# EMBODIEDEVAL: Evaluate Multimodal LLMs as Embodied Agents

Anonymous ACL submission

## Abstract

Multimodal Large Language Models (MLLMs) have shown significant advancements, providing a promising future for embodied agents. Existing benchmarks for evaluating MLLMs primarily utilize static images or videos, limiting assessments to non-interactive scenarios. Meanwhile, existing embodied benchmarks are task-specific and not diverse enough, which do not adequately evaluate the embodied capabilities of MLLMs. To address this, we propose EMBODIEDEVAL, a challenging and comprehensive benchmark to evaluate MLLMs' interactive capabilities in embodied tasks within a unified simulation and evaluation framework tailored for MLLMs. We evaluate the state-of-the-art MLLMs on EMBODIEDEVAL and find that they have a significant shortfall compared to human level on embodied tasks. Our analysis demonstrates the limitations of existing MLLMs in embodied capabilities, providing insights for their future development.


Figure 1: Examples of the five task categories and performance overview of EMBODIEDEVAL. The embodied agent powered by MLLMs is required to finish the given task in a 3D simulation environment.

## 1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) (OpenAI, 2023; Team et al., 2023; Liu et al., 2024a) have demonstrated strong capabilities in understanding and reasoning across vision and language tasks. With the rapid development of MLLMs, a rich set of benchmarks (Yue et al., 2024; Liu et al., 2023b; Fu et al., 2023; Li et al., 2023a) has been developed. Beyond these basic tasks that focus on non-interactive visual scenes, researchers are actively trying to expand MLLMs as embodied agents in interactive environments, which require the model to interpret multimodal inputs into actions (Ahn et al., 2022; Driess et al., 2023; Mu et al., 2024). To accomplish this, MLLMs are expected to integrate a multitude of capabilities that enable them to interact effectively with the environment, including ego-centric perception (Cheng et al., 2024a), visual grounding (Anderson et al., 2018b; Zhang et al., 2024b), spatial reasoning (Chen et al., 2024) and episodic memory (Majumdar et al., 2024).

However, the comprehensive evaluation of MLLMs in embodied tasks remains largely unexplored. First, existing benchmarks for embodied tasks lack diversity in both tasks and scenes. For instance, ALFRED (Shridhar et al., 2020) includes just seven predefined tasks (e.g., "pick and place") within four room types. Second, several benchmarks (Anderson et al., 2018b; Qi et al., 2020a) impose rigid input-output formats, e.g. 3D points, making it inefficient or even infeasible to evaluate mainstream MLLMs. Third, certain benchmarks (Li et al., 2024; Liu et al., 2023a; Jia et al., 2024) try to evaluate LLMs' embodied perfor-

| Benchmark | Scene. | Task. | Disc. | Ego. | Nav. | Obj. | So. | Ans. |
|---|---|---|---|---|---|---|---|---|
| MME (Fu et al., 2023) | - | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| EgoPlan etc. (Chen et al., 2023b; Cheng et al., 2023) | - | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ |
| OpenEQA (Majumdar et al., 2024) | - | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ |
| EQA etc. (Das et al., 2018; Yu et al., 2019; Tan et al., 2023) | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |
| ALFRED (Shridhar et al., 2020) | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ |
| BEHAVIOR(Srivastava et al., 2022) | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ |
| EQA-MX (Islam et al., 2024) | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ |
| EMBODIEDEVAL | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison of EMBODIEDEVAL with previous benchmarks. The abbreviations in the table headers, from left to right, represent: **Scene** diversity (beyond household scenes), **Task** diversity (beyond task templates), **Disc**rete action space (for MLLMs evaluation), **Ego**centric vision, **Nav**igation involved, **Obj**ect interaction involved, **So**cial interaction involved, and **Ans**wering questions involved.

mance by representing environments with textual descriptions, relying heavily on text-based states. This downplays critical embodied skills such as visual grounding and spatial reasoning, which are essential for real-world interaction.

To address this gap, we introduce the first comprehensive benchmark for evaluating MLLMs' embodied capabilities in interactive environments. The key features of EMBODIEDEVAL are as follows: (I) **Diverse Interactions.** EMBODIEDEVAL provides a simulation framework that supports a wide range of interactions with objects and humans in realistic 3D environments. Agents need to interact with the environment to gather information or alter its state to complete tasks. The ego-centric visual information will serve as input to the MLLMs to make the decision. (II) **Diverse Tasks.** Unlike previous work that relied on predefined task templates, our tasks are systematically generated and carefully selected to ensure both high quality and diversity. EMBODIEDEVAL includes novel tasks that assess a broad spectrum of abilities, enabling a more comprehensive evaluation of the model's capabilities. (III) **Diverse Scenes.** Our scenes offer significant diversity in terms of objects and spaces, encompassing house rooms, large residences, and public areas such as gyms, stores, and offices. This variety helps minimize the impact of specific scene types on the model's generalization.

Experiment results on EMBODIEDEVAL reveal that mainstream MLLMs largely fall short of human-level performance on embodied tasks. Model performance varies widely across different task categories, with a notable drop in spatial and long-horizon tasks. EMBODIEDEVAL provides insights for further improvements in MLLMs's capability in grounding, spatial reasoning, planning, and exploration.

## 2 Related Works

**Multimodal Large Language Models.** By connecting vision modules with LLMs, LLaVA (Liu et al., 2024a) pioneers research in MLLMs through visual instruction tuning. Many work further improves the MLLMs from various aspects, including detailed captioning (Chen et al., 2023a), trustworthy response (Yu et al., 2024a,b), multilingual capabilities (Hu et al., 2023), visual grounding (Peng et al., 2023; You et al., 2023) and video understanding (Lin et al., 2023; Liu et al., 2024c).

**Evaluation for MLLMs.** Mainstream benchmarks for MLLMs mainly focus on perception and cognitive evaluation (Fu et al., 2023; Liu et al., 2023b; Yue et al., 2024; Fu et al., 2024) and some benchmarks focus on more challenging tasks (Lu et al., 2024; He et al., 2024; Singh et al., 2019; Liu et al., 2023c; Yang et al., 2024a; Yue et al., 2024). Additionally, certain benchmarks (Fan, 2019; Chen et al., 2023b; Cheng et al., 2023; Majumdar et al., 2024; Szot et al., 2023) have been designed to evaluate the egocentric capabilities of MLLMs using egocentric images or videos. However, these benchmarks use static question-answering pairs without interacting with environments.

**Benchmarks for Embodied Agents.** Existing benchmarks or datasets for embodied agents cover several areas such as embodied question answering (Das et al., 2018; Yu et al., 2019; Tan et al., 2023; Ren et al., 2024; Islam et al., 2024; Dorbala et al., 2024; Gordon et al., 2018), navigation (Anderson et al., 2018b; Jain et al., 2019; Ku et al., 2020; Zhu et al., 2021; Qi et al., 2020b; Ma et al., 2024; Khanna et al., 2024b) and object interaction (Shridhar et al., 2020; Batra et al., 2020a; Weihs et al., 2021; Kant et al., 2022; Misra et al., 2018; Srivastava et al., 2022; Li et al., 2023b).

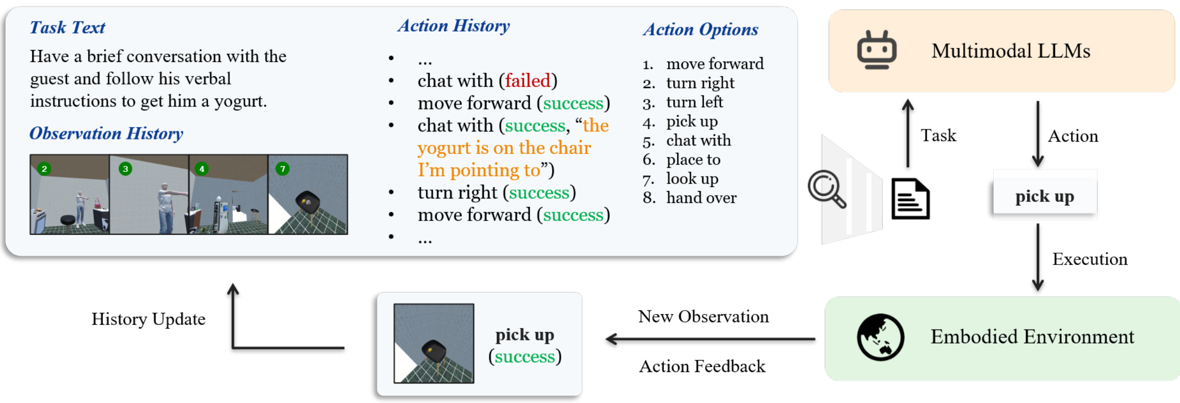



Figure 2: The evaluation process of EMBODIEDEVAL. The task description and ego-centric observation history will be input for the model. The environment will respond to the action from the model output with a new observation.

However, existing embodied benchmarks are limited in task variety, lacking comprehensive assessments of navigation, object interaction, and question-answering. They rely on predefined task templates, failing to adequately capture the wide spectrum of embodied capabilities. Additionally, the task-specific observation spaces and continuous action spaces in many benchmarks are inadequate for effectively evaluating MLLMs. We summarize the comparison between EMBODIEDEVAL and other representative benchmarks in Table 1.

## 3 EmbodiedEval

In this section, we introduce the evaluation process and data collection process of EMBODIEDEVAL.

### 3.1 Evaluation Formulation

EMBODIEDEVAL utilizes LEGENT (Cheng et al., 2024c) platform as simulator, which provide a rich and interactive 3D environment with communicable and actionable agents. We formulate the evaluate process as a decision making problem. As shown in Algorithm 1 and Figure 2, an evaluation episode unfolds as follows: (1) The simulator initializes the 3D scene $x$. The agent $\pi$, powered by the evaluated MLLM, is positioned at a designated starting point, and the initial **first-person** visual observation $o^{(0)}$ provided by the environment is saved into the observation history $\mathcal{M}_o = \{o^{(0)}\}$. (2) At each step $i$, the agent $\pi$ chooses an action $a^{(i)}$ from a given list of options $\mathcal{C}$, which includes movement, interaction and answering, based on both the observation history $\mathcal{M}_o$ and action history $\mathcal{M}_a$. The environment executes the action, changes the state accordingly, and returns new visual observations $o^{i+1}$, along with feedback indicating whether the action was successful. The observation, action, and feedback are then appended to the observation history. (3) This process continues until either all success criteria are met, leading to task completion, or the task fails due to an incorrect answer or exceeding the maximum allowed steps. Task success is determined by the environment based on a set of predefined predicates, which maps the state of the simulation environment to a boolean value indicating success. Further details on the success criteria can be found in Appendix B.4.

**Algorithm 1** EMBODIEDEVAL Evaluation Process

**Input:** A Multimodal LLM $\pi$, a scene $x$, a task description $g$, an option list $\mathcal{C} = \{a_0, a_1, ..., a_n\}$, and a predicate list $\mathcal{P}$.

**Output:** A boolean indicating whether the task was successful $success$.

```
o, s ← E.reset(x)    ▷ E is the simulator, o is the visual observation, s is the world state
M_o ← {o}            ▷ observation history
M_a ← ∅              ▷ action history
for i ← 0 to max steps do
    a ← π.predict(g, C, M_o, M_a)
    o, s ← E.step(a)
    M_o.append(o)
    M_a.append(a)
    done ← P.judge(s)
    if done then
        return true
    end if
end for
return false         ▷ reach the max steps
```

To holistically and effortlessly evaluate MLLMs' embodied capabilities with diverse tasks, rather than focusing on adapting to particular input-output requirements, we define a unified input and output

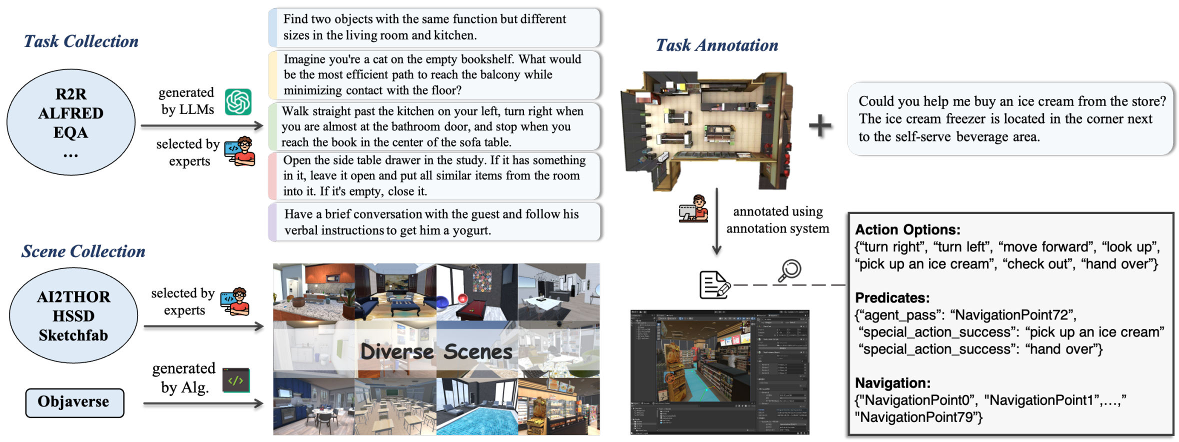



Figure 3: The dataset construction pipeline of EMBODIEDEVAL.

space. The input space consists of textual task descriptions $g$, action option $\mathcal{C}$, and egocentric visual observation $\mathcal{M}_o$ provided by the environment, without any additional environmental state information. This design choice emphasizes visual information as it is both the most accessible and the most general medium connecting the agent to the environment. Also, visual data is the more scalable source for training multimodal foundation models comparing with low-level data. Visual observations can take the form of multiple images representing different states or a video capturing the entire process of state transitions.

The output action space consists of movement, interaction, and answering, which varies in each task instance. For the **movement**, to make the evaluation feasible for current MLLMs, we constrain the movement space of agent on a navigation graph pre-constructed for each scene. MLLMs are not required to make choices from a set of 3D positions, but only need to make directional decisions among navigation points. The details of the movement space can be found in Appendix B.1

For the **interaction**, we utilize the high-level discrete interaction space. We use an open vocabulary for the actions and objects in interactions, where each action has a brief action text, operable objects, and conditions for successful interaction. For example, the "hand over" action requires the agent to hold an object and be next to a person. In a given test case, several interaction actions will be involved. The details of the interaction space can be found in Appendix B.2.

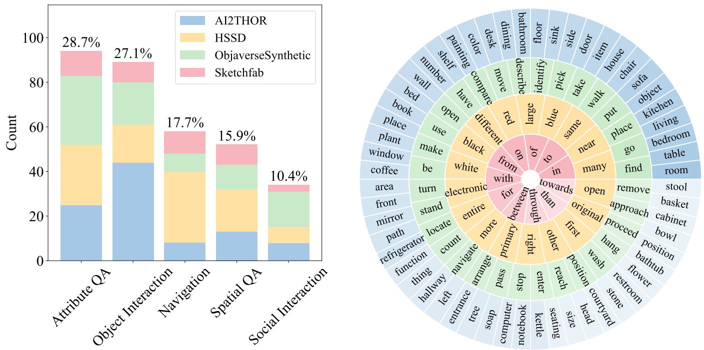
For the **answering**, the agent selects an answer from a set of annotated textual responses. It can continue exploring until it believes it has enough information to make a selection. Once an answer is chosen, the task is immediately judged as correct or incorrect. The options are challenging, closely related to the context, and of high quality, as shown in Appendix B.3.

## 3.2 Task Categories

EMBODIEDEVAL defines five task categories to comprehensively assess the embodied capabilities of MLLMs: (1) **Navigation** tasks involve coarse-grained and fine-grained natural language instructions, requiring the agent to navigate from its initial position to target locations or find specific objects. (2) **Object interaction** tasks require agents to modify the environment through direct interaction with objects, such as moving objects, opening or closing doors and drawers, and operating electrical devices. (3) **Social interaction** tasks involve human-agent interactions, including item delivery, perspective-taking, human feedback interpretation, and non-verbal expression comprehension. (4) **Attribute question answering** tasks require the agent to explore the environment and answer questions related to object and scene attributes. (5) **Spatial question answering** requires agents to answer spatial-related questions through actions and observations, such as queries about size, position, distance, layout, and spatial relationships. Each task type presents challenges that require the agent to integrate various capabilities such as grounding and reasoning. We show the samples from each category in Figure 1 and more detailed examples in Appendix A.

## 3.3 Benchmark Construction

The construction process of EMBODIEDEVAL consists of three parts: scene collection, task collection, and task annotation. Each sample in the dataset requires substantial effort and undergoes rigorous




| Statistics | Number |
|---|---|
| Total Tasks/Predicates | 328/575 |
| Total Options | 1533 |
| *Textual/Interaction* | 1213/320 |
| Total Scenes | 125 |
| Average Task Length | 16.09 |
| Average Option Length | 5.72 |
| Average Step Required | 10.72 |

Figure 4: Dataset statistics of EMBODIEDEVAL. Left: Number of tasks by category for each scene source. Middle: Visualization of vocabulary by part of speech and word frequency.

annotation. Figure 3 illustrates the overview of dataset construction pipeline.

#### 3.3.1 Scene Collection

We construct a diverse collection of scenes from four different sources: Objaverse (Deitke et al., 2023), AI2THOR (Kolve and et al., 2017), Habitat Synthetic Scenes Dataset (HSSD) (Khanna et al., 2024a) and Sketchfab[1]. We use Objaverse to generate indoor scenes, leveraging its extensive collection of 3D assets. First, we filter out unsuitable outdoor objects and manually review rendered assets to remove low-quality assets. See selected assets of Objaverse in Figure 9. Next, we employ ChatGPT to annotate object categories with their typical room placements and functions. Using the procedural generation methods (Deitke et al., 2022), we sequentially placed them based on their attributes. After the scene is generated, we further refine the scenes using a self-developed runtime scene editor. See more details about this synthetic process in Appendix D. In addition, we incorporate indoor room scenes with interactive objects from AI2THOR, and some public spaces, such as stores and supermarkets, from HSSD and Sketchfab. We organize all scenes into the same format.

#### 3.3.2 Task Collection

For task collection, we first gather seed tasks for each of the five task categories from over 30 existing datasets. Using these tasks as seeds, we prompt Claude and ChatGPT to generate diverse task examples. We ask the LLMs to incorporate various capabilities, including complex grounding, episodic memory, spatial reasoning, quantitative reasoning, common sense reasoning, and planning, which resulted in many novel tasks. From this extensive task pool, we select over 300 distinct tasks as the candidate task set. Rather than asking annotators to manually write tasks for given scenes, selecting generated tasks will ensure task diversity, avoid repetition, and reduce the dependency on individual annotators' creativity or preferences.

#### 3.3.3 Task Annotation

After generating the task candidate set, we conduct a manual annotation to finalize each sample. First, the annotators align a suitable scene to the task from the candidate pool. Second, the annotators configure the output space, including movement, interaction, and answering, as introduced in Section 3.1, and define the success criteria. Finally, the annotated tasks are running in the simulator to confirm that the tasks can be successfully finished. We recruit eight expert annotators to perform the annotations. Before beginning the annotation process, we provide systematic training on annotation requirements and system usage. To ensure high dataset quality, each annotated task is independently reviewed for correctness and accuracy by at least three evaluators. Additionally, we validate task feasibility by creating expert demonstrations for each task with expert participants and assessing human performance with non-expert participants. See more details about the annotation process, annotation system, and quality control in Appendix C.

### 3.4 Dataset Statistics

We summarize the statistics of EMBODIEDEVAL in Figure 4. EMBODIEDEVAL consists of 328 tasks in 5 categories across 125 unique scenes, 575 predicate instances, and 1533 varied options including 1213 textual answers and 320 interactions. Each episode requires 10.72 steps on average based on expert demonstrations. Task descriptions average 16.09 words in length, while options average 5.72 words. The left of Figure 4 shows the distribution

[1] https://sketchfab.com

| Model | Attr. QA | Spatial QA | Navigation | | | Object Interaction | | | Social Interaction | | | **Overall** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Succ. | Succ. | Succ. | GcS | SPL | Succ. | GcS | SPL | Succ. | GcS | SPL | Succ. | GcS |
| `Random` | 11.58 | 7.69 | 3.45 | 8.76 | 3.45 | 0.00 | 6.18 | 0.00 | 2.94 | 8.33 | 2.94 | 5.49 | 8.66 |
| `Human` | 98.95 | 92.31 | 96.55 | 97.84 | 82.28 | 97.75 | 99.44 | 90.73 | 100.00 | 100.00 | 89.96 | 97.26 | 97.94 |
| *Proprietary MLLMs* | | | | | | | | | | | | | |
| GPT-4o-Mini | 31.58 | 15.38 | 27.59 | 39.51 | 15.34 | 2.25 | 17.42 | 1.50 | 5.88 | 22.06 | 2.98 | 17.68 | 25.58 |
| GPT-4o | 35.79 | **32.69** | **31.03** | **42.53** | **22.23** | **10.11** | 24.25 | **5.94** | **11.76** | **26.72** | 6.74 | **25.00** | **32.42** |
| Gemini-Flash-1.5 | 26.32 | 13.46 | 5.17 | 17.10 | 3.51 | 2.25 | 7.58 | 0.96 | 2.94 | 12.50 | 1.47 | 11.59 | 16.13 |
| Gemini-Pro-1.5 | 27.37 | 9.62 | 17.24 | 25.86 | 9.78 | 4.49 | 12.36 | 3.00 | 5.88 | 18.14 | 3.44 | 14.33 | 19.26 |
| Qwen-VL-Max | **37.89** | 17.31 | 24.14 | 30.03 | 16.87 | 7.87 | **24.91** | 5.62 | 8.82 | 22.06 | **6.86** | 21.04 | 28.07 |
| *Open-Source Image MLLMs* | | | | | | | | | | | | | |
| InternVL2-8B | 13.68 | 13.46 | 8.62 | 18.25 | 4.04 | 0.00 | 7.43 | 0.00 | **5.88** | 18.63 | **2.45** | 8.23 | 13.27 |
| InternVL2-40B | 14.74 | 5.77 | 6.90 | 12.93 | 3.06 | 0.00 | 7.68 | 0.00 | **5.88** | **19.12** | 2.16 | 7.01 | 11.54 |
| InternVL2-Llama3-76B | 21.05 | 13.46 | 3.45 | 9.48 | 2.18 | 0.00 | 9.08 | 0.00 | 2.94 | 13.73 | 1.14 | 9.15 | 13.79 |
| LLaVA-OneVision-7B | 16.84 | 17.31 | 5.17 | 9.05 | 3.28 | 1.12 | 8.15 | 0.80 | 2.94 | 9.80 | 1.68 | 9.14 | 12.45 |
| LLaVA-NEXT-72B | 23.16 | 5.77 | **12.07** | 22.99 | **7.83** | **3.37** | **9.74** | **2.21** | 0.00 | 12.25 | 0.00 | 10.67 | 15.60 |
| LLaVA-OneVision-72B | **26.32** | **19.23** | 10.34 | **23.28** | 7.53 | 1.12 | 7.81 | 1.12 | 0.00 | 12.75 | 0.00 | **12.80** | **18.23** |
| VILA-8B | 15.79 | 9.62 | 1.72 | 8.91 | 0.96 | 0.00 | 3.46 | 0.00 | 2.94 | 6.37 | 1.68 | 6.71 | 9.27 |
| VILA-40B | 17.89 | 7.69 | 0.00 | 5.75 | 0.00 | 0.00 | 3.93 | 0.00 | 0.00 | 8.58 | 0.00 | 6.40 | 9.53 |
| *Open-Source Video MLLMs* | | | | | | | | | | | | | |
| LLaVA-Video-7B-Qwen2 | 20.00 | **19.23** | 3.45 | 4.89 | 1.88 | 1.12 | **8.80** | 0.27 | 0.00 | 5.15 | 0.00 | 9.76 | 12.63 |
| LLaVA-NEXT-Video-32B-Qwen | 21.05 | 7.69 | 6.90 | 14.08 | 5.34 | 0.00 | 8.61 | 0.00 | 2.94 | **12.01** | 0.98 | 8.84 | 13.39 |
| LLaVA-Video-72B-Qwen2 | **27.37** | 9.62 | **15.52** | **24.28** | **9.62** | 1.12 | 8.05 | 0.86 | 0.00 | 9.80 | 0.00 | 12.50 | **16.95** |
| Oryx-34B | 18.95 | 3.85 | 5.17 | 13.07 | 4.89 | 1.12 | 7.02 | 1.00 | 0.00 | 8.33 | 0.00 | 7.32 | 11.33 |
| VideoLLaMA2-7B | 21.05 | 9.62 | 6.90 | 17.53 | 4.88 | 0.00 | 1.63 | 0.00 | 2.94 | 7.35 | 1.38 | 9.20 | 11.99 |
| VideoLLaMA2-72B | **27.37** | 9.62 | 12.07 | 18.68 | 6.35 | **2.25** | 7.49 | **1.38** | **5.88** | 10.78 | **2.39** | **12.81** | 15.91 |

Table 2: Results of different models on EMBODIEDEVAL (%). Succ., GcS, and SPL mean success rate, goal-condition success, and success weighted by path length, respectively.

of the task across 5 task categories and 4 scene sources, the middle presents a visualization of frequent words categorized by grammatical type. See more task samples in Table 4.

## 4 Experiments

### 4.1 Experimental Setup

We evaluate 19 MLLMs on EMBODIEDEVAL, including proprietary MLLMs GPT-4o/4o-Mini (OpenAI, 2024), Gemini-Pro/Flash-1.5 (Team et al., 2023), Qwen-VL-Max (Bai et al., 2023), open-source image MLLMs Intern-VL-8B/40B/76B (OpenGVLab, 2024), LLaVA-OneVision-7B/72B (Zhang et al., 2024a), LLaVA-NEXT-72B (Zhang et al., 2024a), VILA-8B/40B (Lin et al., 2024), and open-source video MLLMs LLaVA-Video-7B/72B-Qwen2 (Zhang et al., 2024a), LLaVA-NEXT-Video-32B-Qwen (Zhang et al., 2024a), Oryx-34B (Liu et al., 2024d), VideoLLaMA2-7B/72B (Cheng et al., 2024b). Additionally, we introduce two special agents as reference: (1) the *Random* agent, which uniformly samples actions from the option set at each step, and (2) the non-expert *Human* agent, who is unfamiliar with the tasks and performs actions through the simulator's user interface using the same observation and action space as models. For visual observation history, EMBODIEDEVAL provides multiple ego-centric images at each step or, alternatively, videos capturing the entire interaction process. Proprietary and open-source image MLLMs use the former as input, while video MLLMs use the latter. We set the maximum number of attempt steps per task as 24. The image resolution is 448×448 and the field of view is 90 degrees. All models have the temperature set to 0 during evaluation as explained in Appendix E. We prompt the model to output thoughts before deciding options (Yao et al., 2022; Wei et al., 2022).

We evaluate agent performance using three metrics. (1) **Success Rate (Succ.)** (Liu et al., 2024b; Savva and et al., 2019; Kolve and et al., 2017) is the primary metric we use to measure the percentage of tasks that the agent fully completes. (2) **Goal-condition Success (GcS)** (Shridhar et al., 2020; Kim et al., 2023) measures partial success by calculating the proportion of goal conditions achieved, as specified by predicate functions. (3) **Success weighted by Path Length (SPL)** (Anderson et al., 2018a; Batra et al., 2020b) evaluates task execution efficiency in navigation and interaction tasks by considering both task success and the path efficiency relative to the expert demonstration.

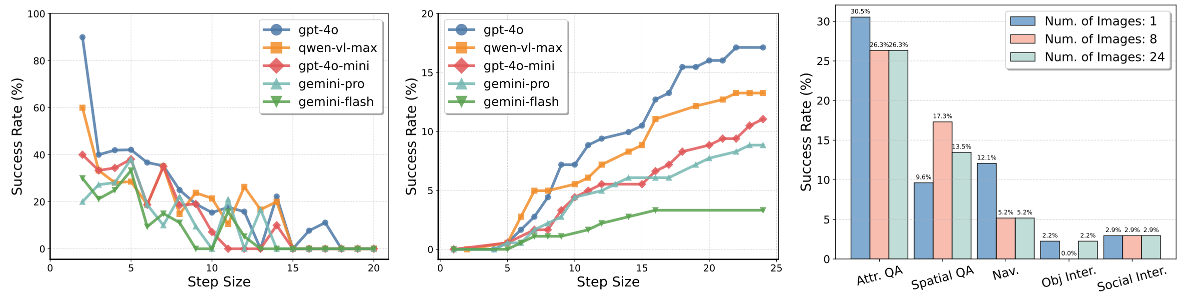



Figure 5: Left: Success rate vs. number of steps required for the task. Middle: Success rate vs. allowed max steps. Right: The success rate of Gemini-Flash with different number of images as input across five task categories.

### 4.2 Main Results

**Performance gap between current MLLMs on Embodied Tasks with Human.** As shown in Table 2, success rates across various models on EMBODIEDEVAL remain consistently low. The best-performing model, GPT-4o, achieves only a 25.00% overall success rate and a 32.42% GcS score. In contrast, non-expert humans reach a near-perfect success rate of 97.26%, highlighting the significant challenges these models face in executing embodied tasks that humans find trivial. This performance gap is further emphasized by lower SPL scores, indicating that the models struggle to find optimal solutions. The performance open-source models show a larger performance gap. The top performing MLLM, LLaVA-OneVision-72B, achieves an overall success rate of 12.80%, barely competitive with proprietary models.

**Model Performance across Different Task Types.** The results highlight a large variation in model performance across different task types. GPT-4o demonstrates relatively strong results in QA and Navigation tasks, but its performance drops notably for interaction tasks. This disparity is even more pronounced among other proprietary models. For instance, most models perform reasonably well in Attribute QA but see a sharp decline in Spatial QA that requires spatial reasoning, often halving their success rates. Overall, the scores for interaction tasks are consistently lower across all models, underscoring the challenge these models face in scenarios that require a deeper understanding of affordance (Gibson, 1977) or social cues.

### 4.3 Performance Analysis

**Challenges in Long-Horizon Tasks.** We shows the trend of success rate under tasks of varying steps required for finishing. Models maintain relatively high success rate in tasks that require fewer steps but shows a decline as the task length increases. The drop in performance can be attributed to the increased complexity of longer tasks and the difficulty in handling long context. The middle of Figure 5 shows the performance curve on the interaction tasks when gradually increasing the max allowed step from 1 to 25. While the success rate improves initially, the gain diminishes as the allowed steps increase, suggesting that the model struggles to manage longer histories effectively. In the right of Figure 5, we show the performance of Gemini-Flash across different tasks with context of varying number of input images. Although increasing the number of images theoretically provides more historical information, the performance decreases except for spatial question answering, which benefits from the additional spatial context. This result indicates that current multimodal models still face challenges when handling multiple egocentric image inputs. These results highlight the difficulty of long-horizon embodied tasks, where longer sequences complicate the agent's ability to plan and act based on historical information.

| Model | Inter. Freq (%) | Inter. Succ (%) |
|---|---|---|
| Random | 41.06 | 2.79 |
| Human | 19.44 | 96.81 |
| GPT-4o-Mini | 26.02 | 11.34 |
| GPT-4o | 40.46 | 9.56 |
| Gemini-Pro-1.5 | 11.46 | 10.03 |
| Gemini-Flash-1.5 | 11.03 | 8.6 |
| Qwen-VL-Max | 26.33 | 8.89 |

Table 3: Statistics of interaction tasks. **Inter. Freq** means interaction frequency and **Inter. Succ** means interaction success rate.

**Challenge in Interaction Tasks.** To further investigate the generally low performance in interaction tasks, we show some statistics of interaction task in Figure 3. Interaction frequency measures the proportion of interaction actions among all executed actions, while interaction success rate reflects how often these actions are invoked under the correct conditions, indicating the model's affordance judgment ability. Humans can generally ensure that only necessary interactions are performed, while models exhibit varying interaction frequencies but relatively low success rates. GPT-4o achieves better performance in interaction tasks by maintaining a comparable success rate with a higher interaction frequency. Existing MLLMs need improvements in spatial distance perception, grounding, and affordance judgment to achieve a higher interaction success rate.

### 4.4 Error Case Analysis

We summarize four primary error categories in MLLM-based embodied agents: **(1) Hallucination in Grounding:** Models misperceive the environment, hallucinating nonexistent objects or overlooking present ones. For example, models may confidently describe absent items or fail to locate small objects like laptops or keys, impacting both QA (e.g., providing answers based on imagined objects) and non-QA tasks (e.g., failing to navigate to or interact with target objects). **(2) Insufficient Exploration:** Agents employ suboptimal exploration strategies, hindering information gathering and goal finding due to incomplete environment coverage. They are often trapped in local areas, or answer before fully exploring the environment due to overconfidence. **(3) Lack of Spatial Reasoning:** Models struggle with understanding spatial relationships. They misinterpret directional instructions (e.g.,"to my left") and face difficulties navigating between locations, even for simple tasks such as moving to or around furniture. **(4) Wrong Planning:** Agents demonstrate poor state estimation and action planning. This results in random or repetitive actions, such as aimless circling or repeatedly picking up objects. They also struggle to understand the outcomes of the action and adapt after failed attempts. Figure 6 provides illustrative examples of these errors. For more detailed examples, please refer to Appendix F and Appendix G.

### 4.5 Future Improvements

Based on the results and error analysis, there are some potential improvements for the development of MLLMs. MLLMs are primarily trained using internet data, lacking training in physical space, which is a significant difference from humans. This leads to poor spatial-related abilities, which could potentially be improved through embodied trajectory data, egocentric video data, synthetic data, and other sources. Egocentric perception and grounding in sequential images or videos should be further explored to reduce hallucination phenomena. Since current models struggle with long-horizon tasks (even those that are just a dozen or so steps), considerable effort is needed to enable them to better understand long multimodal sequences, which is crucial for solving long-horizon visual and embodied tasks. In addition, MLLMs can also be combined with training methods like reinforcement learning to further enhance their ability to explore, reason, and recover from mistakes, building upon their foundational capabilities.


Figure 6: Case study of common error categories.

## 5 Conclusion

In this paper, we propose EMBODIEDEVAL, the first interactive benchmark designed for MLLMs with comprehensive embodied tasks. We provide an efficient framework to interactively evaluate the capabilities of MLLMs on embodied tasks. To ensure the accuracy, diversity, and quality of the dataset, extensive efforts are devoted to the annotation process for each task sample.

Through experiments, we find that current MLLMs perform poorly on embodied tasks. However, we believe there will be more attention to improving the embodied capabilities of MLLMs upon the general capabilities learned from universal multimodal data. We hope EMBODIEDEVAL can help and guide the development of MLLMs to realize their potential in embodied intelligence.

## 6 Limitations

To ensure the quality of the evaluation set, the verification process is time-consuming and involves checking each scene, task, and correctness individually. As a result, our evaluation set contains 327 test cases. We will incorporate more cases in future research.

## 7 Potential Risks

While we aim to advance the capabilities of MLLMs as interactive embodied agents, there are inherent risks that must be acknowledged. One potential risk is the over-reliance on MLLMs for decision-making in critical scenarios, which could lead to biased outcomes due to the models' limitations in understanding contextual nuances. Additionally, the deployment of such advanced systems in real-world environments raises concerns about privacy and data security, as these models often require substantial amounts of personal and environmental data to function effectively.

## References


Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. 2018a. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. 2020a. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020b. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2023b. Egoplan-bench: Benchmarking multi-modal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2023. Can vision-language models think from a first-person perspective? *arXiv preprint arXiv:2311.15596*.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024a. 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.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Zhili Cheng, Zhitong Wang, Jinyi Hu, Shengding Hu, An Liu, Yuge Tu, Pengkai Li, Lei Shi, Zhiyuan Liu, and Maosong Sun. 2024c. Legent: Open platform for embodied agents. *arXiv preprint arXiv:2404.18243*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2022. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994.

Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadhan, and Dinesh Manocha. 2024. S-eqa: Tackling situational queries in embodied question answering. *arXiv preprint arXiv:2405.04732*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Chenyou Fan. 2019. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

JJ Gibson. 1977. The theory of affordances. *Perceiving, acting and knowing: Towards an ecological psychology/Erlbaum*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Md Mofijul Islam, Alexi Gladstone, Riashat Islam, and Tariq Iqbal. 2024. EQA-MX: Embodied question answering using multimodal expression. In *The Twelfth International Conference on Learning Representations*.

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*.

Zixia Jia, Mengmeng Wang, Baichen Tong, Song-Chun Zhu, and Zilong Zheng. 2024. LangSuit·E: Planning, controlling and interacting with large language models in embodied text environments. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14778–14814, Bangkok, Thailand. Association for Computational Linguistics.

Yash Kant, Arun Ramachandran, Sriram Yenamandra, Igor Gilitschenski, Dhruv Batra, Andrew Szot, and Harsh Agrawal. 2022. Housekeep: Tidying virtual households using commonsense reasoning. In *European Conference on Computer Vision*, pages 355–373. Springer.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. 2024a. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393.

Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. 2024b. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383.

Byeonghwi Kim, Jinyeon Kim, Yuyeong Kim, Cheolhong Min, and Jonghyun Choi. 2023. Context-aware planning and environment-aware memory for instruction following embodied agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10936–10946.

Eric Kolve and et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. 2023b. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, et al. 2024. Embodied agent interface: Benchmarking llms for embodied decision making. *arXiv preprint arXiv:2410.07166*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024b. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023a. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023c. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. 2024c. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2024d. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Ji Ma, Hongming Dai, Yao Mu, Pengying Wu, Hao Wang, Xiaowei Chi, Yang Fei, Shanghang Zhang, and Chang Liu. 2024. Doze: A dataset for open-vocabulary zero-shot object navigation in dynamic environments. *arXiv preprint arXiv:2402.19007*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3d environments with visual goal prediction. *arXiv preprint arXiv:1809.00786*.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2024. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2023. Gpt-4v(ision) system card.

OpenAI. 2024. Hello gpt4-o.

OpenGVLab. 2024. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020a. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020b. Reverie: Remote embodied visual referring expression in real indoor environments. *Preprint*, arXiv:1904.10151.

Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. 2024. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*.

Manolis Savva and et al. 2019. Habitat challenge: A photorealistic embodied ai benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9338–9346. IEEE.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. 2022. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, pages 477–490. PMLR.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. 2023. Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*.

Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. 2023. Knowledge-based embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11948–11960.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2021. Visual room rearrangement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5922–5931.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024a. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. 2024b. Holodeck: Language guided generation of 3d embodied ai environments. *Preprint*, arXiv:2312.09067.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. 2023. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. 2019. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: A strong zero-shot video understanding model.

Zhuofan Zhang, Ziyu Zhu, Pengxiang Li, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. 2024b. Task-oriented sequential grounding in 3d scenes. *arXiv preprint arXiv:2408.04034*.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. 2021. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12689–12699.

# A Task Samples

In EmbodiedEval, each category of tasks has sufficient diversity to comprehensively evaluate the model. For example, different from traditional EQA task, attribute question answering tasks in EMBODIEDEVAL encompass a more diverse range of attribute questions about objects and scenes, including but not limited to category, shape, material, color, function, state, location, existence, quantity, comparative analysis, and complex reasoning across multiple attributes and multiple objects. For interaction tasks, the tasks possibly involve multiple objects and multi-step interactions, such as using a tool to manipulate another object or rearranging items to meet specific conditions, which necessitates fine-grained movement planning, reasoning about object affordances, and understanding cause-and-effect relationships. We selected some representative examples to illustrate the diversity of the task set in Table 4.

# B Details of Evaluation Framework

## B.1 Movement Space

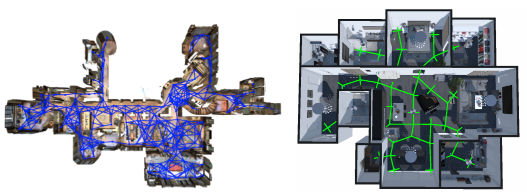


Figure 7: A comparison of navigation graphs between R2R (Anderson et al., 2018b) dataset (left) and EMBODIEDEVAL (right).

We use navigation graph as the movement space where the agent can rotate its view at a point or move between adjacent navigation points. Compared to continuous movement, it discretizes motion without imposing great restrictions on the high-level tasks in practice (Anderson et al., 2018b). Different from grid-world movement, this approach is more natural and adaptable to all kinds of scenes. Through sampling algorithms and manual adjustment, we constructed navigation graphs for each scene. To ensure realism, the navigation points are always walkable locations with no obstacles among them. Due to the greater diversity of our scenes and tasks compared to previous work, the density of navigation points varies based on the size of the scene and the task, ensuring that the number of steps required for tasks remains reasonable. For example, in complex interaction tasks within large scenes, the navigation points are more sparse and critical. In contrast to previous datasets, our navigation points are better organized as shown in Figure 7, and the connections between these points indicate clear semantics. MLLMs are not required to make choices from a set of 3D positions, but only need to make directional decisions among navigation points. Specifically, the action space consists of three types of actions: ***move forward*** (moving to the facing navigation point), ***turn left/right*** (rotating to face a new navigation point), and ***look up/down*** (adjusting the vertical view).

## B.2 Interaction Space

We follows the discrete interaction space of previous embodied AI tasks that involves object interaction such as IQA (Gordon et al., 2018), CHAI (Misra et al., 2018), RoomR (Weihs et al., 2021) and OVMM(Yenamandra et al., 2023) rather than continous space (Shridhar et al., 2020; Srivastava et al., 2022). This choice is based on two main considerations: (1) In continuous spaces, interactions are tightly related to specific methods and types of embodiment, which contradicts the goal of generality in evaluations and goes beyond the core issues of our research. (2) Due to the high complexity of continuous space, MLLMs cannot output reasonable values without being trained on specialized numerical trajectory, leading to infeasible evaluations. In EMBODIEDEVAL, we use an open vocabulary for the actions and objects in interactions to make them as rich as possible. Each interaction action has a brief action text, operable objects, and conditions for successful interaction. For example, the “pick up" action requires the target object to be within sight and very close, the “wash" action requires the agent and the target object to be next to a sink, and the “hand over" action requires the agent to hold an object and be next to a person. In a given test case, several interaction actions will be involved, including those necessary to complete the task and other distracting actions. We provided more examples of the interaction space mentioned in Table 7.

## B.3 Answering Space

Our answering space consists of eight annotated options by annotators. All the options are carefully written and verified to ensure that the answers are challenging, meaningful within the scene, and have a strong distractive capability. We demonstrate some examples in Figure 5.

### B.4 Success Criteria

We automatically and accurately evaluate task completion through predicate functions. Each predicate maps the state of the simulation environment to a boolean value indicating success. For example, the ***agent_at*** predicate requires a designated navigation point as a parameter and returns true when the agent reaches this location at the end of the episode. Beyond evaluating only the final state, EMBODIEDEVAL also includes predicates that assess the entire process, similar to R4R (Jain et al., 2019). For example, the ***agent_pass*** predicate becomes true once the agent passes a specified navigation point. All the predicate are listed in Table 6.

A task is considered successful when all predicates evaluate to be true at the end. Consider the task "*Please go to the kitchen, then come back and tell me if there are any extra cups*". This task involves three predicates: ***agent_pass***, ***agent_at***, and ***choose***. These predicates verify that the agent passes through the kitchen doorway, returns to the initial position in front of the person, and selects the correct answer, respectively.

## C Details of Task Annotation

### C.1 Annotation Process

A task sample includes a scene, task description, output space, and success criteria. The annotators are required to conduct the annotation as following process: (1) Select a task to annotate from the candidate tasks and choose a suitable candidate scene. Nouns, prepositions, adjectives, and other elements in the task text can be slightly adjusted to fit appropriately within the context of the scene, while keeping the core content of the task the same. Each candidate task can only be selected and used once. (2) Annotate the movement space by adjusting the navigation points in the scene. (3) For tasks involving interaction, annotate the interaction space by setting action options, including the action's text, type, and parameters. The interaction space includes the necessary action options for the task, as well as some distracting action options. For certain specific interactions, it is necessary to annotate feedback content. For example, interactions that involve asking humans require annotating the content of human responses. (4) For QA tasks, annotate the answering space by writing challenging answer options. (5) Annotate the success criteria by setting predicate functions, including the predicate's type and parameters. If the task has multiple sub-goals, the predicate function of each sub-goal should also be included. (6) Annotate the agent's initial position and orientation. (7) For social interaction tasks, annotate the initial position, orientation, and body posture of the person, including standing, sitting, lying down, and the direction of the finger pointing, and choose the person's appearance from a selection of characters from Mixamo[2]. (8) Run the annotated tasks in the simulator to confirm that the tasks can be completed without any issues.

### C.2 Annotation Criteria

(1) All tasks must be unambiguous within the given scene. (2) Question-answering tasks must require scene observation, with each task providing eight answer options that vary in difficulty and include misleading options to reduce the chance of guessing the correct answer. (3) Once a task is correctly annotated, the tasks must be executable in the simulator with a well-designed navigation graph and accurate action options. Annotators must verify task feasibility using the same observational constraints as agents.

### C.3 Annotation System

To ensure both efficiency and precision in the complex annotation process, we develop an annotation system based on Unity[3]. The system provides comprehensive function, which encompassing scene and task import/export, flexible content viewing, visualized action space, and a guide annotation workflow that adheres to predefined guidelines:(1) Importing and exporting scenes and tasks, allowing users to freely view the content of scenes and tasks. (2) Enforcing task annotation according to predefined guidelines. The system provides candidate lists for all types of actions and predicates and specifies he parameters that need to be annotated. (3) Generating navigation points and constructing a navigation graph with visualizations, allowing for the addition, deletion, and modification of navigation points. (4) Annotators can visually select 3D objects in the scene as parameters of interactions and predicates. Once one annotation is complete, the task file exported by the annotation system can be loaded by the simulator, starting the simulation and evaluation process.

[2]https://www.mixamo.com/
[3]https://unity.com/

### C.4 Quality Control

Eight expert annotators are recruited to perform the annotations. These standard annotators are from professional data annotation companies. Before starting the annotation process, we conduct systematic training on annotation requirements and system usage. To ensure the dataset's high quality, each annotated task is independently evaluated for correctness and quality by at least three reviewers. There are two rounds for the annotation. In the first round, annotators primarily ensure task completeness. In the second round, expert annotators verify and refine the diversity of task objects, the accuracy and clarity of task descriptions, and task difficulty distinctions. Furthermore, we validate task feasibility by providing expert demonstrations for each task and testing human performance with non-expert participants.

## D Creation of Objaverse Synthetic

We use a wide variety of objects from Objaverse to procedurally generate diverse scenes and further refine them through interactive scene editing.

**Object Selection.** We curated a subset of indoor assets out of Holodeck's (Yang et al., 2024b) annotated realistic and diverse objects choosen from the Objaverse asset library (Deitke et al., 2023). To ensure quality, we employed GPT-3.5 to filter unsuitable outdoor objects and manually reviewed frontal renderings to remove low-quality assets. This process resulted in a database of about 15,000 objects spanning over 500 categories (examples seen in Fig 9).

**Scene Generation.** We leveraged GPT-3.5 to annotate object categories with their typical room occurrences (e.g., inLivingRoom, inKitchen), positions (e.g., onWall, onFloor, onEdge), and functions (e.g., receptacle, pickup). Gemini-1.5-Flash was used to annotate large objects' orientations. Subsequently, a procedural approach was employed to randomly place architectural elements such as walls, doors, and windows. Large objects were then arranged on the floor either against the walls or in the center of the rooms, and smaller items were finally placed on surfaces of large receptacles. Hundreds of scenes were generated randomly, from which we selected 15 living rooms, 15 bedrooms, 10 two-room, 5 three-room, and 5 four-room for further editing, as partly shown in Fig 10.

**Scene Editing.** To make the scene more organized and to avoid errors caused by automatic generation, we also edited the generated scene by developing a runtime scene editor. Users can view the type and description of objects, and adjust their position and orientation (see Fig 8). Once editing is complete, the scene can be saved as a JSON file and imported to reproduce the environment.

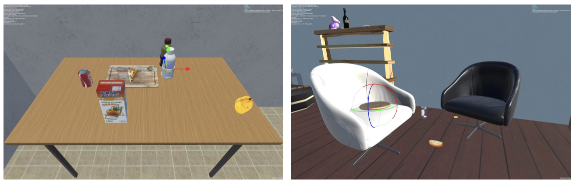


Figure 8: Interactive scene editor: adjust object position (left) and angle (right).

## E Temperature Setting

We find that all models perform slightly better at temperature = 1 compared to temperature = 0. Through observing cases, we believe this is because embodied tasks require a certain level of exploration, and when the temperature is set to 0, the determinism of the output causes the model to more easily get stuck in repetitive errors. However, in this paper, we propose using a temperature of 0 as the evaluation standard, as this removes randomness from the evaluation, improving efficiency and better reflect the model's true capabilities, including its ability to recognize and escape from erroneous trajectories.

| Task | Characteristics |
| --- | --- |
| Please go to the kitchen, then come back and tell me if there are any extra cups. | scene memory |
| Imagine the house is rotated 90 degrees counterclockwise. How would this affect the natural light distribution in the room? | spatial imagination |
| Open a black locked drawer with a key found on the desk. | tool use |
| Pick up the kettle and the box labeled "BREAD" from the kitchen counter and place them on the table with the coffee machine. | optical character recognition |
| Optimize the display of artworks on the shelves as follows: place two items on each shelf, with one shelf featuring two items of the same shape. Complete the requirements in as few steps as possible. | reasoning and planning |
| Grab the object that is cylindrical and silver on the table next to the washing machine. | multiple attribute reference |
| Estimate the percentage of floor space occupied by furniture in the room you're currently in. | area estimation |
| Estimate the straight-line distance from the front door to the TV. Note that each step you take forward is approximately two meters. | distance estimation |
| Which is closer to the drink on the round table, the ginger or the ice cream? | distance comparison |
| Identify an object that is taller than 1 meter. | height estimation |
| If we were to host a birthday party, which area of the house could accommodate the most people while ensuring clear pathways to exits? | logic, space, and common sense |
| Describe the path from the kitchen to the living room. | path description |
| If you were to draw a straight line from the desk with a turned-on laptop to the bookshelf, which pieces of furniture would it intersect? | intersection estimation |
| What is the object I am pointing at? | pointing comprehension |
| Pick up the watermelon on my right. | perspective-taking comprehension |
| My red glasses are missing. Please help me look for them in the room. Once you find them, bring them to me. | object searching and delivering |
| Wake up my dad. He is sleeping in the bedroom. The bedroom is the second room on your right as you walk forward. | social navigation |
| Enter the dining area and see if there is more than one door in the entire house. | object counting |
| Calculate the ratio of seating options to the number of rooms in the house. | counting and calculation |
| Tell me which objects have a handle in the kitchen. | attribute grounding |
| Evaluate whether the painting above the living room sofa is more colorful than the carpet. | attribute comparison |
| How many rooms are there in total? | room counting |
| Confirm if a garbage can is located on the floor in the living room. | object existence |
| Which room has more seating options, the kitchen or the living room? | quantity comparison |
| I'm hungry. Find all objects that can be used as ingredients on the table in this room. | object functionality |
| Count the maximum number of identical clocks among all the rooms. | counting and attribute memory |
| What do you think the owner of this room probably studies? | common sense |
| Is there an egg inside the fridge? | interaction and answering |
| Open the drawer of the side table in the study room. If there is something inside, leave it open and put all similar items from the room into it. If there is nothing inside, close it. | logical execution |
| I just heard something fall to the TV table. What was it? Go check. | object identification |
| Explore the other side of the courtyard thoroughly in a few steps. | scene exploration |
| Imagine the house is rotated 90 degrees counterclockwise. How would this affect the natural light distribution in the room? | spatial imagination and reasoning |
| Navigate to the sofa. | object navigation |
| From the parked car in the garage, walk towards the courtyard, follow the stone path between the two blooming trees, and turn left at the end of the stone path, then walk to the front door of the house. | step-by-step fine-grained navigation |
| You are in the upper right corner of the classroom. Suppose the nearest desk to you is in the first row and first column. Go to your seat in the third row and fourth column and stand at the upper right corner of your desk. | precise navigation |
| Head to the fridge, open the fridge, take out an egg, wash it and crack it into a frying pan to fry it. | sequential interaction |
| Determine the optimal placement of the living room TV to achieve the best viewing experience from multiple seating positions. | multi-object spatial reasoning |
| Move the fruit plate from the kitchen table to the dining table with dishes. Make sure to take it from the side of the kitchen table without chairs, and when placing it, put it in the corner of the dining table closest to your starting position. | fine-grained object interaction |

Table 4: Examples of the diverse tasks in EMBODIEDEVAL.

| Task | Answer Options |
|---|---|
| If we were to host a birthday party, which area of the house could accommodate the most people while ensuring clear pathways to exits? | 1. the master bedroom<br>2. the hallway<br>3. the study room<br>4. the open balcony with some green plants<br>5. the living room<br>6. the large guest room<br>7. the garden next to the living room<br>8. the backyard with a sunshade umbrella |
| Which is closer to the drink on the round table, the ginger or the ice cream? | 1. Both are equally close to the drink.<br>2. The ice cream is closer to the drink.<br>3. The ginger is closer to the drink.<br>4. Neither is close to the drink.<br>5. The ginger is on the other side of the table.<br>6. The drink does not exist.<br>7. The ginger does not exist.<br>8. The ice cream does not exist. |
| Imagine you're a cat on the empty bookshelf. What would be the most efficient path to reach the balcony while minimizing contact with the floor? | 1. Jump onto the sofa, then onto the sofa table, onto the armchair, and finally onto the windowsill to enter the balcony.<br>2. Jump onto the kitchen counter, then onto the dining table, then onto the sofa table, then onto the sofa. Jump from the edge of the sofa armrest to the ground and finally enter through the balcony door.<br>3. Jump onto the sofa. Walk on the top of the sofa back to the end. Jump onto the floor lamp and then onto the windowsill to enter the balcony.<br>4. Jump onto kitchen counter, then onto the dining table, then onto the television, then onto the armchair, and finally onto the windowsill to enter the balcony.<br>5. Turn around and run to the left room at the end. Climb up the toilet and jump into the bathtub.<br>6. Pass through the small path between the sofa and the sofa table. Jump onto the armchair and then onto the windowsill to enter the balcony.<br>7. Pass through the small path between the television and the sofa table. Jump onto the armchair and then onto the windowsill to enter the balcony.<br>8. Jump onto the dining table. Jump from the dining chair to the carpet. Climb onto the sofa. Jump from the edge of the sofa armrest to the windowsill and enter the balcony. |
| Can you describe the type of the paintings in this house? | 1. Pen and Ink Drawing<br>2. Oil Painting<br>3. Charcoal Drawing<br>4. Digital Painting<br>5. Mosaic Art<br>6. Pencil Drawing<br>7. Ink Painting<br>8. Silk Painting |
| Share some information about the numerous red furniture items in the open kitchen. | 1. Only sofa, high stools and pendant lamps are red.<br>2. Only the refrigerator, high stools and pendant lamps are red.<br>3. There are no red objects in the paintings on the wall.<br>4. One piece of red furniture is used to store food.<br>5. All pendant lamps are red.<br>6. Only sitting furniture is red.<br>7. There is no red furniture.<br>8. All the furniture in the room is red. |
| Enter the dining area and see if there is more than one door in the entire house. | 1. Yes, there are two doors.<br>2. No, there is only one door.<br>3. The room is painted blue.<br>4. Cannot determine, not enough information.<br>5. No, there are no doors.<br>6. The dining area has a sliding door.<br>7. Yes, there are multiple doors.<br>8. There is no dining area. |
| Compare the colors of the carpet, the bedside table and the linen basket in the bedroom, and find the one that is most similar in color to the bed. | 1. The basket and the carpet are vibrant. The bedside table's color is muted and most similar to the bed.<br>2. The basket and the carpet are muted. The bedside table's color is vibrant and most similar to the bed.<br>3. The basket and the bedside table are vibrant. The carpet's color is muted and most similar to the bed.<br>4. The basket and the bedside table are muted. The carpet's color is vibrant and most similar to the bed.<br>5. The bedside table and the carpet are vibrant. The basket's color is muted and most similar to the bed.<br>6. The bedside table and the carpet are muted. The basket's color is vibrant and most similar to the bed.<br>7. They are all muted and very similar to the bed.<br>8. They are all vibrant and very similar to the bed. |

Table 5: Examples of the diverse answer options in EMBODIEDEVAL.

| Predicate | Paramters | Success Conditions |
|---|---|---|
| *choose* | The right answer. | When the agent selects the correct answer. |
| *agent_at* | A navigation point. | When the agent finally arrives at this point. |
| *agent_pass* | A navigation point. | When the agent has passed through this point at least once. |
| *at* | An object and a specific point. | When the object is at this point. |
| *grab_once* | An object. | When the agent has picked up this object at least once. |
| *grab* | An object. | When the agent picks up the object. |
| *special_action_success* | An interaction action. | When this interaction action has been successful. |

Table 6: The predicates involved in EMBODIEDEVAL.

| Action Text | Execution Requirements |
|---|---|
| wash | When the agent is holding the target object and stand next to the sink. |
| hand over | When the agent is holding the target object and stand next to the person. |
| sit down | When the agent is next to the target chair. |
| unlock | When the agent is holding the target key and standing next to the drawer |
| greet | When the agent is near the person. |
| ask | When the agent is near the person. |
| mix | When several target beverages are on the table next to the agent. |
| wipe off the table | When the agent is holding an object for cleaning and standing next to the table. |
| check the results of the program | When the agent is next to the computer. |

Table 7: Some cases of the interaction actions involved in EMBODIEDEVAL.


Figure 9: Examples of selected Objaverse assets and views of generated scenes.

Figure 10: Front View of Scenes in Objaverse Synthetic. Top: Multiple rooms. Bottom: Single room.

**Prompt for Multi-image MLLMs**

```
You are an intelligent vision-language embodied agent skilled at solving tasks
     and answering questions in a 3D environment. Your job is to efficiently
    complete a specified task by choosing the optimal action at each timestep
    from a set of available actions. You are given a series of ego-centric
    images, and a history of previous actions with optional feedback (success/
    failure or human response).  Each image shows what you see at a particular
     step in the action history, along with an extra image showing your
    current view.

Current task: {}
Action history (action -> feedback): {}
Visual history: {}
Current view: {}
For the current step, your available options are listed as "[Option␣Number].␣
    Content" as follows: {}
Choose your action from the above options by replying with "Thought:␣Your␣
    reasoning.\nChoice:␣[Option␣Number]␣(e.g.␣[1])".

Note:
- If the task needs more information of the scene, navigate wisely to the
    required targets (objects, places, or people).
- Avoid repeated actions like useless forward motion and circling.
- You can only interact with objects or humans (e.g. pick/place/open/close/
    handover) if they are within your view and very close to you.
- You can only hold one object at a time. Put down any held object before
    picking up another.
- Tasks containing "I" or "me" are requested by a person in the scene.
- Reflect on why previous actions fail to avoid repeating mistakes and ajdust
    your current action.
- You have a limited number of {} steps to complete the task.
```

Figure 11: Prompt for Multi-image MLLMs.

# F Success Cases

We present successful cases accomplished by closed-source MLLMs to gain deeper insights into their current capabilities. As discussed in Section 4.2, the models generally scored low and successfully completed only a limited number of tasks. A closer examination of these successful tasks reveals that they are typically simpler, involve fewer steps and require interaction with fewer objects. To better illustrate these findings, we present representative cases from five task types. These examples highlight the underlying behavioral patterns and reasoning processes of the models during task execution.

## F.1 Attribute QA

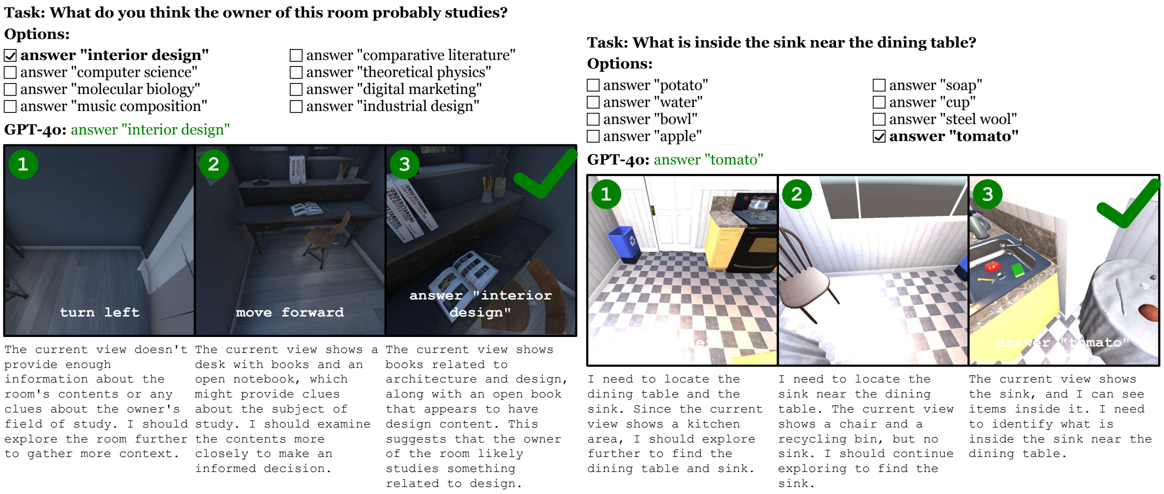



Figure 12

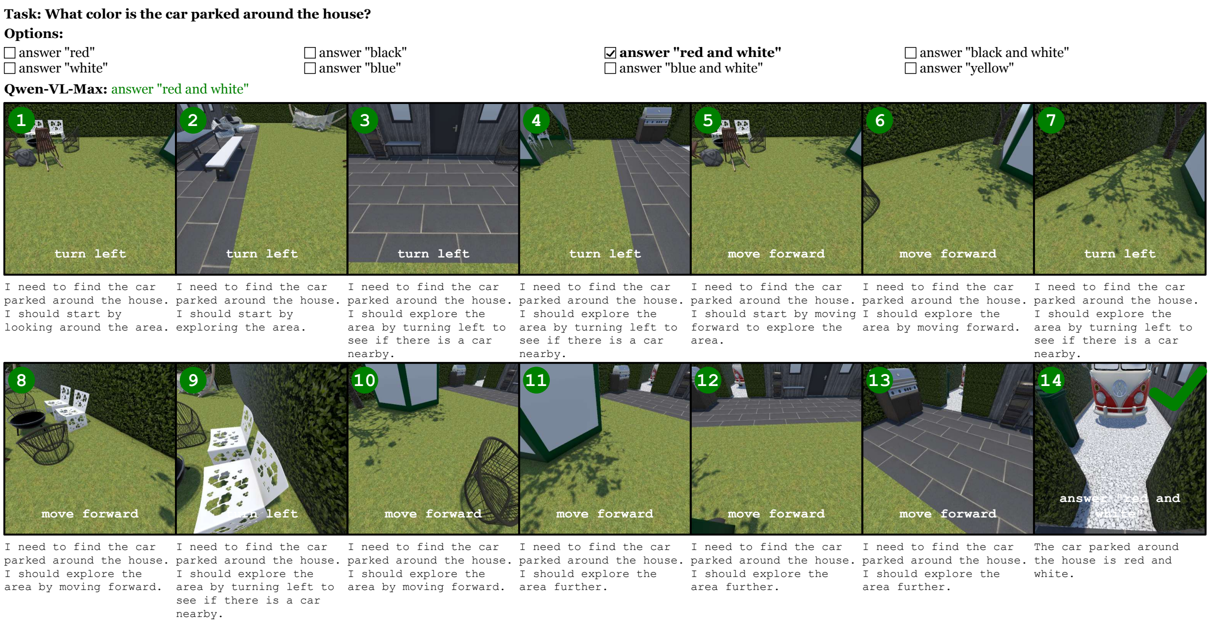



Figure 13

**Task: Compare the sizes of two side tables in the bedroom.**
**Options:**

- ☐ answer "The side table with the lamp is located near the window."
- ☐ answer "Both side tables are of equal size."
- ☐ answer "The side table with the alarm clock is larger than the one with the lamp."
- ☐ answer "The side table with the alarm clock is painted blue."
- ☐ answer "It's difficult to tell the difference in size between the two side tables."
- ☑ **answer "The side table with the lamp is larger than the one with the alarm clock."**
- ☐ answer "The side table with the alarm clock has a drawer."
- ☐ answer "The side table with the lamp has two drawers."

**Gemini-Flash:** answer "The side table with the lamp is larger than the one with the alarm clock."


1: I need to get a better view of both side tables to compare their sizes. I can't tell from the current image.

2: I need to get a better view of both side tables to compare their sizes. I can't tell from this angle.

3: I need to get a better view of both side tables to compare their sizes. I've already moved forward twice, so I should try turning to get a better perspective.

4: I need to get a better view of both side tables to compare their sizes. I've already turned left, so I should turn right to get a different perspective.

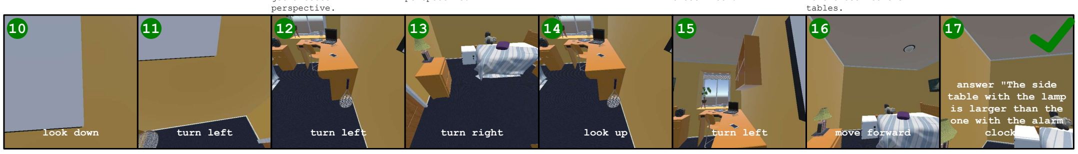
5: I need to get a better view of both side tables to compare their sizes. I've already turned left and right, so I need to move closer to them.

6: I need to get a better view of both side tables to compare their sizes. I've already turned left and right, so I should move forward to get a closer look.

7: I need to get a better view of both side tables to compare their sizes. I've already moved forward and turned around, so I need to move closer to the tables.

8: I need to get a better view of both side tables to compare their sizes. I've been moving around, but I haven't gotten a good angle yet.

9: I need to get a better view of both side tables to compare their sizes. I've been moving around, but I haven't found a good angle yet.




10: I need to get a better view of both side tables to compare their sizes. I've already moved around the room, but I haven't looked down to see the full extent of the tables.

11: I need to get a better view of both side tables to compare their sizes. I've already moved around the room and looked up and down, but I haven't gotten a good angle yet.

12: I need to get a better view of both side tables to compare their sizes. I've already turned left and right, so I need to move closer to them.

13: I need to get a better view of both side tables to compare their sizes. I've already turned left twice, so I should try turning right to see if I can get a better angle.

14: I need to get a better view of both side tables to compare their sizes. I've been turning around a lot, but I haven't gotten a good angle yet. I'll try looking up to see if that helps.

15: I need to get a better view of both side tables to compare their sizes. I've already looked up and down, so I need to move closer to them.

16: I have been moving around the bedroom, trying to get a good view of both side tables. I've seen both tables, but I haven't been able to get a clear view of their sizes. I need to get closer to one of the tables to compare them.

17: I have moved around the bedroom and have a good view of both side tables. I can see that the side table with the lamp is larger than the one with the alarm clock.

Figure 14

## F.2 Spatial QA

**Task: Determine if the sofas on the left will fit through the front door.**
**Options:**
☐ answer "There is no front door."
☐ answer "Yes, the front door is wide enough, and all the sofas can pass through easily in any orientation.."
☐ answer "No, only the smallest sofa can pass through."
☐ answer "Yes, but the sofas need to be stood upright to pass through."
☐ answer "No, the largest sofa cannot pass through."
☐ answer "No, none of the sofas can pass through."
☐ answer "There is no sofa."
☑ **answer "Yes, but the sofas need to be turned sideways to pass through."**
**GPT-4o:** answer "Yes, but the sofas need to be turned sideways to pass through."


I need to gather more information about the sofas to determine if they can fit through the front door. Since the current view is of the front door, I should turn left to check the sofas.

I need to determine if the sofas can fit through the front door. I should compare the size of the sofas with the door. The current view shows the sofas, but I need to see the door again to make a comparison.

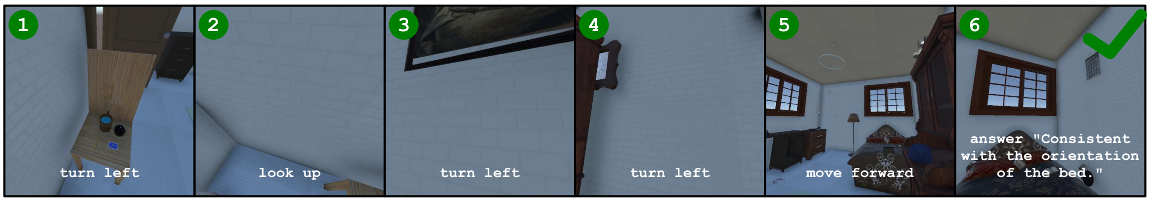
I need to determine if the sofas can fit through the front door. I should first look at the door to assess its size.

I need to assess the size of the sofas and the front door to determine if they can fit through. The current view is focused on the door, but I need to see the sofas again to make a comparison.

I need to determine if the sofas can fit through the front door. I should compare the size of the sofas to the door. I have already seen the door and the sofas, so I can make a decision based on the current view.

Figure 15

**Task: What direction does the bedroom doorway face when looking out?**
**Options:**
☐ answer "Opposite to the orientation of the table."
☐ answer "Opposite to the orientation of the chair with many objects on it."
☐ answer "Consistent with the orientation as the table."
☐ answer "Consistent with the orientation as the toilet."
☐ answer "Consistent with the orientation as the white cabinet."
☑ **answer "Consistent with the orientation of the bed."**
☐ answer "Consistent with the orientation as the fridge."
☐ answer "Opposite to the orientation of the bed."
**GPT-4o:** answer "Consistent with the orientation of the bed."




The current view shows a table with objects on it, and a door is visible. To determine the direction of the bedroom doorway, I need to explore more of the environment to understand the layout better.

I need to gather more information about the room's layout to determine the direction of the bedroom doorway. Since the current view is limited, I should explore further.

I need to determine the orientation of the bedroom doorway. The current view shows a wall with a painting, which doesn't provide enough information about the doorway. I should explore more by turning to get a better view of the room layout.

I need to determine the orientation of the bedroom doorway. The current view doesn't show the doorway, so I should explore further to find it.

I need to determine the orientation of the bedroom doorway. The current view shows the bed and windows, which might help in understanding the room's layout. Since the task is about the doorway's direction, I should explore more to find the doorway.

I need to determine the orientation of the bedroom doorway. Based on the current view and previous images, I should consider the orientation of the bed, as it is a significant feature in the room.

Figure 16

## F.3 Navigation

**Task: Walk straight past the kitchen on your left, turn right when you are almost at the bathroom door, and stop when you reach the book in the center of the sofa table.**
**Qwen-VL-Max:**

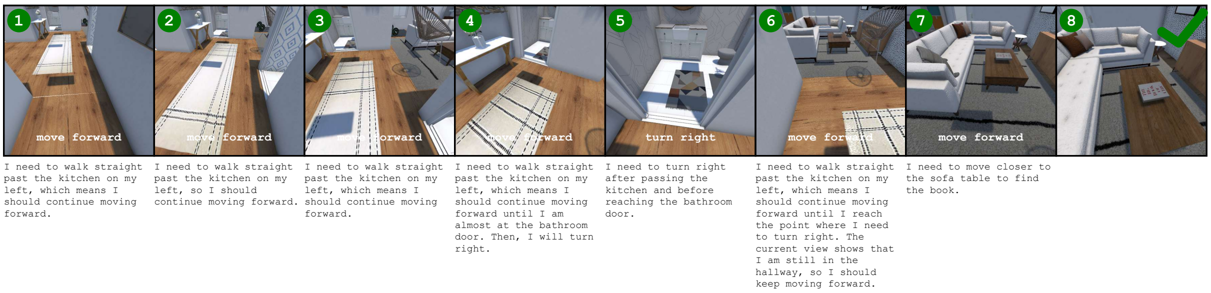



Figure 17

**Task: I want to find a human-shaped statue which is placed on the table below the wall where the iris painting hangs. Please navigate to it.**
**Qwen-VL-Max:**


Figure 18

## F.4 Object Interaction

**Task: Is there an egg inside the fridge?**

**Options:**

☐ answer "Yes, and there are two eggs, one in the upper part and one in the lower part."
☑ **answer "Yes, in the lower part of the fridge."**
☐ answer "Yes, there are two eggs in the upper part of the fridge."
☐ open the lower fridge door
☐ answer "There is no fridge."
☐ answer "No, there are no eggs in the fridge."
☐ answer "Yes, in the upper part of the fridge."
☐ open the upper fridge door
☐ answer "Yes, there are two eggs in the lower part of the fridge."

**Qwen-VL-Max:** answer "Yes, in the lower part of the fridge."

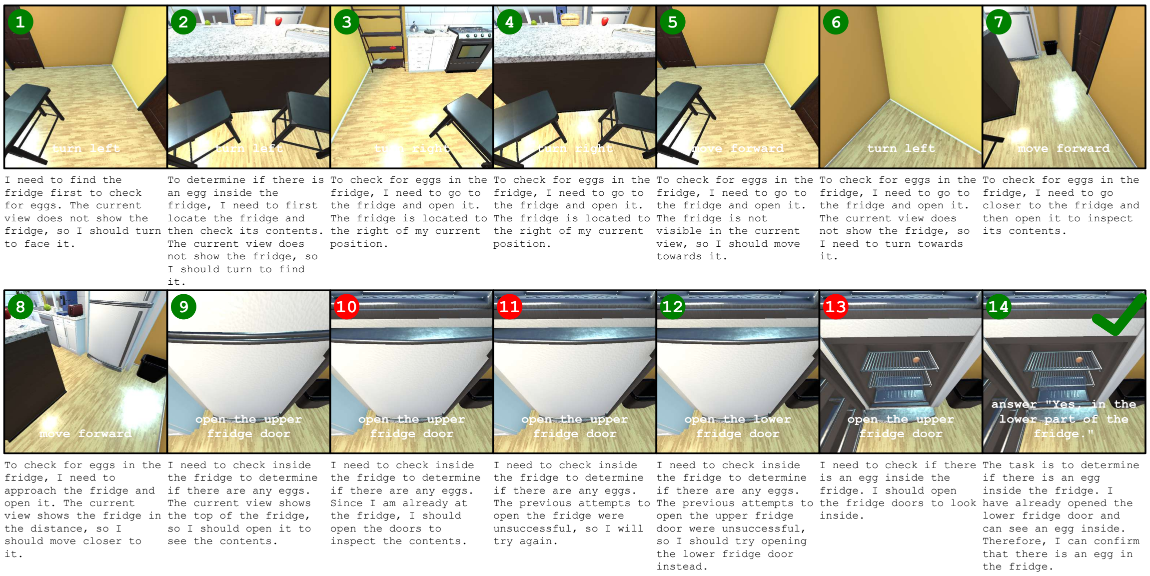



Figure 19

**Task: Pick all round objects into trash can.**
**Options:**
☐ pick up ☐ place to the trash can
**Gemini-Pro:**

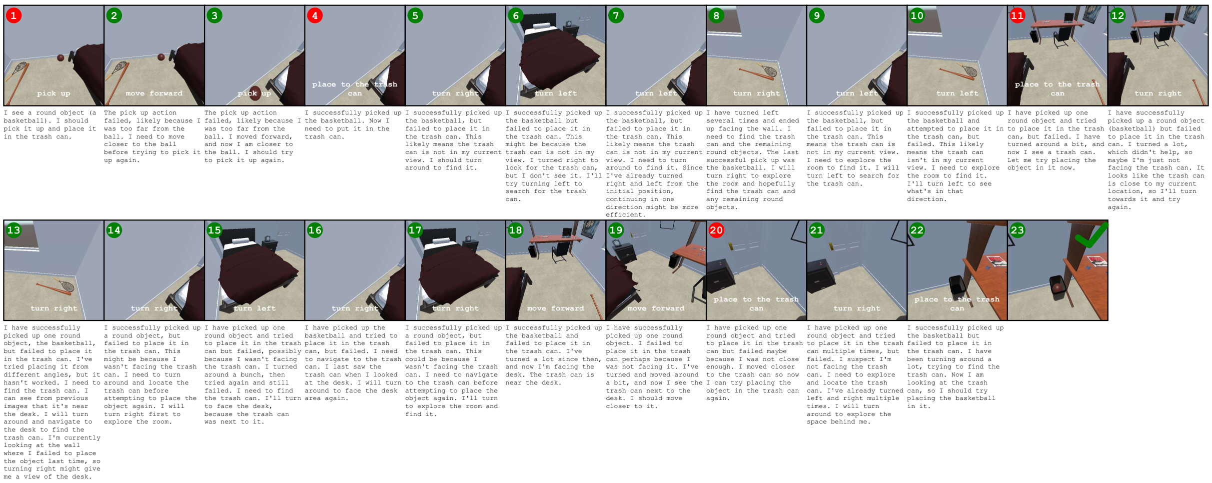



Figure 20

## F.5 Social Interaction

**Task: Wake up my dad. He is sleeping in the bedroom. The bedroom is the second room on your right as you walk forward.**
**Options:**
☐ speak "get up"
**GPT-4o-Mini:**

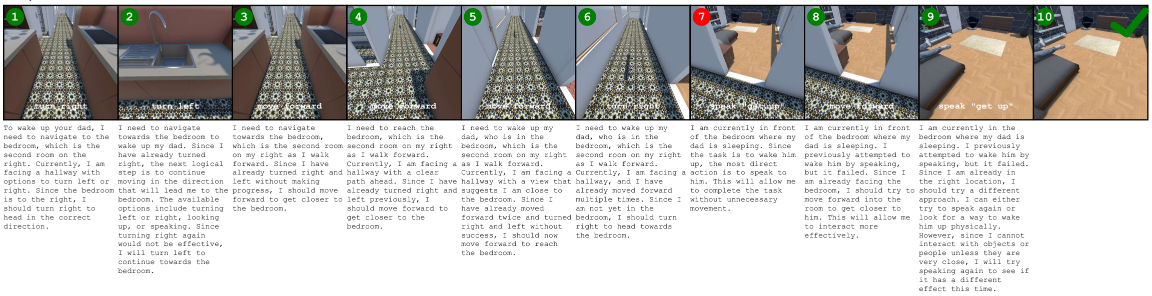



Figure 21

**Task: Retrieve the magazine for the madam sitting on the chair.**
**Options:**
☐ pick up ☐ hand over
**Gemini-Pro:**

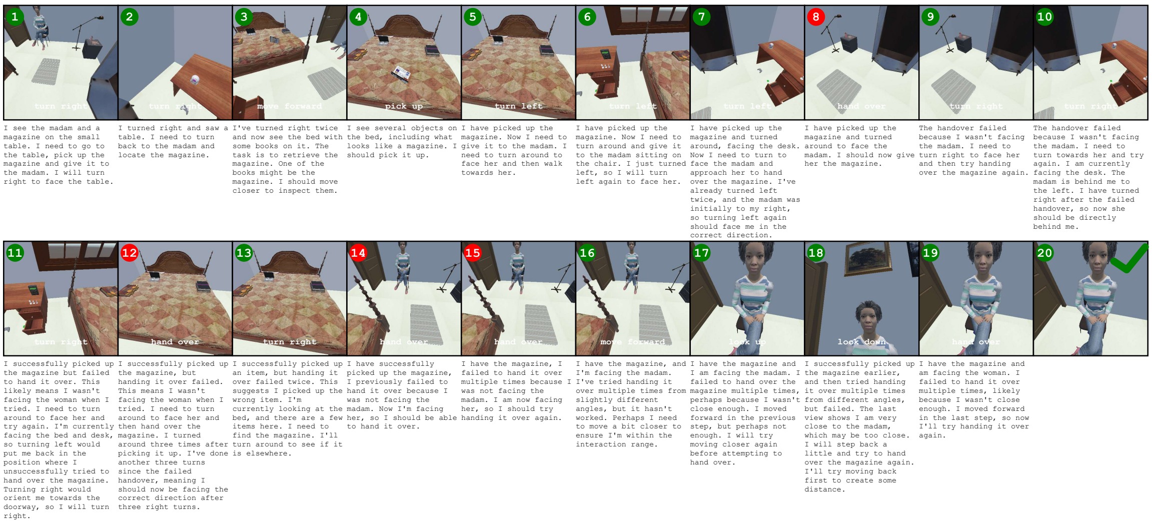



Figure 22

# G Error Cases

## G.1 Attribute QA

**Task: Locate any electronic devices in the bedroom.**

**Options:**

- ☐ answer "flashlight, desk lamp, and phone"
- ☐ answer "laptop, desk lamp, and phone"
- ☑ **answer "laptop, TV, and desk lamp"**
- ☐ answer "TV, laptop, and door"
- ☐ answer "flashlight, desk lamp, and TV"
- ☐ answer "flashlight, desk lamp, and ipad"
- ☐ answer "laptop, desk lamp, and bed"
- ☐ answer "laptop, book, and desk lamp"

**Qwen-VL-Max:** answer "laptop, desk lamp, and phone"


1: I need to explore the room to find electronic devices. I should start by turning left to see what's there.

2: I need to explore the room to find electronic devices. I should start by turning left to see more of the room.

3: I need to locate electronic devices in the bedroom. I have already seen a laptop and a desk lamp. I should explore the room further to find more devices.

4: I need to locate electronic devices in the bedroom. I have already seen a laptop and a desk lamp. I should continue exploring the room to find more devices.

5: I have already explored the room and found a laptop, desk lamp, and phone. I should report these findings to complete the task.

Figure 23

**Task: Are there more flower pots in the living room or the bedroom?**

**Options:**

- ☐ answer "There are more flower pots in the living room, while no one in bedroom"
- ☐ answer "The bedroom has a few flower pots, but the living room has none."
- ☑ **answer "The bedroom has only one flower pot, while the living room has several."**
- ☐ answer "There are no flower pots in the bedroom, only in the living room."
- ☐ answer "The living room has more flower pots, placed near the trash can."
- ☐ answer "There is only one flower pot in each room."
- ☐ answer "Both the living room and bedroom have the same number of flower pots."
- ☐ answer "Both rooms have flower pots, but the bedroom's pots are larger in size."

**Gemini-Pro:** answer "There are no flower pots in the bedroom, only in the living room."


1: I can see a flower pot in the living room. I need to explore the bedroom to compare the number of flower pots.

2: I need to explore both the living room and the bedroom to count the flower pots. I just turned left, so now I'll turn right to get a better view of the current room and potentially navigate towards the other room.

3: I have turned left and right, but still haven't seen all of the flower pots in both rooms. I need to explore more before answering.

4: I have turned left and right twice, but still haven't fully explored the rooms to find the flower pots. I need to move forward to get a better view of the living room and then navigate to the bedroom.

5: I have turned left, right, right, and forward. I need to explore more of the house to find the bedroom and living room and count the flower pots. Moving forward seems like the best option to continue exploration.

6: I have explored the environment a bit. I believe I've seen both the living room and the bedroom. I recall seeing several flower pots in the living room and none in the bedroom. I'm ready to answer.

Figure 24

## G.2 Spatial QA

**Task: Determine the best location for a fire extinguisher that would be quickly accessible from both the kitchen and the living room.**
**Options:**

- ☐ answer "Near the window."
- ☑ **answer "Beside the flowerpot near the round carpet."**
- ☐ answer "On the countertop."
- ☐ answer "Beside the black sofas."
- ☐ answer "In front of the stove."
- ☐ answer "Next to the piano."
- ☐ answer "Beside the white sofa."
- ☐ answer "Next to the irregular table with many books."

**GPT-4o:** answer "Beside the white sofa."

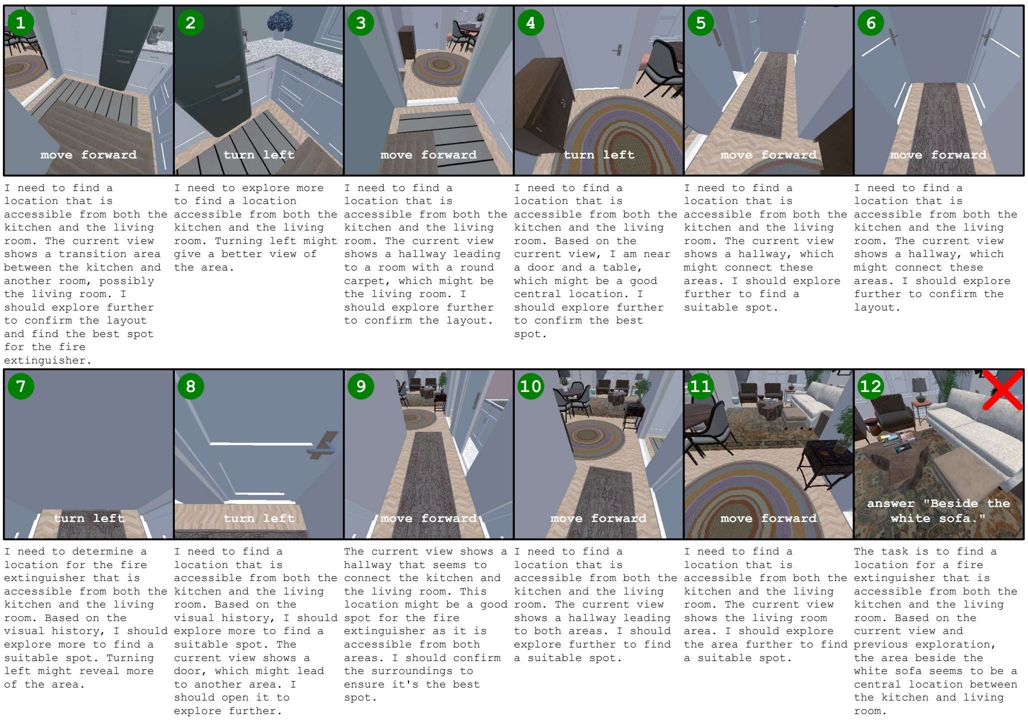



Figure 25

## G.3 Navigation

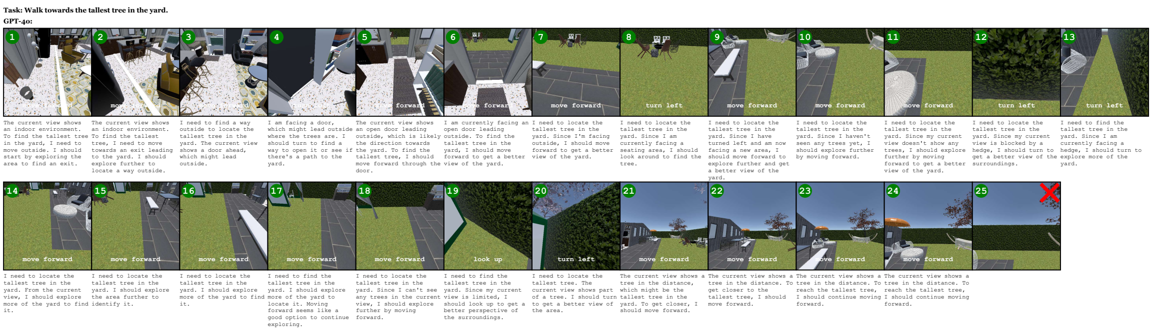
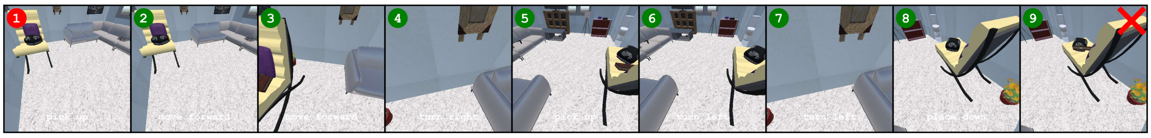



Figure 26

## G.4 Object Interaction

**Task: Retrieve the tied reading materials from the sofa in the living room and place it on the lounge chair.**
**Options:**
☐ place down ☐ pick up
**GPT-4o:**




1. The task is to retrieve the tied reading materials from the sofa and place them on the lounge chair. In the current view, I can see the sofa and the reading materials on it. The lounge chair is also visible. I need to pick up the reading materials first.
2. The task is to retrieve the tied reading materials from the sofa and place them on the lounge chair. The previous attempt to pick up the item failed, possibly because I wasn't close enough. The reading materials are visible on the sofa. I should move closer to the sofa to successfully pick them up.
3. The task is to retrieve the tied reading materials from the sofa and place them on the lounge chair. The current view shows the sofa and the lounge chair. Since the previous attempt to pick up failed, I need to ensure I'm close enough to the reading materials. I should move closer to the sofa to successfully pick up the materials.
4. The reading materials are likely on the sofa to my right, based on the previous images. I should turn right to face the sofa and locate the materials.
5. The tied reading materials are visible on the sofa in the current view. I need to pick them up before moving to the lounge chair.
6. I have successfully picked up the reading materials. Now, I need to place them on the lounge chair, which is visible in my current view. The lounge chair is to my left.
7. I have successfully picked up the reading materials. Now, I need to place them on the lounge chair. The current view shows the wall, so I need to turn to face the lounge chair.
8. I have successfully picked up the reading materials. Now, I need to place them on the lounge chair, which is currently in my view.

Figure 27

## G.5 Social Interaction

**Task: Bring my credit card to me and return it back to its original place.**
**Options:**
☐ pick up the card ☐ place down ☐ hand over ☐ place into trash can
**GPT-4o:**


1. I need to locate the credit card first. The current view shows a trash can and part of a room. I should explore the area to find the card.
2. The current view shows a trash can, but I don't see a credit card. I should turn to explore the area further.
3. I need to locate the credit card before I can pick it up. The current view shows a bed, a chair, and a desk. I should explore the area further to find the card.
4. The task is to bring the credit card to the person and return it to its original place. Currently, I don't see a credit card in my view. I should explore the room further to locate it. Turning right might give a better view of the room.
5. The credit card is visible on the desk in front of me. The next logical step is to pick it up.
6. I have already picked up the card. The next step is to hand it over to the person in the room. The person is visible in the previous images, so I should hand over the card.


7. The task is to bring the credit card to the person and then return it to its original place. The previous attempt to hand over the card failed, possibly because the person was not in the correct position. I need to locate the person and try handing over the card again.
8. I am currently facing the person who requested the credit card. The previous attempt to hand over the card failed, possibly because I wasn't close enough. I should move closer to the person to successfully hand over the card.
9. I am currently facing the person who requested the credit card. The previous attempt to hand over the card failed, possibly because I wasn't close enough. I should try handing over the card again now that I'm closer.
10. The previous attempts to hand over the card failed, possibly because I wasn't close enough to the person. I need to move closer to the person in the current view to successfully hand over the card.
11. I am currently facing the chairs, and the person is likely behind me. I need to turn around to face the person and hand over the card. Previous attempts to hand over the card failed, possibly because I wasn't facing the person directly.

Figure 28