EmbodiedEval: Evaluate Multimodal LLMs as Embodied Agents

Anonymous CVPR submission

Paper ID *****

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 AI 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 comprehensive and interactive evaluation benchmark for MLLMs with embodied tasks.* EMBODIEDEVAL *features 328 distinct tasks within 125 varied 3D scenes, each of which is rigorously selected and annotated. It covers a broad spectrum of existing embodied AI tasks with significantly enhanced diversity, all within a unified simulation and evaluation framework tailored for MLLMs. The tasks are organized into five categories: navigation, object interaction, social interaction, attribute question answering, and spatial question answering to assess different capabilities of the agents. We evaluated the state-of-the-art MLLMs on* EMBODIEDEVAL *and found 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. We open-source all evaluation data and simulation framework at https://github.com/thunlp/EmbodiedEval.*

Figure 1. Examples of the five task categories in EMBODIEDEVAL. On the left are the task text and part of the action space. On the right are observations from specific steps, along with the actions taken in the expert demonstration at those moments.

1. Introduction

In recent years, Multimodal Large Language Models (MLLMs) [53, 67, 82] have demonstrated strong understanding and reasoning capabilities on visual and language tasks. With the rapid development of MLLMs, a rich set of benchmarks for image understanding [25, 45, 56, 91] and video analysis [27, 63] has emerged. However, these benchmarks are non-interactive and insufficient in evaluating MLLMs' ability to handle tasks in real-world scenarios. Beyond these basic tasks which 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 [1, 10, 22, 62, 66, 84, 99]. 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 [15], visual grounding [4, 96], spatial reasoning [8], episodic memory [63], among others.

However, the comprehensive evaluation of MLLMs in embodied tasks remains largely unexplored. Existing benchmarks for embodied AI greatly lack diversity in tasks

and scenes, which limits their evaluation to specific aspects. EQA [18] dataset is restricted to only 9 template questions such as queries about object color and existence, while ALFRED [76] contains only 7 template tasks such as "pick and then place" within 4 room types. Moreover, many of embodied AI benchmarks require specific forms of input and output, making the evaluation of mainstream MLLMs inefficient or even infeasible. For example, R2R [4] uses 3D points as both input and output, REVERIE [71] requires predicting the bounding box of a target object, while ALFRED requires an object segmentation mask for interaction.

To address this gap, we present the first comprehensive evaluation benchmark for evaluating MLLMs' embodied capabilities in interactive environments, distinguishing it from the existing three types of evaluation benchmarks: static MLLMs benchmarks, which are non-interactive; embodied AI datasets, which are not comprehensive or diverse enough; and LLM agent benchmarks [54, 77], which heavily rely on textual environment states and have overly abstract actions (e.g., go to bathroom), thus downplaying or overlooking the critical embodied agent capabilities such as visual grounding and spatial reasoning. The key features of EMBODIEDEVAL are as follows:

Diverse Interactions. EMBODIEDEVAL provides a simulation framework that supports a variety of interactions, such as locomotion, question-answering, and interactions with objects and humans in realistic 3D scenes. Agents interact with the environment to gather new information or change its state to complete the task.

Diverse Tasks. In contrast to previous work that used task templates or tasks with minimal variation, our tasks have been systematically generated and carefully selected for high quality and diversity. EMBODIEDEVAL featuring a large number of novel tasks that involve a wider range of abilities to assess the comprehensive capabilities of the model as shown in Figure 1, which are categorized into five major categories.

Diverse Scenes. Unlike previous work, which was limited to household scenes, our scenes feature significant diversity in terms of objects and spaces, covering small rooms, large residences, and public spaces such as gym, store, office, among others. This approach can reduce the impact of scene types on the evaluation of the model's generalization, making the assessment more comprehensive.

2. Related Works

Multimodal Large Language Models. By connecting vision modules with LLMs, LLaVA [53] pioneers research in MLLMs through visual instruction tuning, and obtains impressive multimodal chat capabilities. Many work further improves the MLLMs from various aspects, including detailed captioning [9], trustworthy response [89, 90], multilingual multimodal capabilities [32, 80] and visual grounding [70, 87]. Beyond single-image understanding, some work explores more complicated tasks. For example, KOSMOS-1 [33] and VILA [49] focus on image-text interleaved understanding, while Video-LLaVA [48] and VideoChat [47] focus on video understanding.

Benchmark	Scene.	Task.	Disc.	Ego.	Nav.	Obj.	So.	Ans.
Video-MME [26]	-	✓	✓	✗	✗	✗	✗	✓
EgoPlan etc. [11, 14]	-	✓	✓	✓	✗	✗	✗	✓
OpenEQA [63]	-	✓	✓	✓	✗	✗	✗	✓
EQA etc. [18, 81, 88]	✗	✗	✓	✓	✓	✗	✗	✓
ALFRED [76]	✗	✗	✗	✓	✓	✓	✗	✗
BEHAVIOR[79]	✗	✓	✗	✓	✓	✓	✗	✗
EQA-MX [35]	✗	✗	✓	✓	✗	✗	✓	✓
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.

Evaluation for MLLMs. Mainstream benchmarks for MLLMs mainly focus on perception and cognitive evaluation, such as MME [25], MMB [56] and MMMU [91]. As existing MLLMs excel in these benchmarks, some benchmarks propose more challenging tasks, such as mathematical reasoning [31, 60, 93], OCR capability [57, 64, 78] and scientific knowledge [59, 91]. However, these benchmarks lack evaluation on egocentric vision, which is essential for broader applications of MLLMs. To address this gap, EgoVQA [24], EgoPlan-Bench [11], EgoThink [14], and OpenEQA [63] propose to evaluate the reasoning and planning capabilities of MLLMs given the first-person perspective images or videos. However, these benchmarks still use static question-answering pairs without interacting with environments.

Benchmarks for Embodied Agents. The datasets for embodied agents cover areas such as navigation, interaction, and question answering: (1) **Navigation** The R2R [4] dataset was the first to evaluate an agent's navigation ability under natural language instructions, followed by R4R [36] and RxR [44], which improve the fine-grained evaluation of the navigation process. In navigation tasks, object navigation is a crucial task because it serves as a prerequisite step for an embodied agent to interact with any object. There are many object navigation datasets with different focus and features, including SOON [98], REVERIE [72], DOZE [61] and GOAT-Bench [40]. (2) **Interaction** ALFRED [76] is the most representative interaction dataset, requiring the agent to follow instructions to complete tasks involving interactions such as picking up and placing ob-

jects. Additionally, there are datasets focusing on moving objects [28, 86], rearrangement [6, 83], tidying up a room [38], and household activities [42, 46, 65, 79]. (3) **Question Answering** EQA [18] first proposes the navigate-then-answer mechanism. Subsequently, more diverse EQA datasets have emerged, including those with questions involving multiple objects [88], requiring knowledge integration [81], set in realistic scenes [73], and handling situational queries [21]. Additionally, IQA [30] requires interaction with the environment to gather more observations in order to answer questions. EQA-MX [34] requires understanding non-verbal human expressions, such as body movements. However, existing benchmarks are limited in task variety, lacking comprehensive assessments of navigation, object interaction, and question-answering. They rely on highly repetitive 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

EMBODIEDEVAL consists of rich scenes and tasks to comprehensively evaluate the capabilities of embodied agents. We have implemented a unified simulation and evaluation framework and conducted meticulous data annotation to create the final dataset. First, we introduce the task categories involved in EMBODIEDEVAL in Section 3.1, followed by a detailed explanation of the simulation and evaluation framework in Section 3.2. After that, we describe the data collection and annotation process for this framework in Section 3.3. Finally, we show the data statistics of EMBODIEDEVAL in Section 3.4.

3.1. Task Categories

To comprehensively assess the capabilities of MLLMs as embodied agents, we synthesized and expanded upon existing embodied tasks while incorporating novel additions, broadly categorizing the tasks into five categories. Each category encompasses a diverse set of tasks and involves various capabilities. The task categories and sample tasks are shown in Figure 1.

Navigation. The navigation task involves coarse-grained and fine-grained natural language instructions, requiring the agent to navigate from its initial position to a target location and find a specific object if the task demands it.

Object Interaction. In object interaction tasks, agents must change the state of the environment through direct interaction with objects, such as moving objects, opening/-closing doors and drawers, and operating electrical devices. These tasks possibly involve multiple objects and require multi-step actions, such as using a tool to operate another object and rearranging items to meet certain requirements.

Social Interaction. Social interaction tasks encompass human-agent interactions, including item delivery, perspective-taking capabilities, human feedback interpretation, and non-verbal communication comprehension such as gestural cues.

Attribute Question Answering (AttrQA). AttrQA tasks necessitate exploring the environment to address questions regarding object and scene attributes. While incorporating traditional EQA tasks, AttrQA significantly broadens the scope of inquiry. It encompasses comprehensive attribute queries of objects and scenes, including but not limited to category, shape, material, color, function, state, location, existence, quantity, comparative analysis and reasoning across multiple attributes.

Spatial Question Answering (SpatialQA). Spatial understanding [8, 13, 23, 37, 51, 75] is a fundamental capability for embodied agents. SpatialQA requires agents to answer spatial-related questions through actions and observations, such as queries about size, height, position, distance, area, path, layout, spatial relationships, and more.

3.2. Evaluation Framework

We implemented a unified simulation framework based on LEGENT [17] platform. The framework aims to holisticly evaluate embodied abilities through diverse tasks, rather than focusing on specific tasks with particular input-output requirements. The observation space consists of task descriptions and egocentric vision. The action space consists of movement space, interaction space and answering space, which varies in each task instance. The action space is reasonably discretized and semantically understandable, while preserving the core evaluation objectives.

3.2.1 Action Space

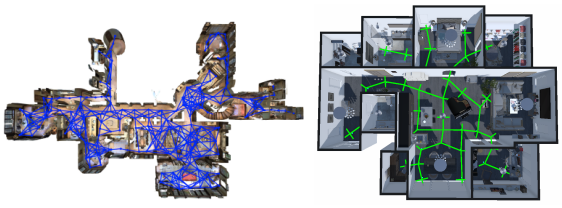


Figure 2. A comparison of navigation graphs between R2R [4] dataset (left) and EMBODIEDEVAL (right).

Movement Space 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 [4]. 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 2, 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).

Interaction Space We follows the discrete interaction space of previous embodied AI tasks that involves object interaction such as IQA [30], CHAI [65], RoomR [83] and OVMM[86] rather than continous space [76, 79]. 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.

Answering Space Unlike EQA and IQA, which require outputs of specific categories or use a very limited vocabulary, our answer space is very open. Responses are written by annotators and rigorously verified, encompassing a variety of possible replies to the question. For QA tasks, before exceeding the maximum allowed steps, there is no specified step at which the agent must respond; it can continue to explore until it believes it can select an answer. Once an answer is chosen, the task is immediately judged as correct or incorrect.

In this way, the action space is discretized and each action has a semantic text. EMBODIEDEVAL directly input the list of actions into the MLLM to decide the next action.

3.2.2 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 [36]. For example, the ***agent_pass*** predicate becomes true once the agent passes a specified navigation point.

A task is considered successful when all predicates evaluate to 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.

3.2.3 Evaluation Process

The process of an evaluation episode is as follows: (1) The simulator initializes the 3D scene and navigation graph. The agent is positioned at a designated starting point, and the initial first-person view image is saved into the observation history. (2) In each subsequent step, the agent chooses an action from a given list of options, including movement, interaction and answering, based on the observation history. The environment executes the action, changes the state accordingly, returns new observations, along with feedback indicating whether the action was successful. The observation, action, and feedback are then appended to the observation history.

This process continues until all success criteria are met resulting in task success, or the task fails due to an incorrect answer or exceeding the maximum allowed steps. For evaluating MLLMs designed for video understanding, image sequences are no longer used in the observation history. Instead, at each step, the simulator will output an egocentric video from the start to the current moment and use it as input for the MLLM.

3.3. Dataset Construction

The dataset construction process of EMBODIEDEVAL consists of three parts: scene collection, task collection, and task annotation. Each sample in the dataset requires sub-

Figure 3. The dataset construction pipeline of EMBODIEDEVAL.

stantial effort and undergoes rigorous annotation. Figure 3 illustrates our dataset construction pipeline.

3.3.1 Scene Collection

The diversity of scenes allows for a more accurate assessment of the agent's generalization capabilities. Our scenes are more varied in source and type than previous benchmarks, curated from four different sources: (1) **Objaverse Synthetic.** To expand our object variety, we utilize the Objaverse [20] dataset as our object database, which contains a vast collection of objects. Using procedural generation methods [19], we created numerous scenes from these objects. From these procedurally generated scenes, we manually selected and refined dozens of high-quality environments. (2) **AI2THOR.** We extracted and structured scenes from AI2THOR [43], focusing on indoor rooms with highly interactive objects. (3) **HSSD.** We included scenes from the Habitat Synthetic Scenes Dataset (HSSD) [39], which provides high-quality, realistic and complex scenes featuring extensive and diverse navigable spaces, including houses, villas, yards, and more. (4) **Sketchfab.** We enhanced our scene collection with selected free 3D scenes from Sketchfab[1]. These scenes are highly realistic and diverse, ranging from classrooms and supermarkets to offices and exhibitions, enriching the scope of our evaluation. These collected scenes serve as the basis for subsequent task annotation.

3.3.2 Task Collection

For task collection, we first gather seed tasks from over 30 existing datasets and benchmarks for embodied agents. Using these tasks as seeds, we prompted several advanced large language models to generate diverse task examples. To enhance task complexity, we designed some tasks 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, experts selected over 300 distinct candidate tasks. During task annotation, each task text could only be selected from this candidate set and used once. We chose this approach rather than allowing annotators to write tasks for given scenes, as it ensures task diversity, prevents repetition, and reduces dependency on individual annotators' creativity or preferences.

3.3.3 Task Annotation

Annoation Content Task sample annotation begins with an annotator selecting a task from the candidate pool and matching it to an appropriate scene. The annotator configures the action space detailed in Section 3.2.1, which includes movement space (via navigation graph adjustment), interaction space (via action options), and answering space (via answer options). Finally, the annotator defines success criteria as outlined in Section 3.2.2 through predicate function instantiation.

To maintain the validity and quality of the dataset, all task annotations must satisfy the following 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

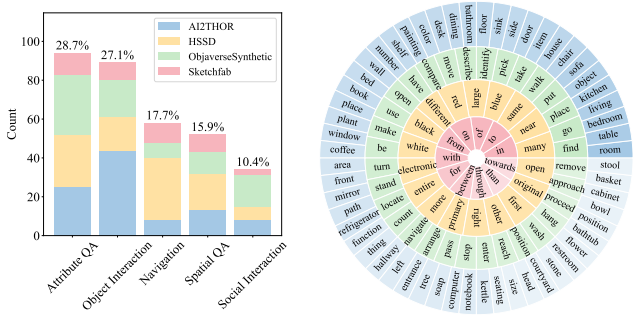
[1]https://sketchfab.com

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.

Annotation System To ensure both efficiency and precision in the complex annotation process, we developed an annotation system based on Unity[2]. The system provides comprehensively function, which encompassing scene and task import/export, flexible content viewing, visualized action space, and a guided annotation workflow that adheres to predefined guidelines (See Section 3.3.3).

Annoation Process Eight expert annotators are recruited to perform the annotations. Before starting the annotation process, we conducted systematic training on annotation requirements and system usage. To ensure the dataset's high quality, each annotated task was independently evaluated for correctness and quality by at least three reviewers. Furthermore, we validated task feasibility by providing expert demonstrations for each task and testing human performance with non-expert participants.

3.4. Dataset Statistics



(a) Number of tasks by category for each scene source.

(b) Visualization of top vocabulary by POS and frequency.

EMBODIEDEVAL consists of 328 human-annotated tasks divided into 5 categories, distributed across 125 unique scenes and demonstrating a high level of visual diversity and complexity. Figure 4a shows the distribution of the task across 5 task categories and 4 scene sources. Overall, the dataset comprises 575 predicate instances and 1,533 action options excluding movement space, including 1,213 answers and 320 interactions, with each episode averaging 10.72 steps based on expert demonstrations. Task descriptions average 16.09 words in length, while options average 5.72 words. To showcase the lexical diversity in EMBODIEDEVAL, Figure 4b presents a visualization of the most frequent words categorized by grammatical type.

[2]https://unity.com/

4. Experiments

4.1. Evaluation Settings

We evaluate the models using the evaluation process described in Section 3.2.3. For multi-image MLLMs, we input the task prompt, history information (egocentric image observation history, action history, feedback history), and the action space (movement, interaction, answering) into the model, which then outputs the corresponding action option to be executed in the environment. For video MLLMs, we replace the image observation history with egocentric video. We set the maximum number of attempt steps for each task to 24, as all tasks are completed within this number of steps in the expert demonstration. We use an image resolution of 448×448.

We evaluate 22 MLLMs on EMBODIEDEVAL including closed-source MLLMs, open-source multi-image MLLMs and open-source video MLLMs demonstrating strong performance on MLLM benchmarks like MMMU [92] and Video-MME [12]. For reference, we also introduce two special agents: (1) *Random* agent uniformly samples actions from the available option set at each timestep. (2) *Human* agent involves non-expert humans unfamiliar with the tasks, performing actions in the simulator's user interface using the same observation and action space as the models.

4.2. Evaluation Metrics

We evaluate agent performance using three metrics. (1) **Success Rate (Succ.)** [2, 43, 55, 74, 97] is the primary metric we use to measure the percentage of tasks that the agent fully completes. (2) For multi-goal tasks, **Goal-condition Success (GcS)** [41, 76] measures partial success by calculating the proportion of goal conditions achieved, as specified by predicate functions. (3) **Success weighted by Path Length (SPL)** [3, 7] is a path-weighted metric to evaluate task execution efficiency in navigation and interaction tasks. SPL assesses an agent's performance by considering both task success and the path efficiency relative to the expert demonstration.

4.3. Main Results

EMBODIEDEVAL Reveals Limitations of Current MLLMs on Embodied Tasks. As shown in Table 2, Success rates across various models on EMBODIEDEVAL remain notably low. The best-performing model, GPT-4o, achieves only a 25.00% overall success rate and a 32.42% GcS score. This is in stark contrast with non-expert humans, who reach a nearly perfect success rate of 97.26%, highlighting the difficulty these models face in executing embodied tasks that humans find trivial. This performance deficit is further underscored by significantly lower SPL scores, indicating that the models struggle to find optimal solutions. This shows the challenges for deploying MLLMs

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
Closed-Source Multi-Image MLLMs													
GPT-4o [68]	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**
GPT-4o-Mini [68]	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
Gemini-Pro [29]	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
Gemini-Flash [29]	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
Qwen-VL-Max [5]	**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
Qwen-VL-Plus [5]	10.53	11.54	3.45	10.49	3.45	0.00	2.43	0.00	2.94	8.82	1.68	5.79	8.31
Open-Source Multi-Image MLLMs													
InternVL2-40B [69]	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
InternVL2-8B [69]	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-Llama3-76B [69]	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-NEXT-72B [12]	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 [52]	**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**
LLaVA-OneVision-7B [52]	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
VILA-40B [50]	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
VILA-8B [50]	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
Open-Source Video MLLMs													
LLaVA-NeXT-Video-32B-Qwen [94]	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 [95]	**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**
LLaVA-Video-7B-Qwen2 [95]	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
Oryx-34B [58]	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
VideoLLaMA2-72B [16]	**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
VideoLLaMA2-7B [16]	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

Table 2. Results of different models on EMBODIEDEVAL (%). The best-performing model in each category is bolded.

in embodied scenarios.

Performance Gap between Best Proprietary and Open-source Models. Proprietary models demonstrate a consistent advantage across all tasks and metrics. GPT-4o leads in overall performance, with the highest success rate on four out of five tasks. Qwen-VL-Max ranks second, achieving the highest success rate in Attribute QA. In contrast, open-source models show a substantial performance gap. The top performing multi-image MLLM, LLaVA-OneVision-72B, achieves an overall success rate of 12.80%, barely competitive with proprietary models. The other multi-image MLLMs have accuracy rates of around 10% at best. Among the video MLLMs, VideoLLaMA2-72B achieves the highest success rate (12.81%), while LLaVA-Video-72B-Qwen2 achieves the highest GcS (16.95%). For open-source models, larger models often exhibit improved performance, but this trend is not universally observed.

Model Performance across Different Task Types. The results highlight a significant 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 commercial 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. The Navigation task shows substantial performance variability across models, such as Gemini-Flash, which achieves 26.32% in Attribute QA but drops to only 5.17% in Navigation. Overall, the scores for Object Interaction and Social Interaction are consistently lower across all models, underscoring the challenge these models face in scenarios that require a deeper understanding of affordance or social cues.

Challenges in Long-Horizon Tasks. The performance of MLLMs shows a significant decline as the number of steps required for the task and the subgoals increase. As illustrated in Figure 5, models maintain relatively high success counts at lower required steps, but their performance drops fluctuatingly as tasks require longer sequences of actions. This drop in performance can be attributed to two primary factors: (1) the increased complexity and reason-

ing demands associated with longer tasks and multiple subgoals, which challenge the models' planning and decision-making capabilities, and (2) the limitations in retaining and processing the long temporal context necessary for successful execution in extended tasks. These challenges highlight the need for further model improvements to enhance MLLM capabilities in handling complex, multi-step objectives over long horizons.

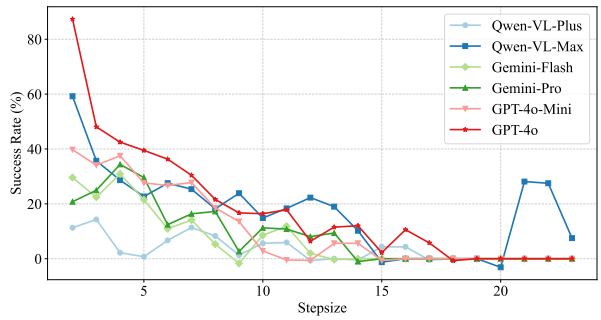


Figure 5. Success rate vs. number of steps required for the task.

Studies on Temperature. 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 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.

4.4. Error 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.

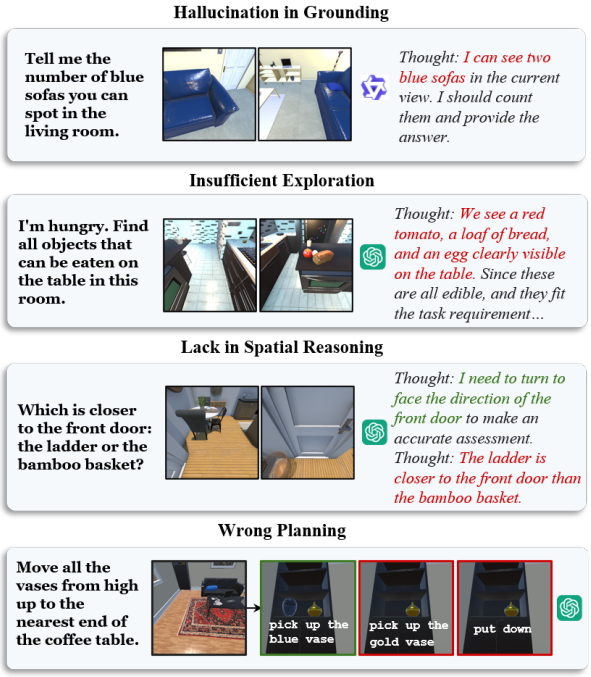


Figure 6. Case study of common error categories. In *Hallucination in Grounding*, the agent mistakenly identified a single blue sofa as two. In *Insufficient Exploration*, the agent failed to look for additional items. In *Lack in Spatial Reasoning*, the agent misestimated the distance between objects. In *Wrong Planning*, the agent did not organize the picking up and putting down of the vases in the proper order and at the correct positions.

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 found 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.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[2] Peter Anderson and et al. Vln-ce: A visually-grounded language navigation dataset for following compositional instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8470–8480, 2020. 6

[3] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 6

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 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, 2018. 1, 2, 3

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 7

[6] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020. 3

[7] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 6

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 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, 2024. 1, 3

[9] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2

[10] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint arXiv:2310.02071*, 2023. 1

[11] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*, 2023. 2

[12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6, 7

[13] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 3

[14] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Can vision-language models think from a first-person perspective? *arXiv preprint arXiv:2311.15596*, 2023. 2

[15] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024. 1

[16] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 7

[17] Zhili Cheng, Zhitong Wang, Jinyi Hu, Shengding Hu, An Liu, Yuge Tu, Pengkai Li, Lei Shi, Zhiyuan Liu, and Maosong Sun. Legent: Open platform for embodied agents. *arXiv preprint arXiv:2404.18243*, 2024. 3

[18] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018. 2, 3

[19] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 5

[20] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 5, 1

[21] Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadhan, and Dinesh Manocha. S-eqa: Tackling situational queries in embodied question answering. *arXiv preprint arXiv:2405.04732*, 2024. 3

[22] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[23] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial un-

derstanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024. 3

[24] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[25] 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. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 2

[26] 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. Video-mme: The first-ever comprehensive evaluation of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2

[27] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1

[28] Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund, Daniel LK Yamins, James J DiCarlo, Josh McDermott, Antonio Torralba, et al. The threedworld transport challenge: A visually guided task-and-motion planning benchmark towards physically realistic embodied ai. In *2022 International conference on robotics and automation (ICRA)*, pages 8847–8854. IEEE, 2022. 3

[29] Google. Our next-generation model: Gemini 1.5, 2024. 7

[30] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018. 3, 4

[31] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. 2

[32] Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*, 2023. 2

[33] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[34] Md Mofijul Islam, Alexi Gladstone, Riashat Islam, and Tariq Iqbal. Eqa-mx: Embodied question answering using multi-modal expression. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[35] Md Mofijul Islam, Alexi Gladstone, Riashat Islam, and Tariq Iqbal. EQA-MX: Embodied question answering using multimodal expression. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[36] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*, 2019. 2, 4

[37] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 3

[38] Yash Kant, Arun Ramachandran, Sriram Yenamandra, Igor Gilitschenski, Dhruv Batra, Andrew Szot, and Harsh Agrawal. Housekeep: Tidying virtual households using commonsense reasoning. In *European Conference on Computer Vision*, pages 355–373. Springer, 2022. 3

[39] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. 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, 2024. 5

[40] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. 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, 2024. 2

[41] Byeonghwi Kim, Jinyeon Kim, Yuyeong Kim, Cheolhong Min, and Jonghyun Choi. 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, 2023. 6

[42] Taewoong Kim, Cheolhong Min, Byeonghwi Kim, Jinyeon Kim, Wonje Jeung, and Jonghyun Choi. Realfred: An embodied instruction following benchmark in photo-realistic environments. *arXiv preprint arXiv:2407.18550*, 2024. 3

[43] Eric Kolve and et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 5, 6

[44] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020. 2

[45] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1

[46] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang,

Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023. 3

[47] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2

[48] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

[49] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 2

[50] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 7

[51] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 3

[52] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 7

[53] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[54] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023. 2

[55] 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. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. 6

[56] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 2

[57] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 2

[58] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 7

[59] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[60] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[61] Ji Ma, Hongming Dai, Yao Mu, Pengying Wu, Hao Wang, Xiaowei Chi, Yang Fei, Shanghang Zhang, and Chang Liu. Doze: A dataset for open-vocabulary zero-shot object navigation in dynamic environments. *arXiv preprint arXiv:2402.19007*, 2024. 2

[62] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024. 1

[63] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 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, 2024. 1, 2

[64] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 2

[65] Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. Mapping instructions to actions in 3d environments with visual goal prediction. *arXiv preprint arXiv:1809.00786*, 2018. 3, 4

[66] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[67] OpenAI. Gpt-4v(ision) system card, 2023. 1

[68] OpenAI. Hello gpt4-o, 2024. 7

[69] OpenGVLab. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 7

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

[71] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 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, 2020. 2

[72] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den

Hengel. Reverie: Remote embodied visual referring expression in real indoor environments, 2020. 2

[73] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024. 3

[74] Manolis Savva and et al. Habitat challenge: A photorealistic embodied ai benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9338–9346. IEEE, 2019. 6

[75] Manasi Sharma. Exploring and improving the spatial reasoning abilities of large language models. In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*, 2023. 3

[76] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 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, 2020. 2, 4, 6

[77] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020. 2

[78] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2

[79] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, pages 477–490. PMLR, 2022. 2, 3, 4

[80] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. Parrot: Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*, 2024. 2

[81] Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. Knowledge-based embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11948–11960, 2023. 2, 3

[82] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[83] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5922–5931, 2021. 3, 4

[84] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*, 2023. 1

[85] 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. Holodeck: Language guided generation of 3d embodied ai environments, 2024. 1

[86] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*, 2023. 3, 4

[87] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2

[88] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019. 2, 3

[89] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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, 2024. 2

[90] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 2

[91] 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. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 1, 2

[92] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 6

[93] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. 2

[94] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 7

[95] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7

[96] Zhuofan Zhang, Ziyu Zhu, Pengxiang Li, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding in 3d scenes. *arXiv preprint arXiv:2408.04034*, 2024. 1

[97] Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Wang. Vlmbench: A compositional benchmark for vision-and-language manipulation. *Advances in Neural Information Processing Systems*, 35:665–678, 2022. 6

[98] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. 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, 2021. 2

[99] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 1

6. Appendix

6.1. Task Samples

We selected some representative examples to illustrate the diversity of the task set in Table 3.

6.2. Details of Evaluation Framework

6.2.1 Evaluation Formulation

We formalize the evaluation process mentioned in Section 3.2.3 as follows. The prompt we used for the MLLMs during the evaluation is shown in Figure 8.

Algorithm 1 EMBODIEDEVAL Evaluation Process

Input: A Multimodal LLM π, an evaluation task including 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 $success$.

$o, s \leftarrow E.reset(x)$ ▷ E is the simulator, o is the observed image, s is the world state
$H_o \leftarrow \{o\}$ ▷ observation history
$H_a \leftarrow \emptyset$ ▷ action history
for $i \leftarrow 0$ to max steps **do**
  $a \leftarrow \pi.predict(g, \mathcal{C}, H_o, H_a)$
  $o, s \leftarrow E.step(a)$
  $H_o.append(o)$
  $H_a.append(a)$
  $done \leftarrow P.judge(s)$
  if $done$ **then**
    return true
  else if a is answer action **then**
    return false ▷ wrong answer
  end if
end for
return false ▷ reach the max steps

6.2.2 Interaction Actions

We provided more examples of the interaction space mentioned in Section 3.2.1 in Table 5.

6.2.3 Predicates

All the predicate functions described in Section 3.2.2 are listed in Table 4.

6.3. 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 [85] annotated realistic and diverse objects choosen from the Objaverse asset library [20]. 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 (see examples in Figure 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.

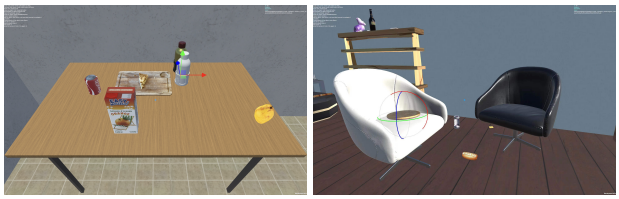


Figure 7. Interactive scene editor: adjust object position (left) and angle (right).

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 7). Once editing is complete, the scene can be saved as a JSON file, which can be imported to reproduce the environment.

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
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?	spatial reasoning
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
Get close to the lady in white and ask if she needs help.	social navigation
Wake up my dad. He is sleeping in the bedroom. The bedroom is the second room on your right as you walk forward.	finding someone
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

Table 3. Examples of the diverse tasks 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 4. 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 5. Some cases of the interaction actions involved in EMBODIEDEVAL.

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 8. Prompt for Multi-image MLLMs.

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

6.4. Success Cases

We present successful cases accomplished by closed-source MLLMs to gain deeper insights into their current capabilities. As discussed in Section 4.3, 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.

6.4.1 Attribute QA

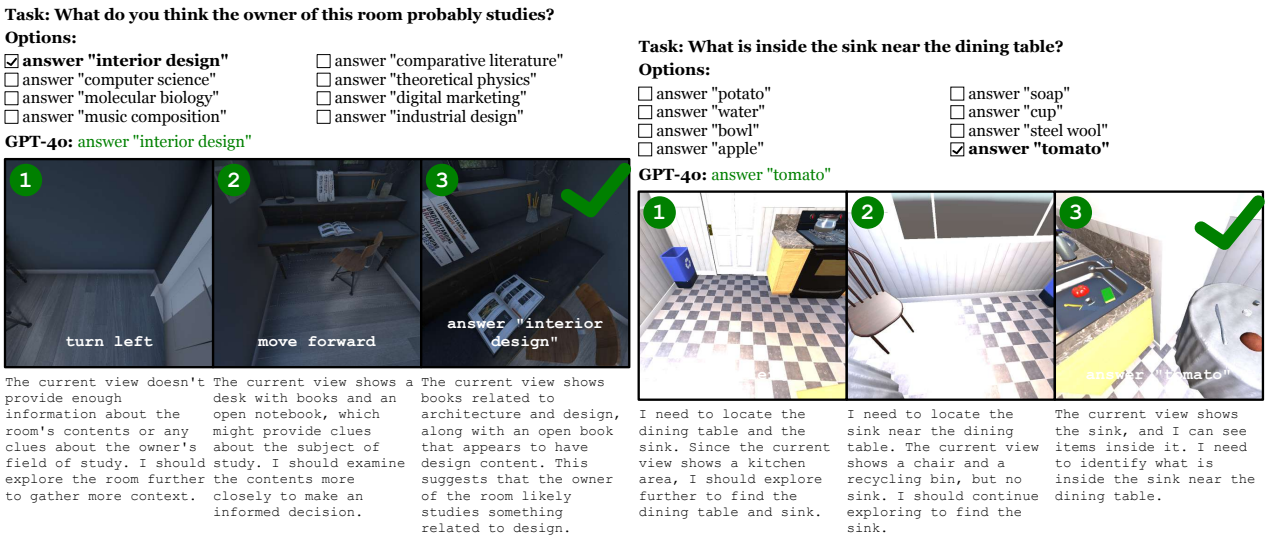


Figure 10

Task: What color is the car parked around the house?
Options:

☐ answer "red"
☐ answer "white"
☐ answer "black"
☐ answer "blue"
☑ **answer "red and white"**
☐ answer "blue and white"
☐ answer "black and white"
☐ answer "yellow"

Qwen-VL-Max: answer "red and white"

Figure 11

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."

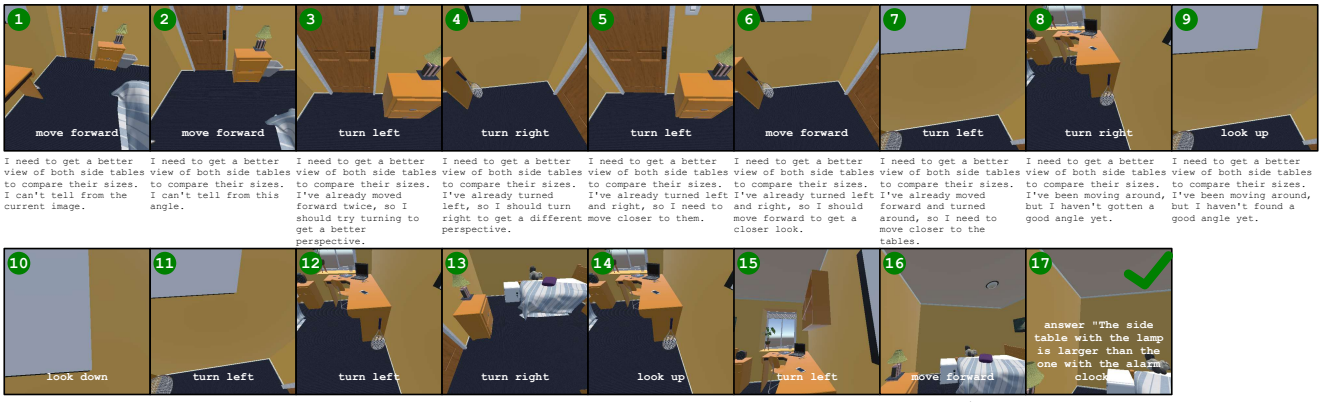


Figure 12

6.4.2 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.

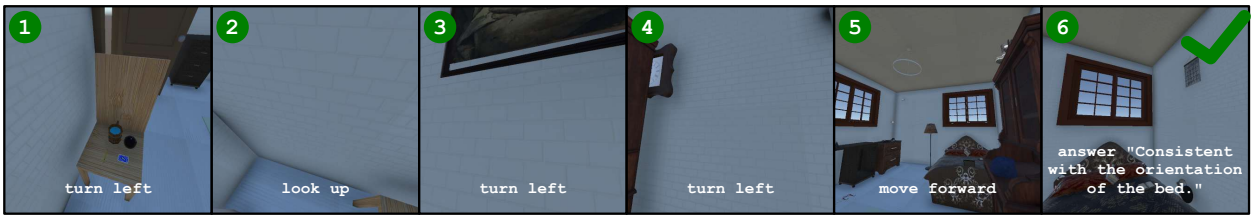
I need to determine if the sofas can fit through the front door. I should compare the size of the sofas to the door. I have already seen the door and the sofas, so I can make a decision based on the current view.

Figure 13

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.

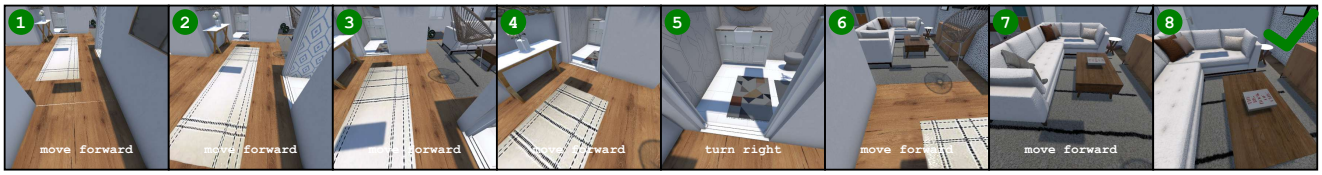
I need to determine the orientation of the bedroom doorway. The current view shows the bed and windows, which might help in understanding the room's layout. Since the task is about the doorway's direction, I should explore more to find the doorway.

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 14

6.4.3 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:



I need to walk straight past the kitchen on my left, which means I should continue moving forward.

I need to walk straight past the kitchen on my left, so I should continue moving forward.

I need to walk straight past the kitchen on my left, which means I should continue moving forward.

I need to walk straight past the kitchen on my left, which means I should continue moving forward until I am almost at the bathroom door. Then, I will turn right.

I need to turn right after passing the kitchen and before reaching the bathroom door.

I need to walk straight past the kitchen on my left, which means I should continue moving forward until I reach the point where I need to turn right. The current view shows that I am still in the hallway, so I should keep moving forward.

I need to move closer to the sofa table to find the book.

Figure 15

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:

I need to find the human-shaped statue on the table below the wall where the iris painting hangs. I should start by looking around to identify any potential locations that match this description.

The task requires me to find a human-shaped statue on a table below a wall with an iris painting. I should start by looking around to identify any potential locations that match this description.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should start by exploring the room and looking for any paintings or statues. Since I am currently facing a blank wall, I should turn left to explore other parts of the room.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the wall with the iris painting to locate the statue.

The task requires me to find a human-shaped statue on a table below the wall where the iris painting hangs. I should move towards the wall with the iris painting to locate the statue.

The task requires me to find a human-shaped statue on a table below the wall where the iris painting hangs. I should move towards the wall with the iris painting to locate the statue.

The task requires me to find a human-shaped statue on a table below a wall with an iris painting. I should move towards the wall with the iris painting to locate the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the current view. Moving forward will bring me closer to the painting and potentially the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the background of the current view. Moving forward will bring me closer to that area.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the background of the current view. Moving forward will bring me closer to the painting and potentially the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the current view. Moving forward will bring me closer to the painting and potentially the statue.

The task requires me to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move towards the area where the iris painting is located, which is visible in the current view. Moving forward will bring me closer to the target location.

The task is to find a human-shaped statue on a table below a wall where the iris painting hangs. I need to move closer to the table where the statue is located.

The task is to find a human-shaped statue on a table below a wall where the iris painting hangs. I should move closer to the table to confirm if the statue is there.

The task requires me to find a human-shaped statue placed on a table below a wall where the iris painting hangs. I have been moving around the room and now I am near the window. I should turn right to face the wall with the iris painting and locate the statue.

The task is to find a human-shaped statue placed on a table below the wall where the iris painting hangs. I should move closer to the table to inspect it.

Figure 16

6.4.4 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."

I need to find the fridge first to check for eggs. The current view does not show the fridge, so I should turn to face it.

To determine if there is an egg inside the fridge, I need to first locate the fridge and then check its contents. The current view does not show the fridge, so I should turn to find it.

To check for eggs in the fridge, I need to go to the fridge and open it. The fridge is located to the right of my current position.

To check for eggs in the fridge, I need to go to the fridge and open it. The fridge is located to the right of my current position.

To check for eggs in the fridge, I need to go to the fridge and open it. The fridge is not visible in the current view, so I should move towards it.

To check for eggs in the fridge, I need to go to the fridge and open it. The current view does not show the fridge, so I need to turn towards it.

To check for eggs in the fridge, I need to go closer to the fridge and then open it to inspect its contents.

To check for eggs in the fridge, I need to approach the fridge and open it. The current view shows the fridge in the distance, so I should move closer to it.

I need to check inside the fridge to determine if there are any eggs. The current view shows the top of the fridge, so I should open it to see the contents.

I need to check inside the fridge to determine if there are any eggs. Since I am already at the fridge, I should open the doors to inspect the contents.

I need to check inside the fridge to determine if there are any eggs. The previous attempts to open the fridge were unsuccessful, so I will try again.

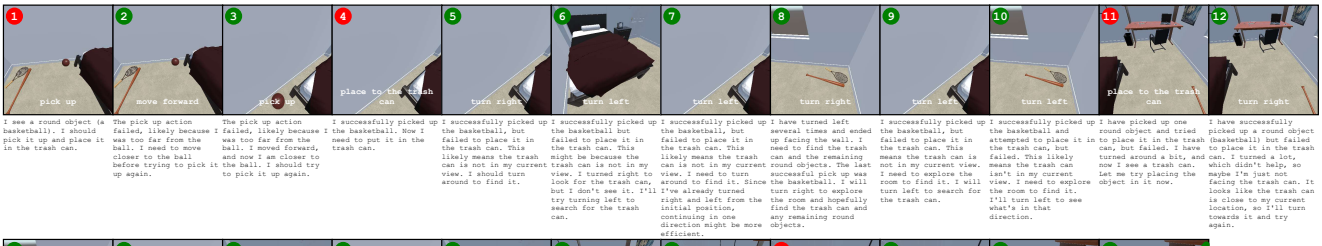
I need to check inside the fridge to determine if there are any eggs. The previous attempts to open the upper fridge door were unsuccessful, so I should try opening the lower fridge door instead.

I need to check if there is an egg inside the fridge. I should open the fridge doors to look inside.

The task is to determine if there is an egg inside the fridge. I have already opened the lower fridge door and can see an egg inside. Therefore, I can confirm that there is an egg in the fridge.

Figure 17

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



I see a round object (a basketball). I should pick it up and place it in the trash can.

The pick up action failed, likely because I was too far from the ball. I need to move closer to the ball before trying to pick it up again.

The pick up action failed, likely because I was too far from the ball. I moved forward, and now I am closer to the ball. I should try to pick it up again.

I successfully picked up the basketball. Now I need to put it in the trash can.

I successfully picked up the basketball, but failed to place it in the trash can. This likely means the trash can is not in my current view. I should turn around to find it.

I successfully picked up the basketball but failed to place it in the trash can. This might be because the trash can is not in my view. I turned right to look for the trash can, but I don't see it. I'll try turning left to search for the trash can.

I successfully picked up the basketball, but failed to place it in the trash can. This likely means the trash can is not in my current view. I need to turn around to find it. Since I've already turned right and left from the initial position, continuing in one direction might be more efficient.

I have turned left several times and ended up facing the wall. I need to find the trash can and the remaining round objects. The last successful pick up was the basketball. I will turn right to explore the room and hopefully find the trash can and any remaining round objects.

I successfully picked up the basketball, but failed to place it in the trash can. This means the trash can is not in my current view. I need to explore the room to find it. I will turn left to search for the trash can.

I successfully picked up the basketball and attempted to place it in the trash can, but failed. This likely means the trash can isn't in my current view. I need to explore the room to find it. I'll turn left to see what's in that direction.

I have picked up one round object and tried to place it in the trash can, but failed. I have turned around a bit, and now I see a trash can. Let me try placing the object in it now.

I have successfully picked up a round object (basketball) but failed to place it in the trash can. I turned a lot, which didn't help, so maybe I'm just not facing the trash can. It looks like the trash can is close to my current location, so I'll turn towards it and try again.

I have successfully picked up one round object, the basketball, but failed to place it in the trash can. I've tried placing it from different angles, but it hasn't worked. I need to find the trash can. I can see from previous images that it's near the desk. I will turn around and navigate to the desk to find the trash can. I'm currently looking at the wall where I failed to place the object last time, so turning right might give me a view of the desk.

I successfully picked up a round object, but failed to place it in the trash can. This might be because I wasn't facing the trash can. I need to turn around and locate the trash can before attempting to place the object again. I will turn right first to explore the room.

I have picked up one round object and tried to place it in the trash can but failed, possibly because I wasn't facing the trash can. I turned around a bunch, then tried again and still failed. I need to find the trash can. I'll turn to face the desk, because the trash can was next to it.

I have picked up the basketball and tried to place it in the trash can, but failed. I need to navigate to the trash can. I last saw the trash can when I looked at the desk. I will turn around to face the desk area again.

I successfully picked up a round object, but failed to place it in the trash can. This could be because I wasn't facing the trash can. I need to navigate to the trash can before attempting to place the object again. I'll turn to explore the room and find it.

I successfully picked up the basketball and failed to place it in the trash can. I've turned a lot since then, and now I'm facing the desk. The trash can is near the desk.

I have successfully picked up one round object. I failed to place it in the trash can perhaps because I was not facing it. I've turned and moved around a bit, and now I see the trash can next to the desk. I should move closer to it.

I have picked up one round object and tried to place it in the trash can but failed maybe because I was not close enough. I moved closer to the trash can so now I can try placing the object in the trash can again.

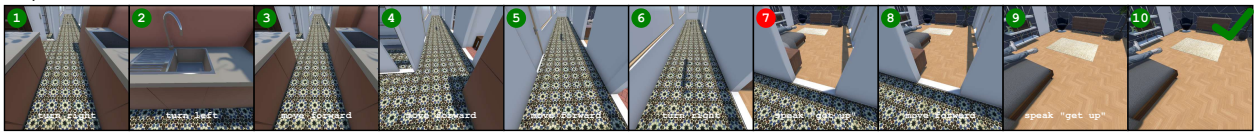
I have picked up one round object and tried to place it in the trash can multiple times, but failed. I suspect I'm not facing the trash can. I need to explore and locate the trash can. I've already turned left and right multiple times. I will turn around to explore the space behind me.

I successfully picked up the basketball but failed to place it in the trash can. I have been turning around a lot, trying to find the trash can. Now I am looking at the trash can, so I should try placing the basketball in it.

Figure 18

6.4.5 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:



To wake up your dad, I need to navigate to the bedroom, which is the second room on the right. Currently, I am facing a hallway with options to turn left or right. Since the bedroom is to the right, I should turn right to head in the correct direction.

I need to navigate towards the bedroom to wake up my dad. Since I have already turned right, the next logical step is to continue moving in the direction that will lead me to the bedroom. The available options include turning left or right, looking up, or speaking. Since turning right again would not be effective, I will turn left to continue towards the bedroom.

I need to navigate towards the bedroom, which is the second room on my right as I walk forward. Since I have already turned right and left without making progress, I should move forward to get closer to the bedroom.

I need to reach the bedroom, which is the second room on my right as I walk forward. Currently, I am facing a hallway with a clear path ahead. Since I have already turned right and left previously, I should move forward to get closer to the bedroom.

I need to wake up my dad, who is in the bedroom, which is the second room on my right as I walk forward. Currently, I am facing a hallway with a view that suggests I am close to the bedroom. Since I have already moved forward twice and turned right and left without success, I should now move forward to reach the bedroom.

I need to wake up my dad, who is in the bedroom, which is the second room on my right as I walk forward. Currently, I am facing a hallway, and I have already moved forward multiple times. Since I am not yet in the bedroom, I should turn right to head towards the bedroom.

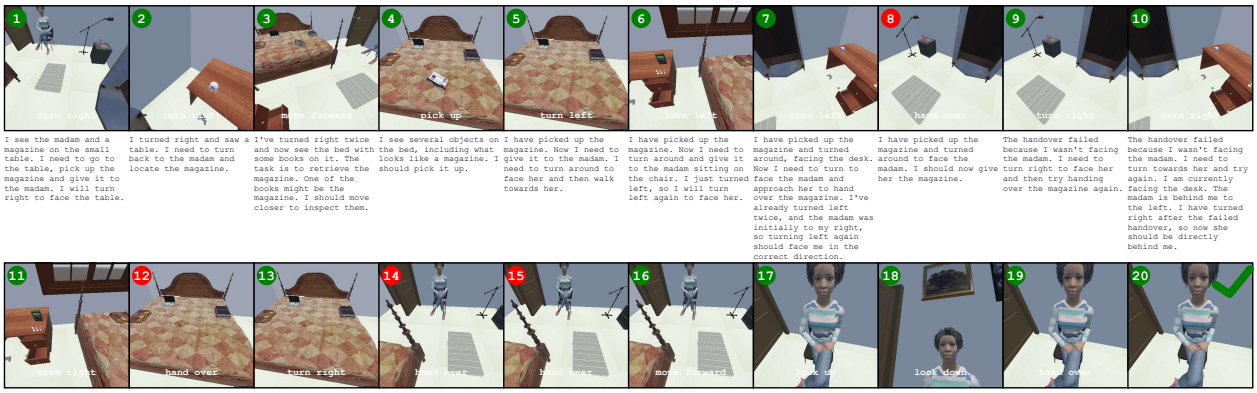
I am currently in front of the bedroom where my dad is sleeping. Since the task is to wake him up, the most direct action is to speak to him. This will allow me to complete the task without unnecessary movement.

I am currently in front of the bedroom where my dad is sleeping. I previously attempted to wake him by speaking, but it failed. Since I am already facing the bedroom, I should try to move forward into the room to get closer to him. This will allow me to interact more effectively.

I am currently in the bedroom where my dad is sleeping. I previously attempted to wake him by speaking, but it failed. Since I am already in the right location, I should try a different approach. I can either try to speak again or look for a way to wake him up physically. However, since I cannot interact with objects or people unless they are very close, I will try speaking again to see if it has a different effect this time.

Figure 19

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



I see the madam and a magazine on the small table. I need to go to the table, pick up the magazine and give it to the madam. I will turn right to face the table.

I turned right and saw a table. I need to turn back to the madam and locate the magazine.

I've turned right twice and now see the bed with some books on it. The task is to retrieve the magazine. One of the books might be the magazine. I should move closer to inspect them.

I see several objects on the bed, including what looks like a magazine. I should pick it up.

I have picked up the magazine. Now I need to give it to the madam. I need to turn around to face her and then walk towards her.

I have picked up the magazine. Now I need to turn around and give it to the madam sitting on the chair. I just turned left, so I will turn left again to face her.

I have picked up the magazine and turned around, facing the desk. Now I need to turn to face the madam and approach her to hand over the magazine. I've already turned left twice, and the madam was initially to my right, so turning left again should face me in the correct direction.

I have picked up the magazine and turned around to face the madam. I should now give her the magazine.

The handover failed because I wasn't facing the madam. I need to turn right to face her and then try handing over the magazine again.

The handover failed because I wasn't facing the madam. I need to turn towards her and try again. I am currently facing the desk. The madam is behind me to the left. I have turned right after the failed handover, so now she should be directly behind me.

I successfully picked up the magazine but failed to hand it over. This likely means I wasn't facing the woman when I tried. I need to turn around to face her and try again. I'm currently facing the bed and desk, so turning left would put me back in the position where I unsuccessfully tried to hand over the magazine. Turning right would orient me towards the doorway, so I will turn right.

I successfully picked up the magazine, but handing it over failed. This means I wasn't facing the woman when I tried. I need to turn around to face her and then hand over the magazine. I turned around three times after picking it up. I've done another three turns since the failed handover, meaning I should now be facing the correct direction after three right turns.

I successfully picked up an item, but handing it over failed twice. This suggests I picked up the wrong item. I'm currently looking at the bed, and there are a few items here. I need to find the magazine. I'll turn around to see if it is elsewhere.

I have successfully picked up the magazine, I previously failed to hand it over because I was not facing the madam. Now I'm facing her, so I should be able to hand it over.

I have the magazine, I failed to hand it over multiple times because I was not facing the madam. I am now facing her, so I should try handing it over again.

I have the magazine, and I'm facing the madam. I've tried handing it over multiple times from slightly different angles, but it hasn't worked. Perhaps I need to move a bit closer to ensure I'm within the interaction range.

I have the magazine and I am facing the madam. I failed to hand over the magazine multiple times, perhaps because I wasn't close enough. I moved forward in the previous step, but perhaps not enough. I will try moving closer again before attempting to hand over.

I successfully picked up the magazine earlier, and then tried handing it over multiple times from different angles, but failed. The last view shows I am very close to the madam, which may be too close. I will step back a little and try to hand over the magazine again. I'll try moving back first to create some distance.

I have the magazine and am facing the woman. I failed to hand it over multiple times, likely because I wasn't close enough. I moved forward in the last step, so now I'll try handing it over again.

Figure 20

6.5. Error Cases

6.5.1 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"

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

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

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.

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.

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

Figure 21

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."

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

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.

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.

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.

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.

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 22

6.5.2 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."

I need to find a location that is accessible from both the kitchen and the living room. The current view shows a transition area between the kitchen and another room, possibly the living room. I should explore further to confirm the layout and find the best spot for the fire extinguisher.

I need to explore more to find a location accessible from both the kitchen and the living room. Turning left might give a better view of the area.

I need to find a location that is accessible from both the kitchen and the living room. The current view shows a hallway leading to a room with a round carpet, which might be the living room. I should explore further to confirm the layout.

I need to find a location that is accessible from both the kitchen and the living room. Based on the current view, I am near a door and a table, which might be a good central location. I should explore further to confirm the best spot.

I need to find a location that is accessible from both the kitchen and the living room. The current view shows a hallway, which might connect these areas. I should explore further to find a suitable spot.

I need to find a location that is accessible from both the kitchen and the living room. The current view shows a hallway, which might connect these areas. I should explore further to confirm the layout.

I need to determine a location for the fire extinguisher that is accessible from both the kitchen and the living room. Based on the visual history, I should explore more to find a suitable spot. Turning left might reveal more of the area.

I need to find a location that is accessible from both the kitchen and the living room. Based on the visual history, I should explore more to find a suitable spot. The current view shows a door, which might lead to another area. I should open it to explore further.

The current view shows a hallway that seems to connect the kitchen and the living room. This location might be a good spot for the fire extinguisher as it is accessible from both areas. I should confirm the surroundings to ensure it's the best spot.

I need to find a location that is accessible from both the kitchen and the living room. The current view shows a hallway leading to both areas. I should explore further to find a suitable spot.

I need to find a location that is accessible from both the kitchen and the living room. The current view shows the living room area. I should explore the area further to find a suitable spot.

The task is to find a location for a fire extinguisher that is accessible from both the kitchen and the living room. Based on the current view and previous exploration, the area beside the white sofa seems to be a central location between the kitchen and living room.

Figure 23

6.5.3 Navigation

Figure 24

6.5.4 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:

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.

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.

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.

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.

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.

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.

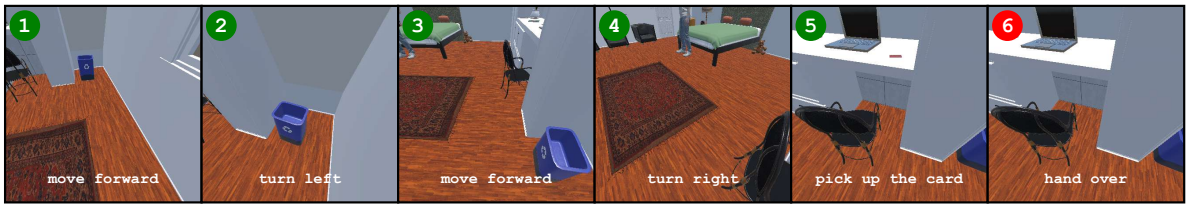
I have successfully picked up the reading materials. Now, I need to place them on the lounge chair. The current view shows the wall, so I need to turn to face the lounge chair.

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 25

6.5.5 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:



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.

The current view shows a trash can, but I don't see a credit card. I should turn to explore the area further.

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.

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.

The credit card is visible on the desk in front of me. The next logical step is to pick it up.

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.

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.

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.

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.

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.

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 26