multi-round VQA in compositional object understanding. The results reveal that GPT-4o demonstrates competent performance in basic object recognition tasks, accurately identifying common household items such as mugs, banana, and tools, as well as packaged goods like Cheez-It boxes and Campbell’s soup cans. The model reliably detects object colors (e.g., distinguishing purple and blue mugs) and provides reasonably accurate descriptions of relative positioning (e.g., “the blue mug is slightly in front of the purple one”). These capabilities suggest that GPT-4o has strong foundational skills in object categorization and relative spatial relationship inference, making it suitable for applications requiring high-level scene understanding.

However, the model exhibits notable limitations in precise object localization and size estimation. Its bounding box predictions often deviate from the actual object boundaries (the Cheez-It box and blue objects in Figure 14), sometimes even failing to fully enclose the target item (the mugs in Figure 13 and the banana in Figure 14). Additionally, GPT-4o’s size comparisons rely solely on pixel-level calculations without accounting for perspective distortion, leading to errors when comparing the widths of two visually dissimilar mugs. These shortcomings highlight gaps in the model’s geometric reasoning, particularly in tasks requiring fine-grained spatial awareness or real-world physical intuition. While GPT-4o performs well in coarse object detection, its current limitations in localization and 3D-aware perception restrict its reliability in robotics or manipulation-focused applications.

## Compositional Object Understanding

### Prompt:

Please detect all the mugs in the given image.



### GPT-4o:

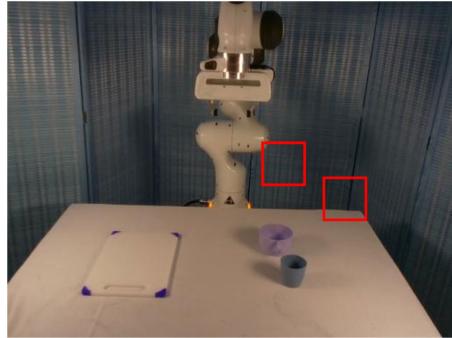
The image shows two mugs on a table:

1. A purple mug located near the center-right of the table.
2. A blue mug located slightly in front and to the right of the purple mug.

### Prompt:

Please draw the bounding boxes of detected mugs in the image and give the accurate coordinates of each bounding box in the following format: (min\_x, max\_x, min\_y, max\_y).

### GPT-4o:



Here are the bounding box coordinates for the detected mugs in the image:

- Purple Mug: (min\_x=290, max\_x=340, min\_y=160, max\_y=210)
- Blue Mug: (min\_x=360, max\_x=410, min\_y=200, max\_y=250)

### Prompt:

If I want to pick the widest mug in the image, which mug should I pick? Ground this mug in the image and provide its bounding box in the following format: (min\_x, max\_x, min\_y, max\_y).

### GPT-4o:

To determine the widest mug, we compare the width of each bounding box:

- Purple Mug: width = 340 - 290 = 50 pixels
- Blue Mug: width = 410 - 360 = 50 pixels

Both mugs have the same width of 50 pixels. However, if we visually inspect the image, the purple mug appears slightly larger in diameter due to its top view and lighting. So, if we prioritize visual appearance and top width, the purple mug should be picked.

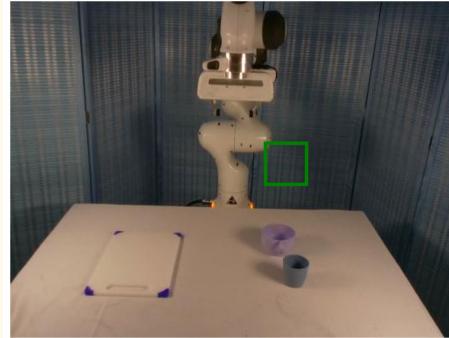
Widest Mug (Grounded):

- Mug: Purple Mug
- Bounding Box: (290, 340, 160, 210)

### Prompt:

Please draw the bounding box of the widest mug in the image.

### GPT-4o:



The green bounding box in the image highlights the widest mug, which is the purple mug. Its bounding box coordinates are:

(min\_x=290, max\_x=340, min\_y=160, max\_y=210)

Figure 13: Illustration of GPT-4o’s ability to ground objects and understand the attributes of objects. Green highlights the right answer in understanding while Red highlights the wrong answer. Check Section 3.1 for detailed discussions.

## Compositional Object Understanding

### Prompt:

Please **detect all the objects** on the tabletop in the given image.



### GPT-4o:

The image shows a tabletop with multiple recognizable objects. Here's a list of detected items:

- **Plastic banana** – near the center top.
- **Gray toy animal** (possibly a rhino) – top-center.
- **Blue plastic cup/mug** – near the banana.
- **Box of Cheez-It crackers** – right side of the image.
- **White bottle with red label** (possibly lotion or cream) – under the blue cup.
- **Can of Campbell's soup** – slightly left of the red-label bottle.
- **Orange and black cordless drill** – lower left side.
- **White cylindrical bottle** (possibly shampoo or body wash) – left side.
- **Small blue cap or tool bit** – far left below the white bottle.

### Prompt:

Please **draw the bounding box** of the banana in the image and give the accurate coordinates of each bounding box in the following format: (min\_x, max\_x, min\_y, max\_y).

### GPT-4o:



The yellow bounding box in the image highlights the banana.

Its bounding box coordinates are:  
(min\_x=330, max\_x=430, min\_y=130, max\_y=250)

### Prompt:

If I want to **pick the biggest object** in the image, which object should I pick? Ground this object in the image and provide its bounding box in the following format: (min\_x, max\_x, min\_y, max\_y).

### GPT-4o:

The **biggest object in the image is the Cheez-It box**, highlighted with a lime green bounding box.

Bounding Box Coordinates:

(min\_x=530, max\_x=640, min\_y=180, max\_y=370)



### Prompt:

If I want to **pick all the blue objects** in the image, can you tell me what I should pick? Ground these objects in the image and provide their bounding boxes in the following format: (min\_x, max\_x, min\_y, max\_y).

### GPT-4o:

**Here are the blue objects grounded in the image with their bounding box coordinates:**

- Blue Cup:  
(min\_x=460, max\_x=520, min\_y=200, max\_y=270)
- Blue Cap:  
(min\_x=200, max\_x=250, min\_y=280, max\_y=330)



Figure 14: Illustration of GPT-4o's ability to ground objects and understand the attributes of objects. **Green** highlights the right answer in understanding while **Red** highlights the wrong answer. Check Section 3.1 for detailed discussions.

### 3.2 Affordance Grounding

Affordance bridges perception to action by encoding "what actions are possible" on objects or "where to place" in spaces. In this section, we assess two kind of affordance grounding: (1) object affordance grounding, which locates the object part that supports the given action; (2) spatial affordance grounding that predicts the free space for object placing.

**Object Affordance Grounding:** We evaluate GPT-4o's object affordance understanding capability by outputting bounding boxes of actionable regions. We use images from the AGD20k [10] test set as input, and provide prompts in the following format:

*"You are a robotic assistant. Now, you are tasked with grasping the knife to cut the bread in the image. Which part of the knife should you grasp? Please mark the object part you can grasp with red bounding box in the image. Also, please provide the bounding box coordinates  $[(x_1, y_1), (x_2, y_2)]$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the top-left and bottom-right corners of the bounding box. Please draw the bounding box on the original image, do not modify the original image."*

As shown in Figure 15-17, GPT-4o demonstrates strong performance in predicting object affordances, correctly identifying all actionable regions in these test cases. However, while the bounding boxes it generates are sometimes accurate, they occasionally exhibit significant deviations (e.g., misaligned bounding boxes in Figure 17). This limitation aligns with its inconsistent performance in precise object localization, suggesting that while GPT-4o excels in high-level understanding and reasoning, it struggles with fine-grained numerical prediction tasks.

**Spatial Affordance Grounding:** We evaluate GPT-4o's spatial affordance understanding by outputting bounding boxes of free space regions. Employing images from the Where2Place Benchmark, we provide prompts containing spatial relational information and one example is as follows:

*"Identify the vacant space that's between the two mugs. Your answer should be a tuple  $(min_x, max_x, min_y, max_y)$  indicating the bounding box of the target region.  $min_x, max_x, min_y, max_y$  are normalized image coordinates between 0 and 1. Please draw the bounding box on the image."*

As shown in Figure 18-20, GPT-4o demonstrates strengths in spatial affordance reasoning. It correctly interprets basic spatial relations like "between," "left," "below," and "above" to identify approximate free-space regions referred by specified objects.

However, GPT-4o exhibits some limitations in spatial affordance prediction:

**(1) Inaccurate free-space localization:** Generated bounding boxes often overlap with objects rather than strictly covering vacant areas.

**(2) Poor performance in complex 3D spatial reasoning:** It struggles with shelf-based scenarios (e.g., "Locate the blank space situated above the apple" and "Locate the vacant area situated below the rightmost item" in Figure 20), failing to confine predicted spaces within shelf boundaries. The model frequently predicts regions outside the shelf despite the contextual expectation of shelf-contained spaces.

In summary, while GPT-4o shows competent 2D planar spatial affordance localization, its performance degrades significantly in 3D-structured environments where depth and containment relationships are critical.

## Object Affordance Grounding

### Prompt:

You are a robotic assistant. Now, you are tasked with **grasping the knife to cut the bread** in the image. Which part of the knife should you grasp? Please mark the object part you can grasp with red bounding box in the image. Also, please provide the bounding box coordinates  $[(x_1, y_1), (x_2, y_2)]$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the top-left and bottom-right corners of the bounding box. Please draw the bounding box on the original image, do not modify the original image.



### GPT-4o:

The part of the knife you should grasp to cut the bread is the **handle**, marked with a red bounding box.  
Bounding box coordinates: [(610, 310), (750, 440)]  
You can download the image with the red bounding box here:



### Prompt:

You are a robotic assistant. Now, you are tasked with **grasping the mug** in the image. Which part of the mug should you grasp? Please mark the object part you can grasp with red bounding box in the image. Also, please provide the bounding box coordinates  $[(x_1, y_1), (x_2, y_2)]$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the top-left and bottom-right corners of the bounding box. Please draw the bounding box on the original image, do not modify the original image.



### GPT-4o:

The part of the mug you should grasp is the **handle**, marked with a red bounding box.  
Bounding box coordinates: [(560, 365), (635, 435)]  
You can download the image with the red bounding box here:



Figure 15: Illustration of GPT-4o's ability to predict object affordance on selected images from AGD20k. **Green** highlights the right answer in understanding. Check Section 3.2 for detailed discussions.

## Object Affordance Grounding

### Prompt:

You are a robotic assistant. Now, you are tasked with **grasping the scissor to cut the paper** in the image. Which part of the scissor should you grasp? Please mark the object part you can grasp with red bounding box in the image. Also, please provide the bounding box coordinates  $[(x_1, y_1), (x_2, y_2)]$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the top-left and bottom-right corners of the bounding box. Please draw the bounding box on the original image, do not modify the original image.

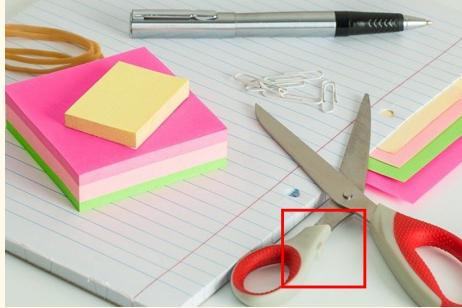


### GPT-4o:

The part of the scissors you should grasp to cut the paper is the **handle area**, marked in red.

Bounding box coordinates:  $[(490, 370), (640, 510)]$

You can download the image with the red bounding box here:



### Prompt:

You are a robotic assistant. Now, you are tasked with **grasping the forks in the image**. Which part of the fork should you grasp? Please mark the object part you can grasp with red bounding box in the image. Also, please provide the bounding box coordinates  $[(x_1, y_1), (x_2, y_2)]$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the top-left and bottom-right corners of the bounding box. Please draw the bounding box on the original image, do not modify the original image.



### GPT-4o:

The part of the forks you should grasp is the **handle area**, marked with a red bounding box.

Bounding box coordinates:  $[(60, 390), (240, 500)]$

You can download the image with the red bounding box here:



Figure 16: Illustration of GPT-4o's ability to predict object affordance on selected images from AGD20k. **Green** highlights the right answer in understanding. Check Section 3.2 for detailed discussions.

## Object Affordance Grounding

### Prompt:

You are a robotic assistant. Now, you are tasked with **open the microwave** in the image. Which part of the microwave should you grasp? Please mark the object part you can grasp with red bounding box in the image. Also, please provide the bounding box coordinates  $[(x_1, y_1), (x_2, y_2)]$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the top-left and bottom-right corners of the bounding box. Please draw the bounding box on the original image, do not modify the original image.



### GPT-4o:

The part of the microwave you should grasp to open it is the **handle, located at the top of the door**.

**Bounding box coordinates:  $[(50, 80), (460, 130)]$**

The red bounding box has been drawn around this area in the image.



### Prompt:

You are a robotic assistant. Now, you are tasked with **open the refrigerator** in the image. Which part of the refrigerator should you grasp? Please mark the object part you can grasp with red bounding box in the image. Also, please provide the bounding box coordinates  $[(x_1, y_1), (x_2, y_2)]$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the top-left and bottom-right corners of the bounding box. Please draw the bounding box on the original image, do not modify the original image.



### GPT-4o:

The part of the refrigerator you should grasp to open it is the **vertical handle on the left side of the door**.

**Bounding box coordinates:  $[(45, 110), (75, 460)]$**

You can download the image with the red bounding box marking the handle here:



Figure 17: Illustration of GPT-4o’s ability to predict object affordance on selected images from AGD20k. **Green** highlights the right answer in understanding while **Red** highlights the wrong answer. Check Section 3.2 for detailed discussions.

## Spatial Affordance Grounding

### Prompt:

Identify the **vacant space that's between the two mugs**. Your answer should be a tuple (min\_x, max\_x, min\_y, max\_y) indicating the bounding box of the target region. min\_x, max\_x, min\_y, max\_y are normalized image coordinates between 0 and 1. Please draw the bounding box on the image.

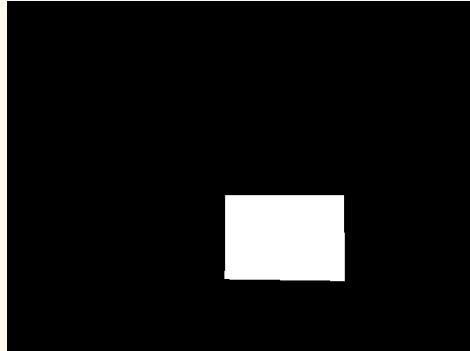


### GPT-4o:

The vacant space between the two mugs is highlighted in red in the image. The bounding box in normalized coordinates is: (min\_x, max\_x, min\_y, max\_y) = (0.38, 0.62, 0.6, 0.8).



### Ground Truth Mask:



### Prompt:

Locate the **vacant space positioned to the left of the yellow mustard bottle**. Your answer should be a tuple (min\_x, max\_x, min\_y, max\_y) indicating the bounding box of the target region. min\_x, max\_x, min\_y, max\_y are normalized image coordinates between 0 and 1. Please draw the bounding box on the image.



### GPT-4o:

The vacant space to the left of the yellow mustard bottle is highlighted in red. The bounding box in normalized coordinates is: (min\_x, max\_x, min\_y, max\_y) = (0.35, 0.55, 0.65, 0.85).



### Ground Truth Mask:

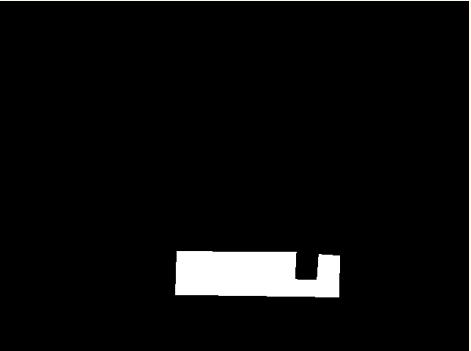


Figure 18: Illustration of GPT-4o’s ability to predict spatial affordance on selected images from Where2Place. We provide ground truth segmentation masks for reference. Check Section 3.2 for detailed discussions.

## Spatial Affordance Grounding

### Prompt:

Find the space that is between the blue mug and orange bowl. Your answer should be a tuple (min\_x, max\_x, min\_y, max\_y) indicating the bounding box of the target region. min\_x, max\_x, min\_y, max\_y are normalized image coordinates between 0 and 1. Please visualize the bounding box on the image.

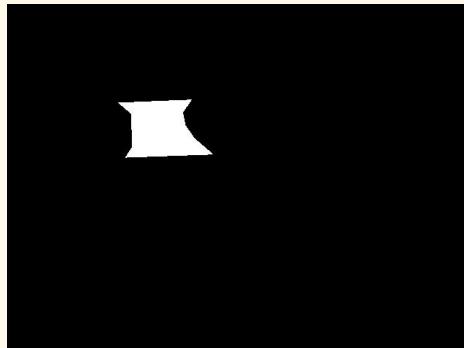


### GPT-4o:

The space between the blue mug and the orange bowl is highlighted in red. The bounding box for this region in normalized coordinates is: (min\_x, max\_x, min\_y, max\_y) = (0.28, 0.42, 0.28, 0.45).



### Ground Truth Mask:



### Prompt:

Pinpoint the vacant space that is in between the phone and measuring tape. Your answer should be a tuple (min\_x, max\_x, min\_y, max\_y) indicating the bounding box of the target region. min\_x, max\_x, min\_y, max\_y are normalized image coordinates between 0 and 1. Please visualize the bounding box on the image.



### GPT-4o:

The vacant space between the phone and the measuring tape is highlighted with a red bounding box in the image. The normalized bounding box coordinates are: (min\_x, max\_x, min\_y, max\_y) = (0.48, 0.60, 0.55, 0.68).



### Ground Truth Mask:

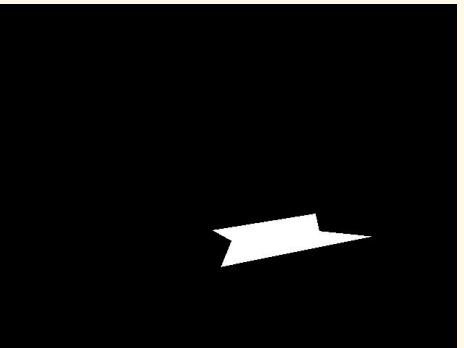


Figure 19: Illustration of GPT-4o's ability to predict spatial affordance on selected images from Where2Place. We provide ground truth segmentation masks for reference. Check Section 3.2 for detailed discussions.

## Spatial Affordance Grounding

### Prompt:

Locate the **blank space situated above the apple**. Your answer should be a tuple (min\_x, max\_x, min\_y, max\_y) indicating the bounding box of the target region. min\_x, max\_x, min\_y, max\_y are normalized image coordinates between 0 and 1. Please draw the bounding box on the image.

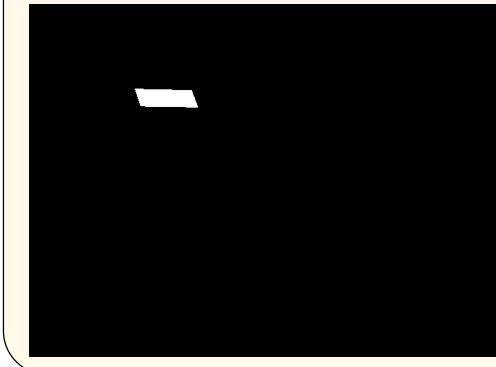


### GPT-4o:

The blank space situated above the apple is highlighted in red in the image. The bounding box in normalized coordinates is: (min\_x, max\_x, min\_y, max\_y) = (0.20, 0.38, 0.35, 0.50).



### Ground Truth Mask:



### Prompt:

Locate the **vacant area situated below the rightmost item**. Your answer should be a tuple (min\_x, max\_x, min\_y, max\_y) indicating the bounding box of the target region. min\_x, max\_x, min\_y, max\_y are normalized image coordinates between 0 and 1. Please draw the bounding box on the image.



### GPT-4o:

The vacant area situated below the rightmost item is highlighted in red in the image. The bounding box in normalized coordinates is: (min\_x, max\_x, min\_y, max\_y) = (0.75, 0.88, 0.55, 0.72).



### Ground Truth Mask:

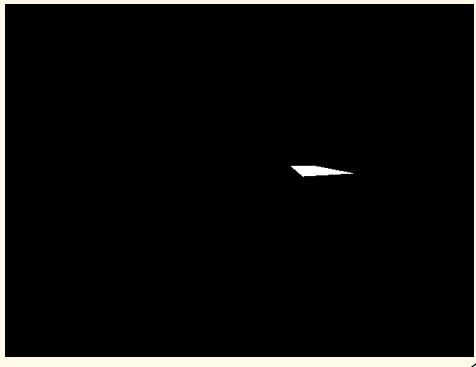


Figure 20: Illustration of GPT-4o’s ability to predict spatial affordance on selected images from Where2Place. We provide ground truth segmentation masks for reference. Check Section 3.2 for detailed discussions.

### 3.3 Task Planning

In this section, we test the GPT-4o's ability on task planning by performing hierarchical decomposition of high-level goals into executable sub-tasks. We provide a third-person perspective image of the scene and prompt GPT-4o with following task planning template:

*"You are a robot specialized in task planning for household activities. Your goal is to decompose complex tasks into executable sub-steps while considering object dependencies, environmental constraints, and logical sequences. Based on the objects in the image, if you are assigned to complete the task 'XXX', provide a detailed task planning with several key steps and descriptions."*

In household task planning tests, GPT-4o demonstrates strong structured task decomposition capabilities in Figure 21-24. The model can break down complex tasks (e.g., "making coffee" in Figure 21 or "setting the dining table" in Figure 22) into logically rigorous substeps while accurately identifying key objects in the environment (e.g., Keurig coffee pods, tableware placement). Its planning solutions exhibit three major strengths: First, GPT-4o demonstrates intuitive step sequencing that aligns with human operational logic, such as the "preheat pan → add oil → flip shrimp" workflow in the cooking task in Figure 24. Second, it shows proactive consideration of the object dependencies (e.g., verifying the coffee machine's power state in Figure 21) and safety constraints (e.g., avoid knocking mug during/after brew for making coffee in Figure 21 and consider timing to avoid burning in cook shrimp in Figure 24). Third, its contingency planning for anomalies is good, such as monitoring towel saturation levels and replacement during spill cleanup.

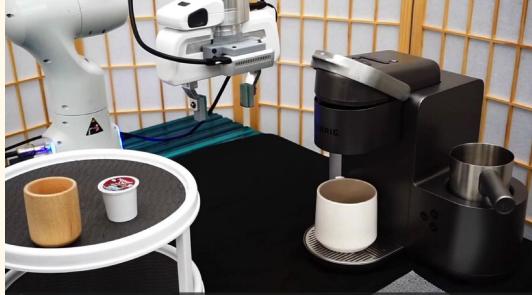
However, GPT-4o still exhibits notable limitations. The most prominent issue is deficient physical commonsense—for instance, GPT-4o assumes the presence of refrigerators and seasonings not visible in Figure 24 and misidentifying the spatula as knife during the "cook shrimp" task, revealing incomplete understanding of visual inputs. Additionally, operational details lack precision: critical parameters like the torque for operating coffee machine handles or the distance in tableware positioning remain unquantified.

Overall, GPT-4o shows promise as a task planning hub, with semantic understanding and workflow decomposition capabilities surpassing traditional scripted robotic systems. However, for real-world physical deployment, it must integrate embodied sensor data to compensate for shortcomings in visual perception, force control, and spatial reasoning in task planning.

## Task Planning

### Prompt:

You are a robot specialized in task planning for household activities. Your goal is to decompose complex tasks into executable sub-steps while considering object dependencies, environmental constraints, and logical sequences. Based on the objects in the image, if you are assigned to complete the task '**make coffee**', provide a detailed task planning with several key steps and descriptions.



### GPT-4o:

The image shows a robot arm next to a Keurig-style coffee maker, with a coffee pod (K-Cup), a cup, and a mug stand. Here's a detailed task planning breakdown for the goal:

#### Task: Make Coffee

#### High-Level Goal

Use the Keurig machine to brew a cup of coffee using the K-Cup pod and serve it in the ceramic mug.

#### Step-by-Step Execution Plan

##### Step 1: Perceive Environment

- 1.1: Use sensors and vision to localize: **Coffee machine, K-Cup pod, Ceramic mug, Machine handle, and Pod chamber**

1.2: Verify the coffee machine is powered on (status light on).

##### Step 2: Prepare Coffee Maker

- 2.1: **Grasp the handle of the Keurig and lift it to open the pod chamber**
- 2.2: Detect whether an old pod is present (if yes, remove and discard).
- 2.3: Place the handle in the fully open position for pod insertion.

##### Step 3: Load K-Cup Pod

- 3.1: Navigate to the table with the K-Cup.
- 3.2: Pick up the K-Cup securely.
- 3.3: **Insert the K-Cup pod into the designated chamber of the machine.**
- 3.4: Lower the handle carefully to close the pod chamber (ensure click-lock).

##### Step 4: Position Mug

- 4.1: Detect the mug on the tray.
- 4.2: **If misaligned, pick and reposition it correctly under the spout.**
- 4.3: Ensure the mug is stable and centered.

##### Step 5: Start Brewing

- 5.1: **Locate and press the brew/start button on the machine.**
- 5.2: Wait for the brewing process to complete (monitor via sensors or time).
- 5.3: Detect end of coffee dispensing (no further liquid flow).

##### Step 6: Post-Brew Cleanup (Optional)

- 6.1: Open the pod chamber.
- 6.2: **Remove used K-Cup pod and place it in trash or recycling.**
- 6.3: Wipe surface if drips detected.

##### Step 7: Serve or Transfer Mug

- 7.1: **Grip the handle of the mug securely.**
- 7.2: Lift and transfer the mug to a designated serving area (e.g., tray or user zone).
- 7.3: Return arm to idle state.

#### Dependencies & Constraints

Mug stability: Ensure cup is ceramic (heavier) and stable on drip tray.

Pod orientation: Insert pod with correct alignment; don't puncture incorrectly.

Handle force: Apply proper torque to close/lock pod chamber securely.

Spill risk: Avoid knocking mug during/after brew.

Thermal safety: Avoid contact with hot liquid or surfaces post-brew.

Figure 21: Illustration of GPT-4o's ability to perform task planning for long-horizon tasks. Green highlights the right answer in understanding. Check Section 3.3 for detailed discussions.

## Task Planning

### Prompt:

You are a robot specialized in task planning for household activities. Your goal is to decompose complex tasks into executable sub-steps while considering object dependencies, environmental constraints, and logical sequences. Based on the objects in the image, if you are assigned to complete the task '**set dinning table**', provide a detailed task planning with several key steps and descriptions.



### GPT-4o:

The image shows a robot arm in a tabletop environment with utensils, a knife, a fork, a plate, a wooden coaster, and a rotating shelf with packaged food items. Given the task "set dining table", the goal is to organize and place tableware appropriately for a meal.

Here's a structured and context-aware plan:

#### Task: Set Dining Table

#### High-Level Goal

Prepare a place setting suitable for an individual meal, including plate, utensils, and potentially a napkin or centerpiece.

#### Step-by-Step Execution Plan

##### Step 1: Perceive & Assess Workspace

1.1: Scan workspace using camera and sensors.

1.2: Identify objects: Plate (metal tray), Fork and knife, Wooden coaster, Fabric placemat, Packaged food (likely not needed for table setting unless it's a dining prop)

1.3: Identify appropriate table center (placemat) for arrangement.

##### Step 2: Place Main Items

###### 2.1: Place Plate

- Pick: Grasp the metal plate securely.
- Place: Center the plate horizontally on the placemat.

###### 2.2: Place Utensils

- Knife:
  - Pick up knife from mat or coaster.
  - Place to the right of the plate, blade facing inward.
- Fork:
  - Pick up fork from coaster.
  - Place to the left of the plate, aligned with the knife.

###### 2.3: Place Coaster

- Pick: Grab wooden coaster.
- Place: Position it above the plate, slightly to the right — where a cup or glass would go.

##### Step 3: Optional Elements

3.1: Add Napkin (if present)

• If a napkin is available (e.g., mat on table), fold and place on top of the plate or left of the fork.

3.2: Decorative or Meal Items (if applicable)

- If the scenario calls for a drink or food item from the rotating shelf:
  - Pick a small, appropriate item (e.g., juice box or raisins box).
  - Place it to the side of the plate or behind the coaster, depending on layout design.

##### Step 4: Final Alignment Check

4.1: Visually verify all items are aligned and within the placemat boundary.

4.2: Adjust for spacing symmetry (centered plate, evenly spaced utensils).

4.3: Confirm stable placement (nothing tipping or off-edge).

#### Dependencies & Constraints

Keep sharp tools (knife) aligned safely.

Ensure plate is centered and not overlapping mat edge.

Avoid obstruction of eating space with packaging.

Prioritize symmetrical and culturally appropriate placement.

Figure 22: Illustration of GPT-4o's ability to perform task planning for long-horizon tasks. Green highlights the right answer in understanding. Check Section 3.3 for detailed discussions.

## Task Planning

### Prompt:

You are a robot specialized in task planning for household activities. Your goal is to decompose complex tasks into executable sub-steps while considering object dependencies, environmental constraints, and logical sequences. Based on the objects in the image, if you are assigned to complete the task '**wipe wine**', provide a detailed task planning with several key steps and descriptions.



### GPT-4o:

Based on the image you provided, the environment appears to be a modern kitchen/lab setting equipped with a robot platform with dual arms, a laptop interface, and a workspace where a wine glass has spilled pink liquid on a white countertop. **The objective is to "wipe wine" – this requires cleanup of the spill.**

Here's a detailed task planning breakdown for the robot to achieve this goal:

#### Task: Wipe Wine (Clean Up Spilled Wine)

#### High-Level Goal

Remove the spilled wine from the countertop surface to leave it dry and clean.

#### Step-by-Step Plan

##### Step 1: Perceive Environment

- 1.1: Activate camera and sensors to locate spill zone.
- 1.2: **Detect and localize the wine spill using visual cues (color, reflection).**
- 1.3: Identify nearby cleaning tools (e.g., paper towels or cloth).
- 1.4: Detect wine glass position to avoid further tipping or ensure stability.

##### Step 2: Prepare Cleanup Tool

- 2.1: **Plan trajectory to approach the paper towel dispenser (visible in the image above the sink).**
- 2.2: Extend arm and grasp one or multiple paper towels.
- 2.3: Ensure firm grip and fold if needed for better absorption.

##### Step 3: Clear Obstacles

- 3.1: **Check if wine glass is unstable or obstructing the wiping area. If yes: Carefully reposition or move the glass to a safer location on the countertop.**
- 3.2: Reconfirm spill area is unobstructed.

##### Step 4: Wipe Surface

- 4.1: **Move cleaning tool to start of spill area.**
- 4.2: Apply gentle downward force and wipe in linear/circular motions.
- 4.3: Monitor saturation level of towel. If saturated: go to Step 2.2 to replace towel.
- 4.4: Repeat until no visible liquid remains.

##### Step 5: Dispose Used Material

- 5.1: Move to trash disposal (if accessible).
- 5.2: **Release the used paper towel into bin.(Optional) If trash is not visible or accessible: Place towel in a predefined holding area.**

##### Step 6: Confirm Completion

- 6.1: Re-scan cleaned surface for residual liquid.
- 6.2: Log task completion status.
- 6.3: Reset arm position to idle mode.

#### Dependencies & Considerations

- Ensure spill is cleaned without spreading it further.  
Maintain safe clearance from glassware.  
Handle wet paper towels with secure grip.  
Avoid dripping during towel transport.  
Surface type (e.g., porous vs. smooth) may affect number of wipes needed.

Figure 23: Illustration of GPT-4o's ability to perform task planning for long-horizon tasks. **Green** highlights the right answer in understanding. Check Section 3.3 for detailed discussions.

## Task Planning

### Prompt:

You are a robot specialized in task planning for household activities. Your goal is to decompose complex tasks into executable sub-steps while considering object dependencies, environmental constraints, and logical sequences. Based on the objects in the image, if you are assigned to complete the task '**cook shrimp**', provide a detailed task planning with several key steps and descriptions.



### GPT-4o:

Based on the environment and objects visible in the image — which includes a pan on a stovetop, cooking oil, a bowl (possibly for ingredients), a knife, and robotic arms — here's a structured task planning for the objective:

🟡 Task: Cook Shrimp

✅ High-Level Goal

Prepare and cook shrimp in a pan using available kitchen tools and ingredients.

📘 Step-by-Step Execution Plan

#### Step 1: Perceive & Prepare Workspace

1.1: Use vision to identify and localize: **Shrimp** (assumed to be available nearby or in fridge), Pan, Bowl, Oil dispenser, **Knife**, Towel

1.2: Check if pan is clean and on stovetop.

1.3: Ensure power access and control to stovetop.

#### Step 2: Fetch Shrimp

2.1: Navigate to refrigerator (if needed).

2.2: Locate container with raw shrimp.

2.3: Grasp container and place it near workspace.

2.4: Open container and transfer shrimp into bowl.

#### Step 3: Season Shrimp (Optional but common)

3.1: Add salt and other available seasonings to bowl.

3.2: Mix gently using utensil or by gently tilting bowl.

#### Step 4: Preheat Pan

4.1: Ensure pan is centered on active burner.

4.2: Turn on stovetop (via knob or digital interface).

4.3: Set to medium-high heat.

4.4: Wait until pan reaches proper cooking temperature (can use thermal sensor if available).

#### Step 5: Add Oil

5.1: Pick up oil dispenser.

5.2: Tilt and pour a controlled amount into heated pan.

5.3: Swirl pan slightly for even coating (or use tool to spread oil).

#### Step 6: Cook Shrimp

6.1: Pick up shrimp using hand or tool (e.g., tongs or spoon).

6.2: Place shrimp individually into pan to avoid overlap.

6.3: Let cook for ~2–3 minutes on one side (until pink/opaque).

6.4: Flip shrimp to cook other side evenly.

6.5: Monitor temperature and appearance to avoid overcooking.

#### Step 7: Finish & Serve

7.1: Once cooked, turn off stovetop.

7.2: Pick up cooked shrimp and place them into a clean serving bowl/plate.

7.3: Move serving bowl to final destination (e.g., dining table).

### 🔴 Dependencies & Constraints

Ensure shrimp is thawed (if frozen).

Avoid cross-contamination between raw shrimp and cooked food.

Do not spill oil during transfer.

Maintain safety near hot surfaces.

Consider timing to avoid burning.

Knife should be used with caution and appropriate safety grip.

Figure 24: Illustration of GPT-4o's ability to perform task planning for long-horizon tasks. **Green** highlights the right answer in understanding while **Red** highlights the wrong answer. Check Section 3.3 for detailed discussions.

### 3.4 Trajectory Prediction

Trajectory prediction aims at generating collision-free motion paths for robots to execute manipulation actions. In this section, we evaluate GPT-4o's trajectory prediction capability for embodied manipulation using images and language instructions from BridgeData V2 [16]. We input the first frame of each video sequence in BridgeData V2 as a first-person perspective image to GPT-4o, requiring it to mark the optimal trajectory in the image, with the final state from the video's last frame serving as reference. An example prompt used in our evaluation is as follows:

*"You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task 'take the banana out of the pot and put it on the left side of the stove', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image."*

Figure 25-28 show the results of GPT-4o in trajectory prediction. We found that GPT-4o demonstrates significant advantages in object recognition and basic action planning. The model can accurately identify various target objects (such as yellow cloth, pot lids, green spoons, blue rectangle etc.) and generate basic action sequences that align with task objectives. Taking the task "take the banana out of the pot" as an example, the model correctly planned the standard steps of "approach → lift → translate → place," reflecting its grasp of simple operational logic. This capability was also validated in tasks like "open the drawer," which require specific operational directions, indicating that the model has acquired basic task decomposition and action sequencing abilities.

However, there are notable shortcomings in trajectory prediction accuracy. Testing revealed significant deviations between the paths generated by the model and the planned action positions. For instance, for the cases in Figure 26-27, the predicted path showed a substantial discrepancy from the actual operational trajectory. In the "sweep into pile" task of Figure 28, the planned sweeping angle failed to fully cover the scattered objects. More critically, the model consistently ignored environmental obstacles. For example, for the task "take the banana out of the pot and put it on the left side of the stove," in Figure 25, it completely overlooked potential collision risks from other stacked blocks along the path.

The model exhibits a core weakness in understanding three-dimensional spatial relationships. When tasks involve multi-layered structures or precise alignment, the predicted waypoints often violate physical constraints. For instance, in the "place the pot lid" task shown in Figure 26, the model's planned horizontal movement height did not account for the protruding handle, which could lead to collisions in real operations. This flaw is particularly pronounced in tasks requiring depth perception, for example, the end waypoint for the case in Figure 28 right column fails to locate on the cylinder, highlighting the inherent limitations of inferring 3D spatial relationships from 2D images.

## Trajectory Prediction

### Prompt:

You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task '**take the banana out of the pot and put it on the left side of the stove**', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.



### GPT-4o:

Here's the predicted optimal trajectory for the robot's end effector to take the banana out of the pot and place it on the left side of the stove:

- Approach the banana (starting point in the pot).
- Lift upward to avoid collisions with the pot's rim.
- Move left above the stove.
- Lower down to place the banana on the left side of the stove.

The red line shows the movement path, and the blue dots mark key trajectory waypoints.



### Ground Truth for Finial State:



### Prompt:

You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task '**put the yellow cloth in the white basket**', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.

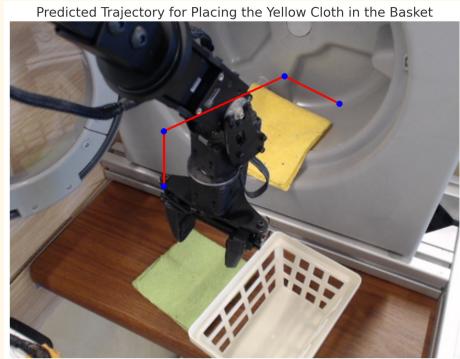


### GPT-4o:

Here's the predicted optimal trajectory for the robot's end effector to pick up the yellow cloth from inside the drum and place it in the white basket:

- Approach the yellow cloth inside the washing machine.
- Lift upward and backward to safely exit the drum.
- Move above the basket while maintaining height.
- Lower down into the basket for placement.

The red line indicates the trajectory, and the blue dots mark the key waypoints.



### Ground Truth for Finial State:



Figure 25: Illustration of GPT-4o's ability to predict trajectory on selected images from BridgeData V2. We provide ground truth of final state for reference. Check Section 3.4 for detailed discussions.

## Trajectory Prediction

### Prompt:

You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task '**put the lid of the pot on the blue cloth**', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.



### GPT-4o:

Here's the predicted optimal trajectory for the robot's end effector to pick up the lid from the pot and place it onto the blue cloth:

- Approach and grasp the lid directly above the pot.
  - Lift upward to clear the pot and handle.
  - Move left while maintaining height to reach the blue cloth area.
  - Lower down and place the lid on the blue cloth.
- Red lines represent the trajectory, and blue dots mark the key trajectory waypoints.



### Ground Truth for Finial State:



### Prompt:

You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task '**place the orange towel to the right of the can**', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.



### GPT-4o:

Here's the predicted optimal trajectory for the robot's end effector to pick up the orange (yellow) towel and place it to the right of the can:

- Approach the towel to grasp it.
- Lift upward to clear the surface.
- Move leftward toward a point just right of the can.
- Lower down to place the towel.

The red path represents the trajectory, with blue dots indicating key trajectory waypoints.



### Ground Truth for Finial State:

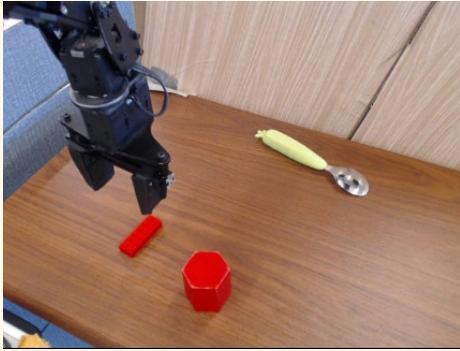


Figure 26: Illustration of GPT-4o's ability to predict trajectory on selected images from BridgeData V2. We provide ground truth of final state for reference. Check Section 3.4 for detailed discussions.

## Trajectory Prediction

### Prompt:

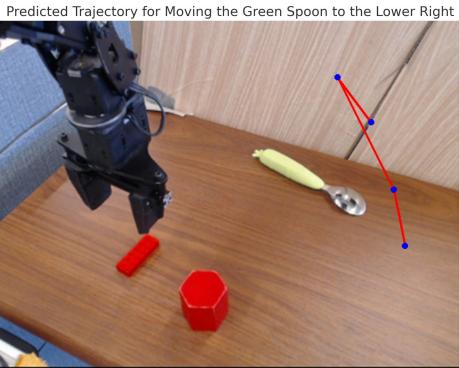
You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task '**move the green spoon to lower right of the table**', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.



### GPT-4o:

Here's the predicted optimal trajectory for the robot's end effector to pick up the green spoon and place it at the lower right of the table:

- Approach the spoon to grasp it.
  - Lift upward to avoid nearby surface or objects.
  - Move diagonally downward to the right, heading toward the target region.
  - Lower down to place the spoon at the desired position.
- The red path represents the trajectory, with blue dots indicating key trajectory waypoints.



### Prompt:

You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task '**open the drawer**', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.



### GPT-4o:

Here's the predicted optimal trajectory for the robot's end effector to grasp the drawer handle and pull it open:

- Approach the drawer handle from above to align for grasping.
  - Move downward to secure a grip on the handle.
  - Pull backward (toward the robot or the bottom of the image) to open the drawer.
  - Continue pulling until the drawer is sufficiently open.
- The red path shows the pulling trajectory, while the blue dots mark the key trajectory waypoints.



### Ground Truth for Finial State:



### Ground Truth for Finial State:



Figure 27: Illustration of GPT-4o's ability to predict trajectory on selected images from BridgeData V2. We provide ground truth of final state for reference. Check Section 3.4 for detailed discussions.

## Trajectory Prediction

### Prompt:

You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task 'sweep into pile', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.

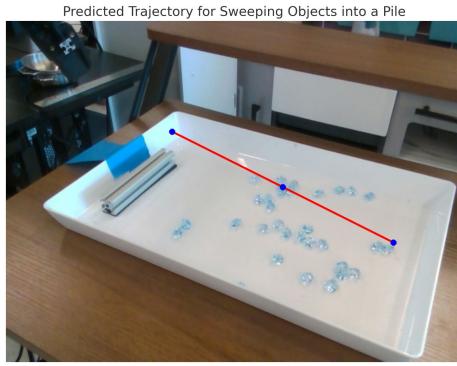


### GPT-4o:

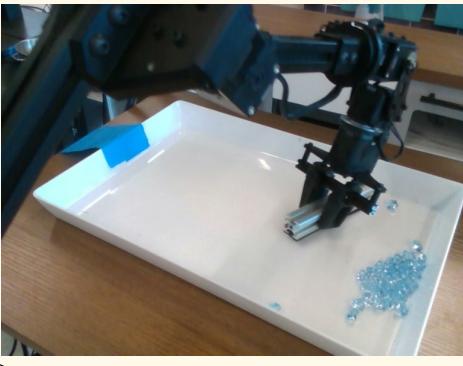
Here's the predicted optimal trajectory for the robot's end effector to sweep scattered objects into a pile:

- Approach the farthest cluster of scattered pieces at the bottom-right of the tray.
- Drag the end effector diagonally across the tray surface toward the upper-left corner.
- Conclude the sweep near the squeegee, forming a concentrated pile of objects.

The red path represents the sweeping trajectory, with blue dots marking the key waypoints.

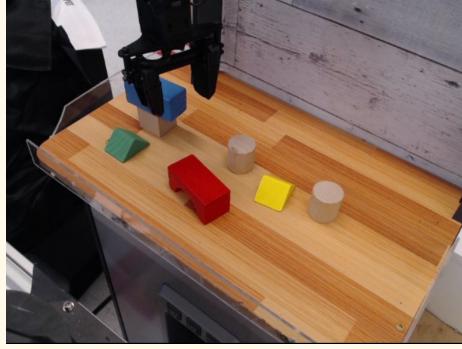


### Ground Truth for Finial State:



### Prompt:

You are a robotic assistant specialized in predicting trajectory for the end effector in robotic manipulation. Based on the given image, if the robot is instructed to complete the task 'put the blue rectangle on top of the cylinder in the middle of the table', what is the optimal trajectory of the robot's end effector? Please draw the trajectory in the original image and do not modify or generate a new image.

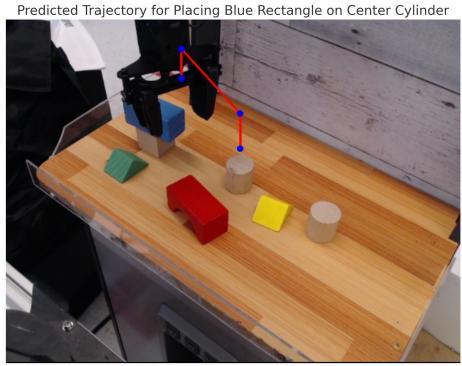


### GPT-4o:

Here's the predicted optimal trajectory for the robot's end effector to pick up the blue rectangle and place it on top of the center cylinder:

- Approach the blue rectangle to prepare for grasping.
- Lift upward to avoid collision with the stacked base.
- Move laterally to position above the center cylinder on the table.
- Lower down to place the blue block securely on the cylinder.

The red path represents the end effector's trajectory, and the blue dots mark key waypoints.



### Ground Truth for Finial State:



Figure 28: Illustration of GPT-4o's ability to predict trajectory on selected images from BridgeData V2. We provide ground truth of final state for reference. Check Section 3.4 for detailed discussions.

## 4 Embodied Navigation

Embodied navigation represents a fundamental task within the broader domain of embodied artificial intelligence. As a specific application of embodied AI, navigation focuses on an agent’s ability to perceive its environment, reason about spatial relationships, and execute movements to reach desired destinations. While embodied AI encompasses a wide range of capabilities including manipulation, interaction, and multimodal perception, embodied navigation specifically addresses the challenge of purposeful movement through space—a capability critical for agents to function effectively in physical or simulated environments.

The embodied AI paradigm emphasizes learning and decision-making through active engagement with environments rather than passive processing of datasets. Within this framework, embodied navigation serves as both a core capability and an enabling technology for more complex tasks. An agent that cannot effectively navigate its surroundings will be limited in its ability to perform manipulation tasks, engage in situated dialogues, or complete goal-directed activities that require movement between locations. The field has evolved through specialized benchmarks such as Vision-and-Language Navigation (VLN) [5, 12, 19, 20, 23], ObjectNav [17, 22, 21], and Mobile Manipulation, each targeting different navigation challenges within the embodied AI ecosystem. Methodological approaches have ranged from reinforcement learning and imitation learning to modular architectures combining classical planning with learning-based components. Recent research has increasingly explored how large language models (LLMs) and vision-language models (VLMs) might enhance embodied navigation through their semantic understanding and reasoning capabilities.

In this paper, we use Vision-and-Language Navigation (VLN) as our evaluation task. VLN requires agents to follow natural language instructions to navigate through realistic 3D environments, making it an ideal testbed for assessing multimodal understanding and spatial reasoning capabilities. This challenging task combines visual perception, language comprehension, and action execution—core components of embodied AI systems. We evaluate GPT-4o’s capabilities across four critical dimensions of embodied navigation:

Sec. 4.1 Navigation Instruction Reasoning evaluates GPT-4o’s ability to analyze navigation videos and generate comprehensive summaries of the navigation process, including identification of key decision points and landmarks encountered along the route.

Sec. 4.2 Navigation Map Reasoning tests GPT-4o’s ability to interpret annotated semantic maps, including understanding object contents, their spatial distribution, room properties, the navigation agent’s current position (represented by a red arrow), and the agent’s past trajectory (represented by a red line).

Sec. 4.3 Navigation Trajectory Inference assesses GPT-4o’s capability to predict and visually mark future navigation trajectories directly on first-person view images when given the next action instruction (e.g., “enter the room”), demonstrating spatial reasoning and path planning abilities.

Sec. 4.4 Navigation Action Prediction examines GPT-4o’s ability to function as a navigation agent by generating appropriate action commands (move forward, turn left, turn right, stop) when provided with a first-person view image and a VLN instruction, bridging perception and decision-making in navigation scenarios.

Through comprehensive evaluation across these dimensions, we aim to understand GPT-4o’s potential contributions to embodied navigation systems and identify both strengths and limitations to guide future research directions.

### 4.1 Navigation Instruction Reasoning

In this section, we evaluate GPT-4o’s ability to analyze and summarize navigation trajectories from first-person perspective videos. We constructed navigation videos by compiling first-person RGB images from complete VLN episodes at a rate of one frame per second and provided them as input to GPT-4o. To assess GPT-4o’s navigation reasoning capabilities, we prompted the model with:

*“Based on this video, tell me what kind of navigation the robot in the video performs and summarize its trajectory. Here is an example: Walk straight down the hallway. Stop just past the double doors. Try to add semantic objects or scenes encountered in some key frames in the picture.”*

As shown in Figure 29-32, our analysis of GPT-4o's responses reveals several limitations in its current capabilities for navigation understanding. While GPT-4o correctly identifies the temporal structure of the video (one frame per second) and attempts to identify key frames, its understanding of the navigation process remains superficial. The model successfully recognizes general environmental features such as hallways, doors, corners, and obstacles. However, it demonstrates significant limitations in two critical areas:

First, GPT-4o fails to identify and interpret many semantic objects present in the images. The model misses important landmarks and environmental features that would be crucial for accurate navigation understanding. This suggests limitations in its visual semantic comprehension abilities when applied to navigation contexts.

Second, GPT-4o struggles to construct coherent trajectory summaries. The model cannot reliably describe navigation paths in terms of sequential actions (e.g., "move forward, then turn right"). This indicates difficulties in understanding the spatial relationships between sequential frames and inferring the agent's movement patterns.

These limitations highlight the challenges GPT-4o faces in integrating temporal visual information with spatial reasoning for navigation understanding. While the model can identify basic environmental structures, it lacks the comprehensive spatial-semantic understanding required for detailed navigation analysis from first-person video sequences.

## Navigation Instruction Reasoning

### Prompt:

Based on this video, tell me what kind of navigation the robot in the video performs and summarize its trajectory. Here is an example: Walk straight down the hallway. Stop just past the double doors. Try to add semantic objects or scenes encountered in some key frames in the picture.



### GPT-4o:

From the key frames, I observed the following:

- Frame 0 (Start): The robot begins by moving along a hallway, facing forward.
- Frame 1 (1/4th of the way through): The robot approaches a set of obstacles or turns, possibly indicating a change in direction or response to an object.
- Frame 2 (Midway): The robot is likely navigating around a more complex scene or avoiding an object that may have been detected.
- Frame 3 (3/4th of the way through): The robot makes further adjustments to its path, continuing its movement along the corridor.
- Frame 4 (End): The robot stops at a specific location or obstacle, potentially indicating the end of the route or a stopping point.
- The robot's navigation appears to involve moving along a hallway with periodic adjustments based on the environment (such as objects or changes in the path).



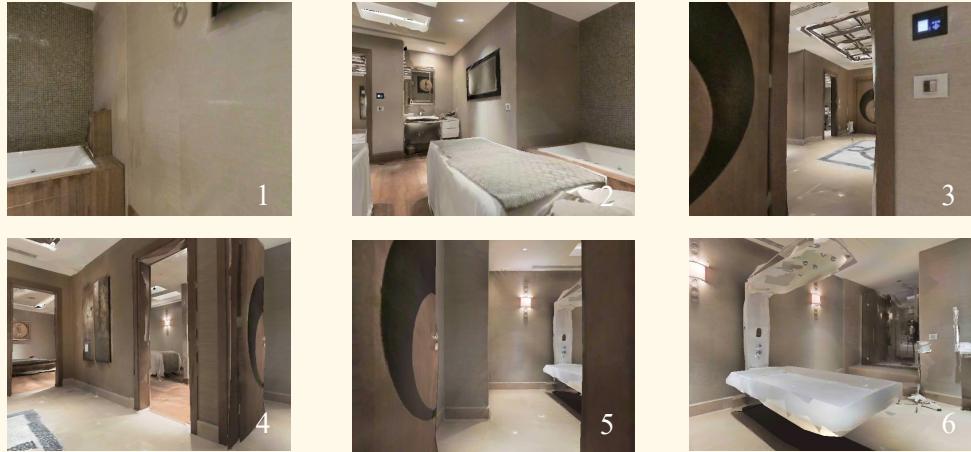
**Ground Truth:** Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it.

Figure 29: Illustration of GPT-4o's navigation instruction reasoning capabilities on a VLN task sequence. The model attempts to interpret key frames from first-person navigation but provides only general descriptions without capturing specific semantic objects or directional details. The highlighted **Orange** shows the ground truth instruction from the VLN task. Check Sec. 4.1 for detailed discussions.

## Navigation Instruction Reasoning

### Prompt:

Based on this video, tell me what kind of navigation the robot in the video performs and summarize its trajectory. Here is an example: Walk straight down the hallway. Stop just past the double doors. Try to add semantic objects or scenes encountered in some key frames in the picture.



### GPT-4o:

Based on the key frames from the video, I can describe the robot's navigation and trajectory as follows:  
 Frame at 0.0 seconds: The robot begins its movement, likely in a hallway or open space with some visible walls and objects.  
 Frame at 14.2 seconds: The robot appears to be moving forward, passing by some obstacles or furniture, such as a table or chair.  
 Frame at 28.4 seconds: It continues down the path, perhaps nearing a corner or door, with the environment becoming clearer (e.g., a door frame or open space visible).  
 Frame at 42.6 seconds: The robot may be nearing a turn or encountering a more significant obstacle, potentially near the end of the hall.  
 Frame at 57.0 seconds: The final frame shows the robot near its stopping point, possibly just past a set of double doors or a large obstacle.  
 From these frames, it seems the robot moves straight down a hallway, encounters some obstacles, and possibly stops just past a significant feature such as a door or a corner.



**Ground Truth:** Exit the beginning room and enter the open area. Turn right and enter the farthest right room.  
 Stop right in front of the table.

Figure 30: Illustration of GPT-4o’s navigation instruction reasoning capabilities on a VLN task sequence. The model attempts to interpret key frames from first-person navigation but provides only general descriptions without capturing specific semantic objects or directional details. The highlighted **Orange** shows the ground truth instruction from the VLN task. Check Sec. 4.1 for detailed discussions.

## Navigation Instruction Reasoning

### Prompt:

Based on this video, tell me what kind of navigation the robot in the video performs and summarize its trajectory. Here is an example: Walk straight down the hallway. Stop just past the double doors. Try to add semantic objects or scenes encountered in some key frames in the picture.



### GPT-4o:

Based on the key frames from the video, the robot's navigation and trajectory appear as follows:

Frame at 0.0 seconds: The robot begins moving in an environment that seems to have a corridor or open space, possibly near some walls or furniture.

Frame at 12.8 seconds: The robot moves forward, possibly avoiding an obstacle or navigating around an object like a table or chair.

Frame at 25.6 seconds: The robot continues straight, with a clear view of what seems to be a narrow path, nearing a wall or an end of the hallway.

Frame at 38.4 seconds: The robot might be approaching a turn or encountering a more significant obstacle, such as a door or a larger piece of furniture.

Frame at 51.0 seconds: The robot reaches its stopping point, potentially near a door or just past an obstacle, bringing its journey to a close.

Summary of the trajectory: The robot moves straight through a hallway or corridor, encounters some obstacles such as tables or chairs, and eventually stops near a door or the end of the hallway.



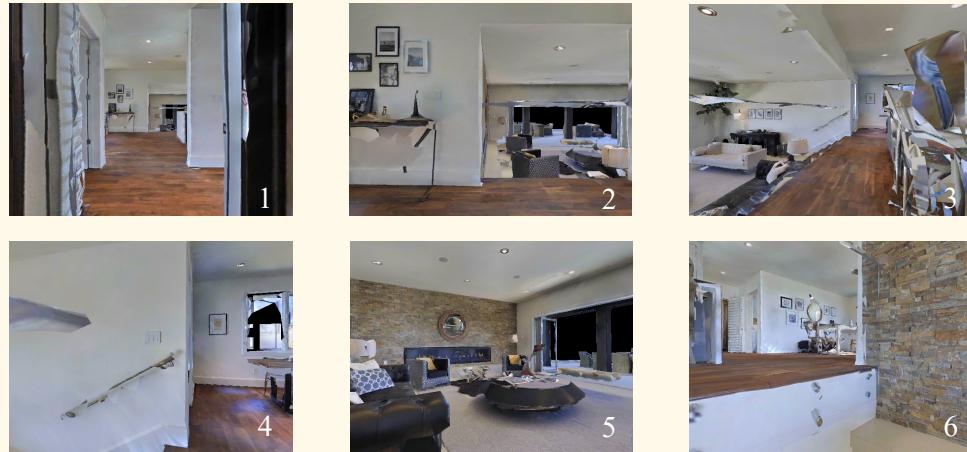
**Ground Truth:** Turn around so you are facing the kitchen. Move forward towards the kitchen island. Move through the kitchen, staying to the left of the kitchen island so that the island is on your right and the counters are on your left. When you get close to the large sink (not the one in the island) there will be an arched passageway to the left. Go through this passageway and towards the dining room. Stop in the second arched doorway leading to the dining room.

Figure 31: Illustration of GPT-4o's navigation instruction reasoning capabilities on a VLN task sequence. The model attempts to interpret key frames from first-person navigation but provides only general descriptions without capturing specific semantic objects or directional details. The highlighted **Orange** shows the ground truth instruction from the VLN task. Check Sec. 4.1 for detailed discussions.

## Navigation Instruction Reasoning

### Prompt:

Based on this video, tell me what kind of navigation the robot in the video performs and summarize its trajectory. Here is an example: Walk straight down the hallway. Stop just past the double doors. Try to add semantic objects or scenes encountered in some key frames in the picture.



### GPT-4o:

Based on the key frames from the video, I can summarize the robot's navigation and trajectory as follows:

Frame at 0.0 seconds: The robot starts its movement in what appears to be a wide, open space or corridor. The environment has some walls and possible objects like chairs or tables.

Frame at 21.8 seconds: The robot continues down the path, encountering a set of obstacles such as a table or some furniture. It seems to follow a linear trajectory with minor adjustments to avoid objects.

Frame at 43.6 seconds: The robot might be nearing a turn or facing a more significant obstacle, possibly a corner or a larger object in its path.

Frame at 65.4 seconds: The robot appears to have completed a portion of its journey and is likely near the end of the hallway or another key feature, such as a door or turn.

Frame at 109.0 seconds: The robot reaches its final position, possibly having stopped just past an obstacle or arrived at a designated stopping point.

Summary of the trajectory: The robot moves straight through a corridor, adjusting its path to avoid obstacles like tables or chairs. It likely navigates a turn or corner before reaching a stopping point near the end of the space.



**Ground Truth:** Walk into the elevator area. Walk out of the elevator and down the hall into the kitchen area.

Walk around the kitchen counter tops towards the living room area. Walk up to the backside of the couches near the table lamp.

Figure 32: Illustration of GPT-4o's navigation instruction reasoning capabilities on a VLN task sequence. The model attempts to interpret key frames from first-person navigation but provides only general descriptions without capturing specific semantic objects or directional details. The highlighted **Orange** shows the ground truth instruction from the VLN task. Check Sec. 4.1 for detailed discussions.

## 4.2 Navigation Map Reasoning

In this section, we evaluate GPT-4o’s ability to interpret and reason about annotated semantic maps (ASMs) from the MapNav methodology [20]. These maps provide a structured representation of the navigation environment, including semantic object annotations, room layouts, and agent trajectory information.

We provided GPT-4o with these annotated semantic maps and prompted the model with:

*"Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line."*

As shown in Figure 33-36, our analysis of GPT-4o’s responses reveals strong capabilities in map reasoning and interpretation. The model demonstrates several key competencies:

First, GPT-4o can accurately identify and locate semantic objects annotated on the map. It successfully recognizes various household objects and their spatial arrangements within the environment. This suggests robust visual perception abilities when interpreting structured spatial representations.

Second, the model effectively infers room properties based on the distribution of objects within defined areas. GPT-4o can determine room types (e.g., kitchen, bedroom, living room) by analyzing the collection of objects typically found in these spaces, demonstrating an understanding of functional spatial semantics.

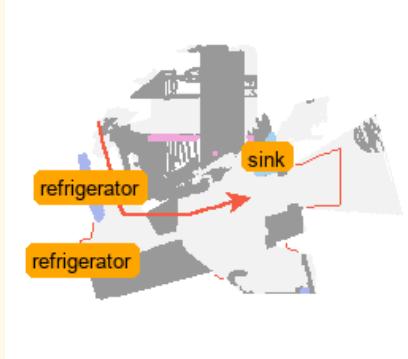
Third, GPT-4o accurately interprets the agent’s position and movement history as represented by the red arrow and trajectory line. The model can describe the agent’s current location and reconstruct its past path through the environment with reasonable precision.

These results indicate that GPT-4o possesses strong capabilities in understanding structured spatial representations. The model’s ability to interpret annotated semantic maps suggests potential applications in high-level navigation planning and spatial reasoning tasks. The performance contrast with the previous Sec. 4.1 highlights that GPT-4o may be better suited for reasoning about structured spatial representations than for interpreting temporal sequences of first-person views.

## Navigation Map Reasoning

### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### GPT-4o:

#### Objects in the Map

Refrigerators (x2): Identified with yellow labels and the word “refrigerator”.

Sink: Also labeled in yellow as “sink”.

#### Object Location Distribution

One refrigerator is located towards the lower left part of the scene.

Another refrigerator is located near the center, slightly to the left.

The sink is positioned toward the right-center of the scene.

#### Room Properties

The objects and general layout suggest that this is a kitchen area.

The environment appears to be 3D-structured with walls and counters, indicated by the shading and contours.

#### Navigation Agent

The red arrow represents the current location and facing direction of the navigation agent.

The agent is facing toward the sink, and is located roughly at the center of the image.

#### Agent's Trajectory

The red line traces the past trajectory of the navigation agent.

It shows a path:

Starting from the bottom left near one refrigerator,  
Moving upwards toward the second refrigerator,  
Then heading to the right, ending near the sink.

### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### GPT-4o:

#### Objects in the Map

TV (x2): Labeled with the word “tv” and marked with yellow labels.

Bed: Labeled with “bed,” positioned at the center of the map.

Hair: Labeled with the word “hair,” possibly indicating the location of a mirror or a personal item near the bed.

#### Object Location Distribution

The bed is located towards the center of the map.

TV objects are positioned:

One TV on the left side of the bed.

Another TV positioned towards the right of the bed.

The hair object is placed above the bed, possibly indicating a personal item or something associated with grooming.

#### Room Properties

The layout suggests this might be a bedroom or living room environment. The general area appears spacious, with visible walls and furniture items like the bed and TVs.

#### Navigation Agent

The red arrow represents the current location and facing direction of the navigation agent. The agent is located to the left of the bed and is facing towards the right side of the room.

#### Agent's Trajectory

The red line traces the past trajectory of the navigation agent: It starts near the bottom-left corner of the map,

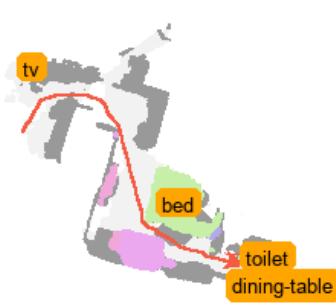
Moves towards the TV on the left, then curves around the bed and hair objects, Ends near the TV on the right, indicating the agent's path through the room.

Figure 33: Illustration of GPT-4o’s map reasoning capabilities on annotated semantic maps from MapNav. The model accurately identifies semantic objects, their spatial relationships, room functions, and the agent’s position and trajectory. Check Section 4.2 for detailed discussions.

## Navigation Map Reasoning

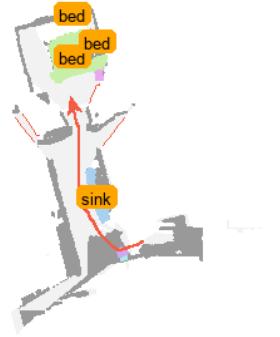
### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### GPT-4o:

#### ⌚ Objects in the Map

TV – Marked with a yellow label, located near the top left.  
Bed – Clearly labeled and located roughly in center of the map.  
Toilet – Identified near the bottom right.

Dining Table – Also located in the same region as the toilet, bottom right.

#### 📍 Object Location Distribution

The TV is positioned toward the top-left of the map.

The bed is located near the center, slightly to the right.

Both the toilet and dining table are situated at the bottom-right section of the map, likely indicating a shared or adjacent space.

#### 🏡 Room Properties

The layout appears to be a long, connected indoor environment, possibly a small apartment or suite:

The left section seems to be a living area (with the TV).

The middle section functions as a bedroom.

The right section combines bathroom and dining features, which is uncommon but possible in small studio-type layouts.

#### 🕒 Navigation Agent

The red arrow indicates the current location and facing direction of the agent. The agent is located at the bottom-right, facing slightly upward/left, near the toilet and dining-table.

#### 🏃 Agent's Trajectory

The red line shows the path the agent has taken: Starting near the TV (top-left), Moving downward and curving toward the bed in the center, Then continuing to the bottom-right, ending near the toilet and dining table.

### GPT-4o:

#### ⌚ Objects in the Map

Bed (x3): Three instances labeled “bed” in the top area.  
Sink: Labeled near the lower part of the map, likely part of a kitchen or bathroom.

#### 📍 Object Location Distribution

All three beds are clustered toward the top-center of the map, possibly indicating:

A shared sleeping area, or

A large bedroom with multiple beds.

The sink is placed in the lower-middle portion of the map, hinting at a kitchen or bathroom zone.

#### 🏡 Room Properties

The structure is narrow and vertical, resembling a hallway-connected apartment or dorm-like layout:

Top section: Clearly a sleeping area with multiple beds.

Middle to lower section: Possibly kitchen or wash area, where the sink is located. Side extensions may be separate sub-rooms or connecting corridors.

#### 🕒 Navigation Agent

The red arrow represents the current location and orientation of the agent. The agent is in the upper-middle part, just below the three beds, facing upward, possibly toward the beds.

#### 🏃 Agent's Trajectory

The red line marks the agent's travel path: Starts at the bottom, near the sink. Moves upward through the narrow hallway. Veers briefly left and right at a junction (suggesting room exploration). Ends at the upper center, near the bed cluster, where the agent is currently located.

Figure 34: Illustration of GPT-4o’s map reasoning capabilities on annotated semantic maps from MapNav. The model accurately identifies semantic objects, their spatial relationships, room functions, and the agent’s position and trajectory. Check Section 4.2 for detailed discussions.

## Navigation Map Reasoning

### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### GPT-4o:

#### Objects in the Map

Clock – Marked near the top center-left.

Sink – Located toward the left-central region.

Oven – Placed in the right-central part of the map.

#### Object Location Distribution

The sink and clock are positioned toward the left half of the map: Sink is near the middle-left.

Clock is located slightly above and right of the sink, possibly on a wall or shelf.

The oven is further to the right, in the middle-right area, likely part of a kitchen.

#### Room Properties

The map suggests a linear layout, such as a hallway or open-plan design: The left section likely includes a kitchen or washroom (sink and clock present). The right section includes cooking appliances (oven), again supporting a kitchen zone.

The segmented gray structures imply distinct rooms or obstacles (e.g., counters, walls, or dividers).

#### Navigation Agent

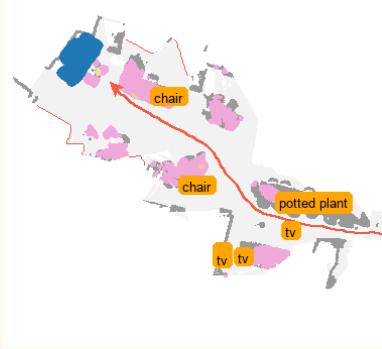
The red arrow indicates the current position and orientation of the agent. The agent is located in the far-right area, facing slightly upward/left, near the oven.

#### Agent's Trajectory

The red line shows a long path the agent has traveled: Starting from the far left, likely near an entrance or another object not labeled, Moving steadily through the center-left area near the sink and clock, Then continuing toward the oven, where the agent now resides.

### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### GPT-4o:

#### Objects in the Map

TVs (x3): Labeled in yellow, clustered in the lower right.

Chairs (x2): Positioned in central and upper-middle areas.

Potted Plant: Located near the TVs, on the mid-right side.

#### Object Location Distribution

TVs: Three instances located toward the bottom-right quadrant, forming a dense entertainment area.

Chairs: One chair is in the center-left portion. Another is near the upper-middle part of the map.

Potted Plant: Located in the mid-right zone, adjacent to the TV cluster, suggesting decorative furnishing in the living area.

#### Room Properties

The environment is elongated and multi-zoned, likely including: Entertainment area on the right side (TVs + plant).

Seating areas in the middle. Possibly a bedroom or bathroom on the left side, as suggested by the large blue patch, which could be a bed or tub (not labeled but visually prominent). The pink/purple shapes may indicate object affordances (e.g., interactive zones or detected regions).

#### Navigation Agent

The red arrow represents the current position and facing direction of the agent. The agent is in the upper-middle area, near the top chair, facing upward-left.

#### Agent's Trajectory

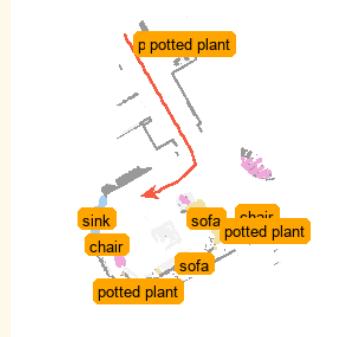
The red line traces a complex navigation path: Starts from the bottom-right, passing through the TV and plant zone, Moves diagonally upward and left, passing another chair. Ends near the upper-middle chair, where the agent currently is.

Figure 35: Illustration of GPT-4o’s map reasoning capabilities on annotated semantic maps from MapNav. The model accurately identifies semantic objects, their spatial relationships, room functions, and the agent’s position and trajectory. Check Section 4.2 for detailed discussions.

## Navigation Map Reasoning

### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### GPT-4o:

#### Objects in the Map

Potted Plants (x4) – Several are distributed throughout the space.

Sofas (x2) – Positioned centrally.

Chairs (x2) – One near the sofas, another in the upper-right area.

Sink – Located toward the left-central portion.

#### Object Location Distribution

Potted Plants: One in the upper-left corner (labeled “p potted plant”). Three clustered in the lower half, forming part of a well-decorated seating area.

Sofas: Both are located in the lower-central area, forming a typical lounge setup.

Chairs: One chair is placed near the sofas, possibly as extra seating. Another is to the right, near a potted plant, suggesting a separate sitting nook.

Sink: Found in the middle-left area, suggesting a kitchen or wet bar adjacent to the living area.

#### Room Properties

The map shows a partially open, L-shaped layout, suggesting:

A living/social zone in the bottom portion (sofas, plants, chair).

A transition or hallway area leading to the top-left. The sink may indicate a small kitchen or utility area branching off the main room.

#### Navigation Agent

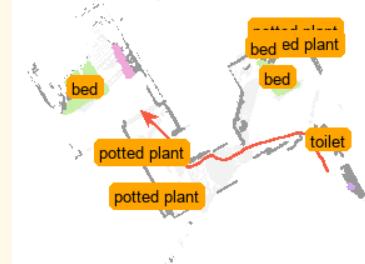
The red arrow shows the agent's current location and facing direction. The agent is located in the central upper-left region, just below the top-left potted plant, facing downward and slightly right.

#### Agent's Trajectory

The red line shows the agent's navigation path: Starts in the lower portion, near the sofas and multiple objects. Moves upward, through a corridor-like path. Ends near the top-left potted plant, where the agent currently resides.

### Prompt:

Please describe this navigation annotated semantic map, including what objects it contains, the location distribution of the objects, the room properties, the location of the navigation agent represented by the red arrow, and the past trajectory of the navigation agent represented by the red line.



### GPT-4o:

#### Objects in the Map

Beds (x2) – Indicating two sleeping zones.

Potted Plants (x3) – Scattered across the space, suggesting interior decoration.

Toilet – Identified in the lower-right part of the map.

#### Object Location Distribution

Beds: One bed is placed in the top-right corner, alongside a potted plant (partially labeled as “ed plant”). Another bed is located in the left-central part of the map.

Potted Plants: One near the top-right bed. One centrally located. One more in the lower-middle region, forming a line toward the center.

Toilet: Located in the bottom-right corner of the map, most likely marking a dedicated bathroom area.

#### Room Properties

The layout is branched and partially open, suggesting:

Two bedroom areas on opposite sides (left and right). A central zone with decorative potted plants. A corridor-like connector between rooms. A bathroom zone at the bottom-right (toilet present).

#### Navigation Agent

The red arrow represents the agent's current location and facing direction. The agent is now positioned centrally, near two potted plants, and is facing upward, toward the left bed.

#### Agent's Trajectory

The red line traces the navigation path: Starts at the bottom-right, near the toilet. Moves left through a hallway-like connector. Passes near a potted plant and approaches the central region. Ends in front of the central potted plants, where the agent currently resides.

Figure 36: Illustration of GPT-4o’s map reasoning capabilities on annotated semantic maps from MapNav. The model accurately identifies semantic objects, their spatial relationships, room functions, and the agent’s position and trajectory. Check Section 4.2 for detailed discussions.

### 4.3 Navigation Trajectory Inference

In this section, we evaluate GPT-4o’s ability to predict and visualize future navigation trajectories based on first-person observations and directional instructions. We provided GPT-4o with a first-person RGB image and a trajectory instruction, then asked the model to plan and visualize the path directly on the image.

One of the prompt used in our evaluation was:

*“If I want to go through this place and turn left at the cabinet ahead, please help me plan the path and indicate it with lines on this picture.”*

As shown in Figure 37-40, our analysis of GPT-4o’s responses reveals both capabilities and limitations in navigation trajectory inference. The model demonstrates several interesting properties:

First, GPT-4o successfully plans approximately accurate trajectories that follow the given directional instructions. The model’s ability to project potential paths from first-person perspective suggests a fundamental understanding of spatial relationships and navigation planning. The general direction of the planned trajectories aligns with the instructions, indicating that GPT-4o can interpret directional guidance in the context of visual observations.

However, we observed several limitations in GPT-4o’s implementation:

Rather than annotating the original image as requested, GPT-4o regenerates the entire image with the trajectory overlay. While this regenerated image often has higher resolution than the original, it frequently contains inconsistencies and inaccuracies compared to the source image. The model sometimes modifies environmental elements or generates entirely new features not present in the original observation.

More concerning for navigation applications, GPT-4o demonstrates incomplete perception of critical environmental constraints. The model shows limited awareness of walls and other obstacles that would impede navigation, sometimes planning trajectories that would require passing through solid barriers. Similarly, the model struggles with accurately perceiving and planning around stairs and multi-level structures, suggesting limitations in understanding 3D spatial relationships from 2D images.

Despite these issues, the overall trajectory prediction quality is promising. The model’s ability to generate spatially coherent paths from first-person views, even with limitations in environmental constraint perception, indicates potential for future developments in vision-language models for navigation planning.

## Navigation Trajectory Inference

### Prompt:

If I want to go through this place and turn left at the cabinet ahead, please help me plan the path and indicate it with lines on this picture.



### Prompt:

If I want to keep going and stop after going through the door, please help me map out the path and show it with lines on this picture.



### GPT-4o:



### GPT-4o:



Figure 37: Illustration of GPT-4o's trajectory inference capabilities on first-person navigation images. When provided with directional instructions such as “turn left at the cabinet” or “stop after going through the door”, the model generates visual path predictions with colored arrows. Check Section 4.3 for detailed discussions.

## Navigation Trajectory Inference

### Prompt:

If I want to keep going and go through the door, please help me map out the path and show it on this picture with lines.

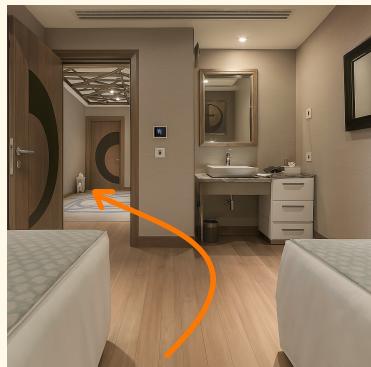


### Prompt:

If I want to continue through this area and enter the door on the left, please help me map out the path and draw lines on this picture.



### GPT-4o:



### GPT-4o:

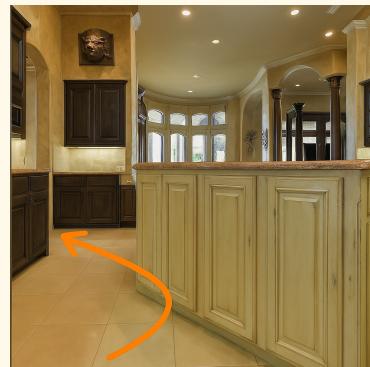


Figure 38: Illustration of GPT-4o’s trajectory inference capabilities on first-person navigation images. When provided with directional instructions such as “turn left at the cabinet” or “stop after going through the door”, the model generates visual path predictions with colored arrows. Check Section 4.3 for detailed discussions.

## Navigation Trajectory Inference

### Prompt:

If I want to continue through this door and stop in front of the mirror at the end, please help me plan the path and indicate it with lines on this picture.

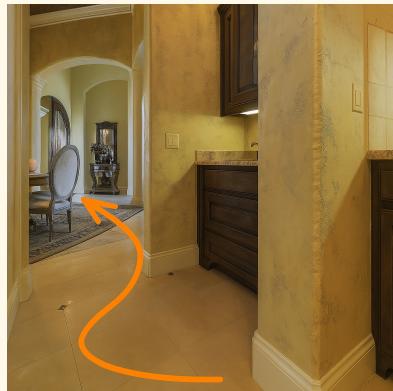


### Prompt:

If I want to keep going and go down to the next level via the stairs ahead, please help me map out the path and show it with lines on this picture.



### GPT-4o:



### GPT-4o:

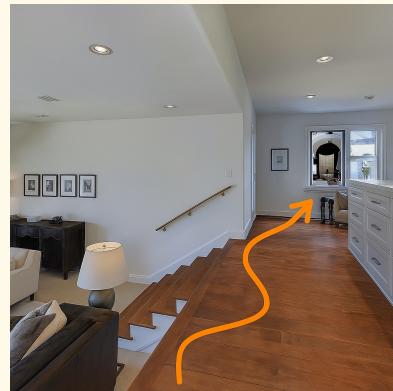


Figure 39: Illustration of GPT-4o’s trajectory inference capabilities on first-person navigation images. When provided with directional instructions such as “turn left at the cabinet” or “stop after going through the door”, the model generates visual path predictions with colored arrows. Check Section 4.3 for detailed discussions.

## Navigation Trajectory Inference

### Prompt:

If I want to go through this area and turn left at the end, please help me map out the path and show it on this picture with lines.



### Prompt:

If I want to enter the room through the door on the front right and stop in front of the sofa, please help me plan the path and show it with lines on this picture.



### GPT-4o:



### GPT-4o:



Figure 40: Illustration of GPT-4o's trajectory inference capabilities on first-person navigation images. When provided with directional instructions such as “turn left at the cabinet” or “stop after going through the door”, the model generates visual path predictions with colored arrows. Check Section 4.3 for detailed discussions.

#### 4.4 Navigation Action Prediction

In this section, we evaluate GPT-4o’s capability to function directly as a navigation agent by predicting the next action based on a first-person perspective image and navigation instructions. This assessment tests the model’s ability to integrate visual perception, language understanding, and action planning in real-time navigation scenarios.

We provided GPT-4o with first-person RGB observations and VLN instructions using the following prompt format:

*“You are now a robot that is performing a navigation task. Your task is to follow this instruction: Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it. Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.”*

As shown in Figure 41- 44, our analysis reveals significant limitations in GPT-4o’s action prediction capabilities. While the model demonstrates reasonable understanding of the overall navigation goal and can conceptualize approximate future trajectories, its ability to select appropriate immediate actions is considerably weaker.

The actions predicted by GPT-4o frequently diverge from ground truth actions, showing poor alignment with optimal navigation strategies. This suggests a disconnect between the model’s high-level spatial understanding and its ability to translate this understanding into concrete, step-by-step navigation commands.

This performance gap indicates that while GPT-4o possesses strong perception capabilities—evidenced by its ability to interpret environments and understand navigation goals—it struggles with the specific requirements of embodied navigation planning. The model appears to lack the precise spatial reasoning needed to determine optimal immediate actions that would efficiently advance the agent toward its goal.

These findings highlight an important distinction between general spatial understanding and the more specialized skills required for effective embodied navigation. While GPT-4o can interpret environments and conceptualize paths, the translation of this understanding into discrete, optimal navigation actions remains a significant challenge for the model.

## Navigation Trajectory Inference

### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it. Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it. Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### GPT-4o:

Based on the image you provided, I can see that there is a center unit in the middle and an open area on the right where a dining table and mirror can be seen.

In order to follow the instruction "Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it", I need to go around the right side of the center unit and approach the doorway on the right with the dining table and mirror.

Next action: **Move forward** (to go around the right side of the center unit).

**Ground Truth: Turn Right.**

### GPT-4o:

According to the image you provided, I have now bypassed the center console and can see a right doorway area ahead with a dining table and mirror. In order to continue the mission instruction "Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it", I have reached the right doorway of the target location.

Next action: **Stop.**

**Ground Truth: Move Forward.**

Figure 41: Illustration of GPT-4o's action prediction capabilities in VLN tasks. When provided with navigation instructions and first-person observations, the model attempts to interpret the environment and determine appropriate next actions. The results suggest limitations in translating visual perception into precise navigation decisions despite generally understanding the high-level goal. Orange shows the ground truth action. Check Section 4.4 for detailed discussions.

## Navigation Trajectory Inference

### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it. Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Exit the beginning room and enter the open area. Turn right and enter the farthest right room. Stop right in front of the table. Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### GPT-4o:

Based on the image you provided, it seems that I have successfully reached the right doorway and can see the dining table and mirror in the doorway area.

Therefore, my task is completed.

Next action: **Stop**.

Ground Truth: **Stop**.

### GPT-4o:

Based on the image you provided, the room I'm currently in looks like a bathroom or bedroom. I need to exit this room and enter the open area, then turn right and enter the far right room and stop at the table. Next action: **Move forward**, exit the current room and enter the open area.

Ground Truth: **Turn Left**.

Figure 42: Illustration of GPT-4o’s action prediction capabilities in VLN tasks. When provided with navigation instructions and first-person observations, the model attempts to interpret the environment and determine appropriate next actions. The results suggest limitations in translating visual perception into precise navigation decisions despite generally understanding the high-level goal. Orange shows the ground truth action. Check Section 4.4 for detailed discussions.

## Navigation Trajectory Inference

### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Go around the right side of the center unit and stop by the right side doorway with the dining table and mirror in it. Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Exit the beginning room and enter the open area. Turn right and enter the farthest right room. Stop right in front of the table. Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### GPT-4o:

Based on the image you provided, it appears that I have entered the open area and can see doorways with multiple rooms. At this point, I need to turn right and enter the farthest right room. Next move: **Turn right** and enter the farthest right room.

**Ground Truth: Move Forward.**

### GPT-4o:

Based on the image you provided, I am currently in a room that looks like a bathroom, but it is not the farthest right room. Therefore, I need to adjust my direction, leave this room and enter the farthest right room.

Next action: **Turn left**, leave the current room and enter the right room.

**Ground Truth: Stop.**

Figure 43: Illustration of GPT-4o’s action prediction capabilities in VLN tasks. When provided with navigation instructions and first-person observations, the model attempts to interpret the environment and determine appropriate next actions. The results suggest limitations in translating visual perception into precise navigation decisions despite generally understanding the high-level goal. Orange shows the ground truth action. Check Section 4.4 for detailed discussions.

## Navigation Trajectory Inference

### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Walk into the elevator area. Walk out of the elevator and down the hall into the kitchen area. Walk around the kitchen counter tops towards the living room area. Walk up to the backside of the couches near the table lamp.

Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### Prompt:

You are now a robot that is performing a navigation task. Your task is to follow this instruction: Exit the beginning room and enter the open area. Turn right and enter the farthest right room. Stop right in front of the table.

Now your first-person perspective is this picture. Please tell me your next move. The optional moves are: Move forward, turn left, turn right, and stop.



### GPT-4o:

Based on the image you provided, my current location looks like a hallway with multiple doorways in front, and some furniture and decorations can be seen.

To continue the mission instructions, I need to enter the elevator area, then walk out of the elevator, into the hallway and into the kitchen, and finally go around the kitchen counter to the living room area, and finally reach the table lamp behind the sofa.

Next action: **Move forward** and enter the elevator area.

**Ground Truth: Move Forward.**

### GPT-4o:

According to the image you provided, I am now close to the kitchen area, with a wide space in front of me where I can see furniture and a dining table.

According to the mission instructions, I need to go around the kitchen countertop and walk towards the living room area, eventually approaching the table lamp behind the sofa.

Next action: **Move forward**, continue walking towards the kitchen, and move towards the living room.

**Ground Truth: Move Forward.**

Figure 44: Illustration of GPT-4o’s action prediction capabilities in VLN tasks. When provided with navigation instructions and first-person observations, the model attempts to interpret the environment and determine appropriate next actions. The results suggest limitations in translating visual perception into precise navigation decisions despite generally understanding the high-level goal. **Orange** shows the ground truth action. Check Section 4.4 for detailed discussions.

## 5 Conclusions

### 5.1 Capabilities and Limitations of GPT-4o in Embodied Intelligence

In this paper, we conducted a comprehensive evaluation of GPT-4o across various scenarios in Embodied Intelligence. The results indicate that GPT-4o exhibits capabilities that have the potential to surpass those of existing Embodied Intelligence in aspects such as embodied reasoning, embodied manipulation, and embodied navigation.

- **Embodied Reasoning** GPT-4o exhibits notable strengths in multimodal comprehension and embodied reasoning tasks. The model demonstrates a strong ability to integrate visual and textual modalities, leveraging this fusion to support robust commonsense inference and goal-oriented understanding. It shows promising performance in perceiving physical scales, such as estimating relative object sizes and distances, and maintains temporal coherence when reasoning over sequences of actions or events. These capabilities allow the model to engage in context-aware and logically grounded reasoning, especially in tasks that require aligning perceptual input with embodied knowledge. Nonetheless, GPT-4o encounters several limitations in more complex embodied reasoning scenarios. When operating in simulated environments lacking extensive fine-tuning data, the model’s comprehension becomes less reliable, often failing to accurately interpret scene-specific cues. It also struggles with spatial imagination tasks, such as mentally projecting itself into a particular location or perspective within a 3D scene. Moreover, in reasoning tasks that require iterative exploration—such as those demanding trial-and-error deduction or long-horizon planning—the model’s predictions often lack robustness, leading to suboptimal or implausible conclusions. Overall, GPT-4o demonstrates a commendable grasp of high-level embodied reasoning, supported by its multimodal fusion and commonsense capabilities. However, its performance still lags in situations demanding fine-grained spatial inference, scene-specific grounding, and iterative reasoning. These gaps suggest that future improvements may depend on enhanced simulation grounding and embodied learning from interactive experiences.
- **Embodied Manipulation** GPT-4o demonstrates strong capabilities in high-level cognitive reasoning for embodied manipulation tasks. The model performs well in object recognition, accurately identifying common items and inferring their attributes. It shows competent understanding of affordances by recognizing actionable object parts and basic spatial relationships. For task planning, GPT-4o effectively decomposes complex goals into logical sequences while considering object dependencies and safety constraints, and can generate reasonable action sequences for trajectory prediction. However, the model shows significant limitations in precise spatial reasoning required for embodied manipulation. It struggles with accurate object localization and size estimation, particularly in three-dimensional environments. The model’s affordance grounding becomes unreliable when predicting specific object parts or free spaces in complex scenes. Additionally, it sometimes makes unrealistic assumptions during task planning by overlooking critical scene information. Most notably, its trajectory predictions often fail to account for environmental obstacles and spatial constraints, demonstrating limited adaptability in dynamic settings. These findings reveal GPT-4o’s characteristic “strong in semantics but weak in execution” performance pattern for embodied manipulation. While excelling at high-level planning and understanding, the model underperforms in tasks requiring precise physical interaction. This gap highlights the need for integrating multimodal sensory data to enhance the model’s execution capabilities for practical embodied applications.
- **Embodied Navigation** GPT-4o demonstrates varied capabilities across different aspects of embodied navigation. Its primary strengths lie in structured spatial reasoning tasks, particularly when interpreting annotated semantic maps where it accurately identifies objects, infers room properties, and tracks agent positions. It also shows promising ability in trajectory planning from first-person views, generally understanding directional instructions and projecting potential paths. However, GPT-4o exhibits significant limitations that impact its effectiveness for comprehensive embodied navigation. It struggles with temporal-spatial reasoning from video sequences, often missing key semantic objects and failing to construct coherent trajectory summaries. When planning trajectories, it tends to regenerate images rather than annotate them directly and sometimes ignores physical constraints like walls. Most critically, despite understanding high-level navigation goals, GPT-4o performs poorly

at predicting optimal immediate actions in navigation scenarios, revealing a disconnect between its conceptual spatial understanding and the practical requirements of embodied navigation planning. These findings suggest that while GPT-4o possesses strong general visual-spatial perception capabilities, it lacks the specialized reasoning required for effective embodied navigation, particularly in translating environmental understanding into precise sequential actions. This highlights the gap between general spatial comprehension and the specific skills needed for autonomous navigation in complex environments.

Overall, the performance of GPT-4o demonstrates the significant potential of Vision-Language Models (VLMs) to tackle complex challenges in the field of Embodied Intelligence.

## 5.2 Future Work

Future research should focus on addressing the current limitations of GPT-4o in Embodied Intelligence by exploring several key directions. First, enhancing the model's spatial reasoning and fine-grained perception through improved multimodal training on 3D environments and interactive simulations could bridge the gap between high-level planning and precise execution. Second, incorporating reinforcement learning from embodied interactions may strengthen the model's ability to perform iterative reasoning and adapt to dynamic scenarios. Additionally, developing specialized architectures that integrate real-time sensory feedback could improve navigation and manipulation capabilities in physical or simulated environments. Finally, advancing the model's temporal coherence and long-horizon planning through memory mechanisms and hierarchical task decomposition would further solidify its applicability in complex embodied tasks. These improvements could unlock new possibilities for deploying VLMs in robotics, autonomous systems, and other embodied AI applications.

## References

- [1] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.
- [2] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024.
- [3] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.
- [4] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.
- [5] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7606–7623, 2022.
- [6] Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*, 2025.
- [7] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025.
- [8] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120, 2020.
- [9] Dingzhe Li, Yixiang Jin, Yuhao Sun, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Fuchun Sun, Jianwei Zhang, et al. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*, 2024.
- [10] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.
- [11] OpenAI. Introducing 4o image generation. <https://openai.com/index/introducing-4o-image-generation/>, 2025.
- [12] Sang-Min Park and Young-Gab Kim. Visual language navigation: A survey and open challenges. *Artificial Intelligence Review*, pages 365–427, 2023.
- [13] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.
- [14] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.
- [15] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778*, 2025.

- [16] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.
- [17] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3554–3560, 2023.
- [18] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.
- [19] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. In *Proceedings of Robotics: Science and Systems*, 2024.
- [20] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451*, 2025.
- [21] Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. Multi-floor zero-shot object navigation policy. *arXiv preprint arXiv:2409.10906*, 2024.
- [22] Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. Trihelper: Zero-shot object navigation with dynamic assistance. *arXiv preprint arXiv:2403.15223*, 2024.
- [23] Qiang Zhang, Zhang Zhang, Wei Cui, Jingkai Sun, Jiahang Cao, Yijie Guo, Gang Han, Wen Zhao, Jiaxu Wang, Chenghao Sun, et al. Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots. *arXiv preprint arXiv:2503.09010*, 2025.
- [24] Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*, 2025.