OST-Bench: Evaluating the Capabilities of MLLMs in Online Spatio-temporal Scene Understanding

Jingli Lin[1,2*], **Chenming Zhu**[1,3*], **Runsen Xu**[1,4], **Xiaohan Mao**[1,2], **Xihui Liu**[3]
Tai Wang[1†], **Jiangmiao Pang**[1†]

[1]Shanghai AI Laboratory, [2]Shanghai Jiao Tong University,
[3]The University of Hong Kong, [4]The Chinese University of Hong Kong
[*]Equal contribution [†]Co-corresponding

`https://rbler1234.github.io/OSTBench.github.io/`

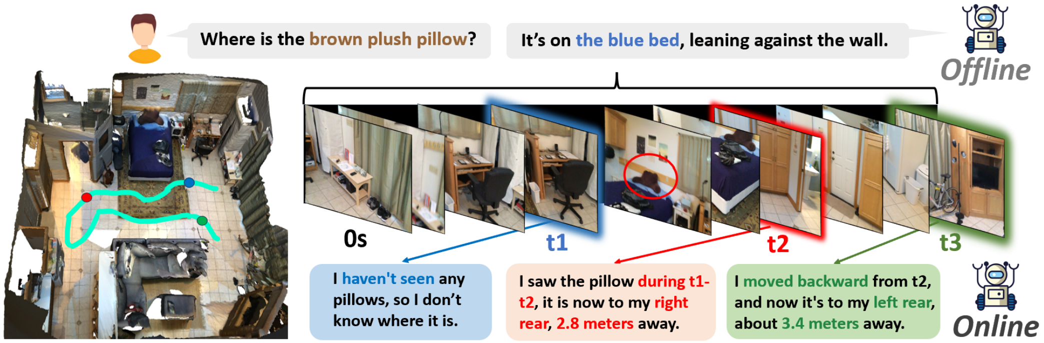

Where is the brown plush pillow?
It's on the blue bed, leaning against the wall.
Offline
0s
t1
t2
t3
I haven't seen any pillows, so I don't know where it is.
I saw the pillow during t1-t2, it is now to my right rear, 2.8 meters away.
I moved backward from t2, and now it's to my left rear, about 3.4 meters away.
Online

Figure 1: **OST-Bench** is designed from the perspective of an embodied agent dynamically exploring static indoor environments, with a focus on **online** and **spatio-temporal** understanding. Compared to the conventional offline setting (top right), which answers questions based on a fixed-length video of the scene, the bottom section illustrates our online setting: for the same question, the agent's answers evolve as it explores the scene, changing from blue (t1) to red (t2) to green (t3), reflecting its continuously updated understanding.

Abstract

Recent advances in multimodal large language models (MLLMs) have shown remarkable capabilities in integrating vision and language for complex reasoning. While most existing benchmarks evaluate models under offline settings with a fixed set of pre-recorded inputs, we introduce OST-Bench, a benchmark designed to evaluate Online Spatio-Temporal understanding from the perspective of an agent actively exploring a scene. The "Online" aspect emphasizes the need to process and reason over incrementally acquired observations, while the "Spatio-Temporal" component requires integrating current visual inputs with historical memory to support dynamic spatial reasoning. OST-Bench better reflects the challenges of real-world embodied perception. Built on an efficient data collection pipeline, OST-Bench consists of 1.4k scenes and 10k question-answer pairs collected from ScanNet, Matterport3D, and ARKitScenes. We evaluate several leading MLLMs on OST-Bench and observe that they fall short on tasks requiring complex spatio-temporal reasoning. Under the online setting, their accuracy declines as the exploration horizon extends and the memory grows. Through further experimental analysis, we identify common error patterns across models and find that both complex clue-based spatial reasoning demands and long-term memory retrieval requirements significantly drop model performance along two separate axes, highlighting the

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

core challenges that must be addressed to improve online embodied reasoning. To foster further research and development in the field, our codes, dataset, and benchmark are available at `https://github.com/InternRobotics/OST-Bench`.

1 Introduction

In the real world, humans continuously perceive and update their understanding of the environment through sequential visual observations. At every moment, we are aware of our spatial state and how it evolves with respect to surrounding objects and scenes. We expect embodied agents to possess similar online scene understanding capabilities. For instance(Fig. 1), in an embodied navigation task[6, 52, 12, 28, 44], an agent should be able to incrementally construct a representation of its surroundings (*"I have seen a brown pillow on the bed in the bedroom."*), track its current status (*"I am now in the living room next to the bedroom, facing south."*), and reason about dynamic spatial relationships (*"The brown pillow is now on my rear left."*). Such awareness enables the agent to instantly respond to commands (*"Go and get the brown pillow."*) and take correct actions.

Recent advances in multimodal large language models (MLLMs)[9, 18, 30, 39, 29, 49, 26] have shown remarkable capabilities in integrating vision and language for complex reasoning. However, most existing benchmarks [15, 31, 32, 8, 24, 5], evaluate models under **offline** settings, where reasoning is performed over a fixed set of pre-recorded inputs, such as reconstructed 3D scenes or images and videos, this does not capture the online nature of embodied tasks.

To address this gap, we introduce **OST-Bench**, a benchmark designed to evaluate **Online Spatio-Temporal understanding** from the perspective of an agent actively exploring a scene. The term *Online* emphasizes the agent's need to perceive, remember, and reason over incrementally received observations, rather than complete, pre-recorded scene data. The term *Spatio-Temporal* highlights the need to integrate current visual observation with historical memory to support dynamic spatial reasoning. To more accurately simulates real-world embodied perception, OST-Bench defines tasks from these perspectives: the **agent**, its surrounding **environment**, and their **relationship**, contains three main categories(Fig.2): (1) **Agent State**: the agent's understanding of its own state, (2) **Agent Visible Info**: the agent's dynamic interpretation of visible scene information, and (3)**Agent-object Spatial Relationship**: the agent's dynamic understanding of spatial relationships with objects, all posed in an online, temporally grounded fashion. OST-Bench comprises 1.4k real-world scenes sourced from the test and validation splits of ScanNet[20], Matterport3D[14], and ARKitScenes[11], accompanied by 10k QA pairs covering a diverse range of subtypes. Our benchmark provides a rigorous testbed for assessing the online spatio-temporal reasoning ability of MLLMs in realistic, embodied settings.

We evaluate leading MLLMs on OST-Bench and find its online spatio-temporal nature poses significant challenges for the models, even the most advanced models lag behind human performance by over 30%. Models perform poorly on tasks requiring complex spatio-temporal reasoning, with accuracy declining as exploration steps increase and memory grows under the online setting. Based on an in-depth experimental analysis, we observe a phenomenon which we term *Spatio-temporal Reasoning Shortcut*-when reasoning over long-term memory, models tend to avoid retrieving key information, instead taking shortcuts and relying on shallow, unsupported inferences; further, we design four tasks with different levels of difficulty to better delineate the models' capability limits, along both the spatial dimension (from single- to multi-step spatial reasoning) and the temporal dimension (from keyframe- to sequence-baesd context), and observe a clear performance drop on both dimensions. This reveals that both complex clue-based spatial reasoning and long-term memory retrieval are two distinct weaknesses that hinder the model's performance on OST-Bench, highlighting the core challenges that must be addressed to advance online embodied reasoning. Moreover, our fine-tuning analysis shows that data-driven training alone yields only limited improvement, suggesting that further progress will likely require advances in model architecture and training methodology rather than sheer data scaling.

2 Related Work

Spatial Reasoning Benchmarks. Early scene understanding benchmarks[8, 32, 24, 31, 50, 54] introduced diverse task taxonomies to comprehensively evaluate various aspects of visual scene interpretation, with spatial understanding consistently recognized as the most fundamental component.

Dataset	Input Modality	Settings	Spatio-Temporal Awareness	Output Format	
				Text	Num.
ScanQA [8]	Video/PC.	Offline	✗	✓	✗
SQA3D [32]	Video/PC.	Offline	✗	✓	✗
SceneVerse [24]	Video/PC.	Offline	✗	✓	✗
MMScan [31]	Video/PC.	Offline	✗	✓	✗
SpatialRGPT-Bench [19]	Image	Offline	✗	✓	✓
CV-Bench [38]	Image	Offline	✗	✓	✓
VSI [46]	Video	Offline	✗	✓	✓
OST-Bench	Video	Online	✓	✓	✓

Table 1: **Comparison with other spatial reasoning datasets.** "PC." abbrev for "Point cloud". "Text" and "Num." represent whether the output is a string or a numerical value. Compared to other benchmarks, ours is clearly distinguished by its focus on the online setting and the requirement for spatio-temporal awareness in models.

Benchmarks such as ScanQA[8], SQA3D[32], SceneVerse[24], and MMScan[31] emphasized semantic understanding and incorporated object locations and spatial relations, they largely treated spatial relationships as semantic attributes, focusing primarily on complex spatial semantics rather than explicitly targeting spatial reasoning. With the rapid advancement of Multimodal Large Language Models (MLLMs), recent benchmarks have begun to place greater emphasis on spatial reasoning evaluation, SpatialR-GPT[19] and CV-Bench[38] require models to reason about 3D information, such as depth and distance from a single image, VSI[46] proposed a finer-grained categorization of spatial reasoning tasks, systematically evaluating models' ability to infer 3D scene layouts from 2D video inputs, covering both relative and absolute spatial relationships. Existing spatial reasoning benchmarks predominantly operate in an offline setting, focusing on static scenes and requiring models to perform reasoning over a fixed set of images or videos of predefined length. In contrast, our OST-Bench adopts an online setting, emphasizing dynamic scene understanding from an agent-centric perspective, and offers an alternative perspective for evaluating spatial reasoning capabilities. It includes a wider range of complex question types to assess more diverse and fine-grained spatio-temporal reasoning abilities.

Video Benchmarks for Temporal Understanding. Video benchmarks for temporal understanding require models to reason over both temporal and visual dimensions. Early efforts in video temporal understanding primarily focused on semantic comprehension from a third-person perspective[45, 47, 22, 43], mostly without considering 3D spatial perception. More recent benchmarks, driven by embodied task settings, have introduced characteristics such as: (1) egocentric perspective[23, 33, 13], where tasks are presented from a first-person viewpoint, (2) online inference[48, 16, 27, 16], requiring online processing of continuously streaming video input, and (3) spatial understanding[33, 27, 13], which evaluates models' awareness of spatial elements. However, spatial tasks in these benchmarks are often limited to 2D relationships or short-term motion cues, reflecting more of a content-level understanding rather than deeper spatial reasoning, lacking complex 3D spatial reasoning that requires integrating multi-view 2D observations into a coherent 3D representation. In contrast, OST-Bench is an egocentric, online temporal video benchmark that uniquely emphasizes 3D spatial reasoning, a core ability for real-world embodied tasks such as navigation and exploration.

3 OST-Bench

In this section, we present our comprehensive methodology for establishing OST-Bench, which comprises three core components: task formulation, the data collection and processing pipeline, and benchmark sample generation.

3.1 Task Formulation

Before introducing OST-Bench, we clarify the assumptions underlying our formulation of scene understanding. (1) While existing real-world datasets predominantly feature static scenes, we specifically focus on static environments in our current benchmark design, meaning the positions and states of objects remain unchanged during exploration; the agent is the only dynamic element. (2) There is no defined absolute coordinate system in the scene, so all spatial references are defined relative to an anchor such as an object, viewpoint, or the agent itself. As a result, the position of any object or agent cannot be defined in isolation. All spatial measurements fall into four categories: relative distance, absolute distance, relative direction, and absolute direction.

A static scene consists of a set of immobile objects, and understanding such a scene involves reasoning about individual entities and their relationships[31], such as the **object attribute**(intrinsic properties of individual objects, including category, color, material, shape, size, and function), **the attribute / spatial relationship between objects**, and **the spatial relationship between objects and a given viewpoint** (provided either textually or via a virtual camera input). When a dynamic agent is introduced, it introduces new relational dynamics and opens up additional avenues for investigation. These can be categorized into three main categories that form the core focus of our benchmark evaluation: (1) **Agent state**: The position and orientation of the agent, which continuously change as the agent explores. (2) **Agent visible info**: The perceptual information available from the agent's point of view at a given moment includes the existence of visible objects, their count, diversity, and the timing of their appearance. The information visible to the agent is continuously updated as the agent explores the scene. (3) **Agent-object spatial relationship**: 3D spatial relations between the agent and objects, described by relative or absolute distance/direction, constantly change as the agent explores.

3.2 Meta-dataset Collection and Processing

Base Dataset Acquisition. The three real scene datasets, ScanNet[20], ARKitScenes[11] and Matterport3D[14], contain rich scene information along with RGB-D videos/images and their corresponding camera information, totaling 7.6k scenes. Building on this foundation, EmbodiedScan[41] provided a large number of high-quality 9-DOF bounding box annotations for the objects in these scenes. MMScan[31] further enriched these scenes with a large number of highly quality, manually annotated object- and region-level semantic annotations. We selected a total of 1.4k scenes from the validation/test splits of these three datasets and constructed our dataset based on the annotations from EmbodiedScan and MMScan.

Exploration Route Generation. To construct an agent-centric exploration dataset, we require first-person videos of environments accompanied by camera parameters. While ScanNet and ARKitScenes provide such first-person videos along with camera pose data suitable for modeling agent trajectories, Matterport3D offers only multi-view images without continuous exploration paths. To address this limitation, we generate synthetic exploration trajectories within Matterport3D by constructing a graph of camera viewpoints and applying the minimum-spanning tree algorithm[36]. This ensures coherent movement and obstacle-free transitions between connected nodes. To maintain observation continuity, we enforce an image-overlap threshold between adjacent viewpoints. This approach enables us to simulate first-person exploration videos for Matterport3D scenes, complete with associated camera parameters.

Visible Information Processing. OST-Bench requires fine-grained visibility annotations at the frame level, which we define in two forms: attribute visibility and spatial visibility. Attribute visibility refers to the ability to determine the existence of an object based on a single frame. Even if an object is partially visible in a frame, as long as its visible portion is sufficiently large, you can infer attributes such as the object's type or color. Spatial visibility is used to generate questions in the OST-Bench that are related to the object's 3D spatial information. Therefore, for spatially visible objects, in addition to being attribute visible, we require that their center position, size, shape, and other spatial information can be inferred from observation. In practice, the attribute and spatial visibility of an object are determined by thresholding the projected area of its point cloud and the visibility of the vertices of its 9-DoF bounding box. Additional implementation details are provided in Appendix A.2.

3.3 Benchmark Samples Generation

Rule-based Generation. OST-Bench is designed in a multi-round dialogue format. In each round, the model receives a sequence of newly observed, temporally ordered frames, which are appended to all previously seen frames to simulate a streaming video input. At the end of the round, a new question is posed based on the accumulated observations. As the dialogue progresses, the input sequence grows incrementally, requiring the model to perform reasoning over an expanding spatio-temporal context. All questions are framed from an online perspective, grounded in the agent's current situation.

Our questions span three major categories: *Agent State, Agent Visible Info,* and *Agent–Object Spatial Relationships*. As illustrated in Fig.2, each main category contains multiple subtypes. Across all

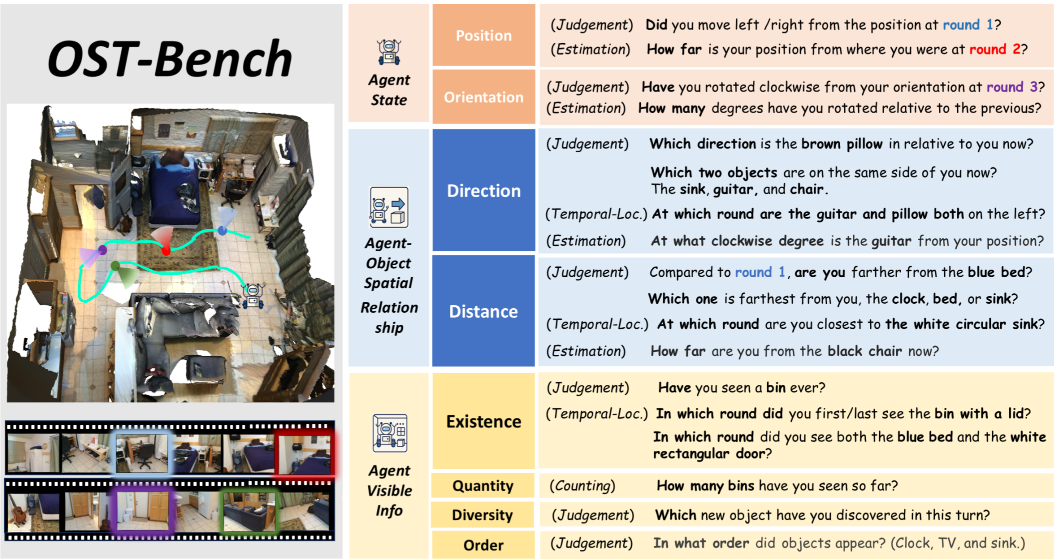

OST-Bench
Agent State
Position
(Judgement) Did you move left /right from the position at round 1?
(Estimation) How far is your position from where you were at round 2?
Orientation
(Judgement) Have you rotated clockwise from your orientation at round 3?
(Estimation) How many degrees have you rotated relative to the previous?
Agent-Object Spatial Relationship
Direction
(Judgement) Which direction is the brown pillow in relative to you now?
Which two objects are on the same side of you now? The sink, guitar, and chair.
(Temporal-Loc.) At which round are the guitar and pillow both on the left?
(Estimation) At what clockwise degree is the guitar from your position?
Distance
(Judgement) Compared to round 1, are you farther from the blue bed?
Which one is farthest from you, the clock, bed, or sink?
(Temporal-Loc.) At which round are you closest to the white circular sink?
(Estimation) How far are you from the black chair now?
Agent Visible Info
Existence
(Judgement) Have you seen a bin ever?
(Temporal-Loc.) In which round did you first/last see the bin with a lid?
In which round did you see both the blue bed and the white rectangular door?
Quantity
(Counting) How many bins have you seen so far?
Diversity
(Judgement) Which new object have you discovered in this turn?
Order
(Judgement) In what order did objects appear? (Clock, TV, and sink.)

Figure 2: OST-Bench categorizes questions into three main categories. Each main category includes several subtypes; in total, the benchmark comprises 15 fine-grained question subtypes.

categories, the questions fall into four general formats: *Judgment*, *Counting*, *Temporal Localization*, and *Estimation*. Judgment questions evaluate the model's qualitative understanding of facts—whether something is true or not, or whether something has occurred; Counting questions assess the model's ability to quantitatively enumerate information; Temporal Localization questions test the model's ability to locate events along the time axis. (We use the round index as a discrete timestamp in OST-Bench); Estimation questions evaluate the model's ability to approximate measurable quantities, such as physical distances or angular differences.

Based on the processed meta-datasets, we define dedicated rule-based generation templates to construct corresponding data samples for each subtype within the three main categories. Detailed generation procedures for each fine-grained subtype, including rule definitions and templates used, are provided in Appendix A.3. Several representative samples of these subtypes are illustrated in Fig.2. Our benchmark comprises approximately 1.4k test and validation scenes selected from ScanNet, Matterport3D, and ARKitScenes. For each scene, we generate a single agent exploration trajectory. Along each trajectory, multiple dialogue rounds are defined, each containing a single question, resulting in a total of 10k questions across the dataset.

Data Quality. Ensuring high-quality benchmark data is crucial. Based on the high-quality manual annotations from Embodiedscan and MMScan, we design and iteratively refine tailored rule-based generation strategies for each subtask to ensure semantic validity, robustness, and clarity, avoiding common corner cases and ambiguities. To assess dataset quality, we employ a rigorous validation protocol in which questions are randomly sampled for manual review. Samples lacking sufficient information or containing incorrect answers are marked as invalid. Human evaluation results confirm that the dataset meets our strict quality standards, with an error rate below 5%, thereby ensuring a reliable and high-quality benchmark.

4 Experiments

4.1 Benchmark Models & Evaluation Metrics

We evaluate the performance of multiple multi-modal large language models (MLLMs), including both proprietary models (Claude-3.5-Sonnet[7], GPT-4o[34], GPT-4.1[35], Gemini-2.0-Flash[37], and its thinking variant) and open-source models (InternVL-2.5[17], QwenVL-2.5[10], LLaVA-Onevision[25], and LLaVA-Video[51] of different scales). Each model is tested in a zero-shot setting and conducts inference in a multi-turn dialogue format. (In addition to these general-purpose VLMs, we also include several models specifically designed with explicit spatial grounding or memory mechanisms to varying degrees, their results are reported in Appendix C.1.) To establish performance boundaries, we include two baselines: a human baseline and a chance-level baseline. For the

Methods	**Agent State**				**Agent Visible Info**					**Agent-object Spatial Relationship**					
	Position		Orientation		Existence		Quantity	Diversity	Order	Direction			Distance		
	JUD.	EST.	JUD.	EST.	JUD.	TEMP.	CNT.	JUD.	JUD.	JUD.	TEMP.	EST.	JUD.	TEMP.	EST.
Proprietary															
Claude-3.5-Sonnet	**65.1**	**36.2**	50.6	30.3	85.8	67.0	57.9	57.1	60.0	39.2	18.3	21.9	43.8	54.9	19.0
Gemimi-2.0-Flash	59.7	27.3	56.7	**36.5**	89.6	70.8	59.4	78.7	55.6	42.0	17.8	21.5	45.5	48.9	**27.1**
Gemimi-2.0-Flash(Thinking)	57.4	36.3	**61.1**	33.4	88.1	74.8	**61.9**	63.6	73.1	50.9	**51.7**	23.3	**51.1**	**56.8**	22.7
GPT-4o	55.6	20.5	45.6	33.6	90.6	75.6	59.8	78.2	59.6	46.1	19.5	21.4	43.1	50.6	20.4
GPT-4.1	64.2	30.7	60.8	33.2	**90.8**	**78.0**	60.6	**82.1**	70.8	**51.5**	23.4	**28.6**	44.9	53.6	23.9
Open-source															
InternVL-2.5-8B	51.7	26.1	49.4	40.4	86.3	51.3	56.4	60.7	38.4	37.2	**33.8**	22.8	43.0	42.9	27.9
InternVL-2.5-38B	56.7	31.8	**54.6**	38.4	**91.7**	**74.7**	61.1	**79.8**	**62.1**	42.1	20.6	27.7	42.7	42.5	28.1
InternVL-2.5-78B	**60.8**	**34.4**	49.9	40.7	90.7	74.4	65.9	77.9	61.2	**43.4**	22.4	17.8	46.7	44.4	22.9
QwenVL-2.5-7B	49.8	19.3	51.8	40.8	78.6	37.3	62.1	56.3	28.5	41.0	28.9	12.2	44.9	43.6	18.6
QwenVL-2.5-32B	51.0	31.1	53.5	39.4	85.3	64.8	59.2	73.4	41.8	39.5	24.9	25.7	43.6	39.1	20.3
QwenVL-2.5-72B	57.0	27.6	52.2	37.1	86.1	64.5	61.5	75.7	34.5	41.4	21.1	8.2	44.5	39.3	18.7
LLaVA-Video-7B	50.4	25.4	46.1	12.1	90.4	32.3	63.1	66.5	39.3	35.4	27.3	16.2	41.3	41.8	10.8
LLaVA-Video-72B	51.0	18.0	49.2	**41.6**	88.0	38.8	51.0	70.9	53.7	35.5	27.7	30.9	43.8	46.2	26.3
LLaVA-Onevision-7B	53.8	11.6	51.2	7.7	90.0	34.8	**66.9**	51.1	33.4	35.7	27.0	**38.1**	43.5	35.6	21.9
LLaVA-Onevision-72B	53.8	13.9	51.6	36.2	89.0	41.8	45.8	74.8	56.6	37.8	28.9	27.3	**48.2**	**47.0**	**28.2**
Baseline															
Human-Level	93.2	58.9	92.8	54.4	95.7	94.7	91.3	94.4	90.9	90.5	93.3	54.3	93.4	94.5	60.1
Chance-Level	50.0	37.8	50.0	39.3	50.0	29.1	25.0	33.0	25.0	36.0	33.2	47.6	36.0	31.2	30.3

Table 2: **Full evaluation results of OST-Bench.** This table reports the performance of each model across all fine-grained question subtypes, "JUD."/ "CNT." / "TEMP." / "EST." abbrev for "judgement","counting","temporal-localization", and "estimation".

Methods	Avg	A. State	A. Info	AO.	JUD.	TEMP.	CNT.	EST.
Proprietary								
Claude-3.5-Sonnet	47.8	45.6	65.6	32.9	57.4	46.7	57.9	26.9
Gemimi-2.0-Flash	49.5	45.1	70.8	33.8	61.1	45.8	59.4	28.1
Gemimi-2.0-Flash(Thinking)	54.2	47.1	72.3	**42.8**	63.6	**61.1**	**61.9**	28.9
GPT-4o	48.7	38.8	72.8	33.5	59.8	48.6	59.8	24.0
GPT-4.1	53.4	**47.2**	**76.5**	37.7	**66.4**	51.7	60.6	**29.1**
Open-source								
InternVL-2.5-8B	44.6	41.9	58.6	34.6	52.4	42.7	56.4	29.3
InternVL-2.5-38B	50.8	45.4	73.9	34.0	61.4	45.9	61.1	31.5
InternVL-2.5-78B	51.1	**46.5**	**74.0**	32.9	**61.5**	**47.1**	65.9	29.0
QwenVL-2.5-7B	41.2	40.4	52.6	31.5	50.1	36.6	62.1	22.7
QwenVL-2.5-32B	46.9	43.8	64.9	32.2	55.4	42.9	59.2	29.1
QwenVL-2.5-72B	45.6	43.5	64.5	28.9	55.9	41.6	61.5	22.9
LLaVA-Video-7B	39.3	33.5	58.3	28.8	52.8	33.8	63.1	16.1
LLaVA-Video-72B	43.2	40.0	60.5	**35.1**	56.0	37.6	51.0	29.2
LLaVA-Onevision-7B	40.4	31.1	55.2	33.6	51.2	32.5	66.9	19.8
LLaVA-Onevision-72B	43.4	38.9	61.6	36.2	58.8	39.2	45.8	26.4
Baseline								
Human Level	83.5	74.8	93.4	81.0	93.0	94.2	91.3	56.9
Chance Level	36.9	44.3	32.4	35.7	40.0	31.2	25.0	38.8

Table 3: **Model performance across main categories and question formats.** "A." abbrev for "Agent" and "AO." abbrev for "Agent-Object Spatial Relationship". The open-source and proprietary models with the highest and second-highest overall average scores are highlighted with bright green and light green marks.

human baseline, we ensure that participants have no prior exposure to the test scenes. The chance-level method adopts a random selection approach, randomly picks one answer from all possible choices for Judgement/Counting/Temporal-Localization questions, and for Estimation questions, it always outputs the mean value calculated from all potential numeric. As for the evaluation metrics, Judgement questions are considered correct if the model selects the same option as the ground truth. For Counting and Temporal Localization questions, the model's output—whether a number or a turn index—must exactly match the ground truth to be deemed correct. For Estimation questions, we adopt the Mean Relative Accuracy (MRA) metric from VSI[46] to score the similarity between the model's floating-point output and the ground truth.

4.2 Main Results

We report the performance of various models on our benchmark. Tab. 2 presents the performance of each model in all different subtype. Tab.3 summarizes the models' overall performance, including their overall average scores, average scores for each of the three main categories, and performance across different question formats. Additional model results and further analysis are provided in Appendix C.1.

Substantial Gap between MLLMs' and Human's Performance. Model accuracy lags significantly behind human performance, with consistent gaps across all question types as shown in Tab.2. According to Tab.3, even the most advanced models lag behind human performance by nearly 30% in the overall average score. This performance gap remains substantial across three main task categories and all question formats. These findings suggest that current MLLMs fall short on OST-Bench, illustrating how our benchmark presents a novel challenge, demanding stronger online spatio-temporal perception and reasoning capabilities. This observation motivates us to investigate further the reasons for the subpar performance of the models on this benchmark.

Weak Spatio-Temporal Reasoning in MLLMs. As shown in Tab. 2 and 3, a striking contrast can be observed across the three main task categories. Although most models achieve average scores close to 70% in the Agent Visible Info category, with performance in each subtype significantly above chance level, their scores in the Agent State and Agent-Object Spatial Relationship categories remain near chance level across all subtypes. This suggests that current models are capable of dynamically perceiving scene information with temporal awareness, but lack the ability to perform complex spatio-temporal reasoning.

Performance Drop During Exploration. As illustrated in Fig. 3, we observe a significant decline in model accuracy as the agent continues to explore with an increasing number of sequential observations in the online setting. This is expected: for each question, the agent must reason based on both the current observation and its historical memory. As the number of exploration turns grows, the amount of relevant past information the agent needs to retain also increases, naturally raising the difficulty of both perception and reasoning. We further analyze how performance evolves for two representative models, InternVL-2.5-38B and GPT-4.1, across the three main question categories. For Agent Visible Info questions, accuracy declines gradually and consistently over turns. In contrast, for Agent-Object Spatial Relationship and Agent State questions, performance drops sharply within the first few steps (typically within 2 to 4 turns) to near chance level, and remains low in subsequent turns.

Comparison of Different Models. When comparing the performance of different models in Tab. 2 and 3, we find that proprietary models demonstrate significantly stronger performance compared to open-source ones. For different variants of the same open-source model, scaling from smaller configurations (7B/8B) to larger ones (>32B) consistently leads to notable performance gains, particularly on the questions under the Agent visible info category; Enabling the "thinking" mode in Gemini-2.0-Flash results in substantial improvements over the original version, especially on the questions with Temporal Localization format and the those under the Agent-Object Spatial Relationship category. This suggests that the thinking mode effectively enhances both spatial and temporal awareness.

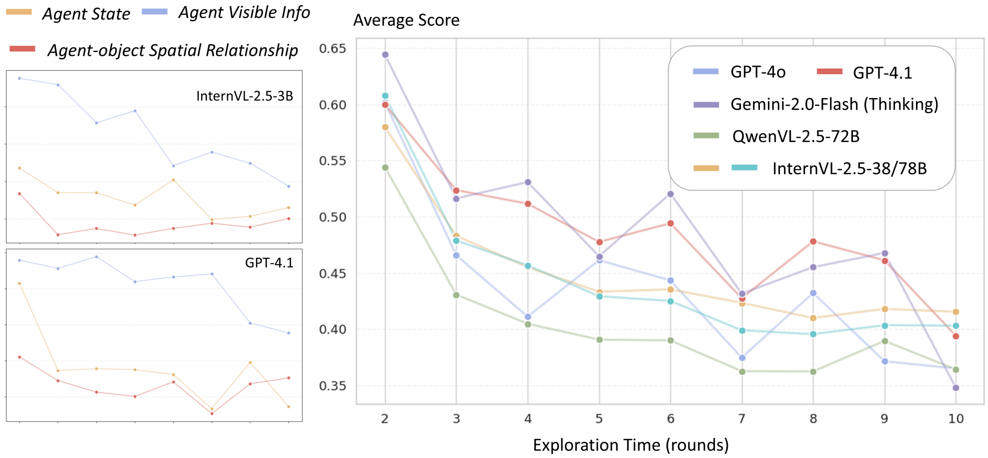


Figure 3: **Model performance over exploration time.** The right side shows a general decline in answer accuracy for all models; the left side illustrates the accuracy trends across three main categories for InternVL-2.5-38B and GPT-4.1.

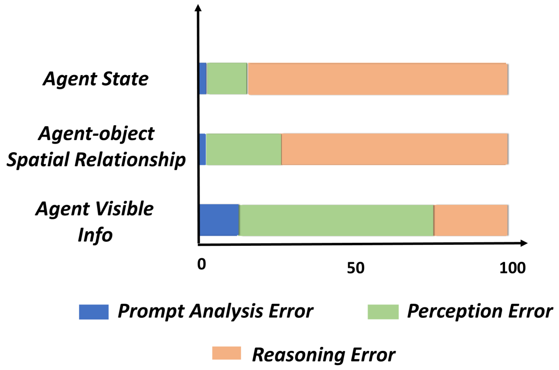
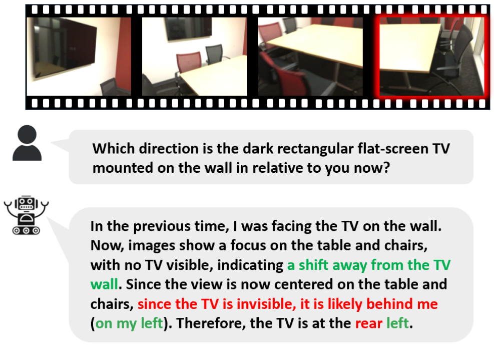


Figure 4: Distribution of three error types across the three task categories in OST-Bench.



Which direction is the dark rectangular flat-screen TV mounted on the wall in relative to you now?

In the previous time, I was facing the TV on the wall. Now, images show a focus on the table and chairs, with no TV visible, indicating a shift away from the TV wall. Since the view is now centered on the table and chairs, since the TV is invisible, it is likely behind me (on my left). Therefore, the TV is at the rear left.

Figure 5: An example of Spatio-temporal Reasoning Shortcut, the green text indicates correct reasoning by the model, while the red text highlights wrong reasoning.

4.3 Experiment Analysis

4.3.1 Insights from Model Explanations

To gain deeper insight into the weaknesses of models on OST-Bench, we prompt them to output not only their final answers but also their reasons. We manually examine the model outputs and identify the sources of errors. Since the model's inference process involves three key stages: understanding and following prompts, extracting information from observations, and performing spatio-temporal reasoning. Based on the stage at which the failure occurs, we categorize errors into three types: (1) **Prompt Analysis Error**, arising from the model's failure to correctly interpret the task setup or follow the given instructions; (2) **Perception Error**, where the model fails to accurately extract information from the visual observations by overlooking or misidentifying objects; (3) **Reasoning Error**, caused by incorrect spatio-temporal reasoning based on the information perceived. These three error types exhibit a clear progressive relationship. We select several representative open-source/proprietary models (GPT-4o, Gemini-2.0-Flash-Thinking, InternVL-2.5-78B) and examine 30 error cases for each major category per model, totaling 270 manual in-depth inspections.

Error Distribution Statistics on OST-Bench. The statistical results in Fig.4 show that Prompt Analysis Errors are relatively rare across all three major task categories, indicating that models generally understand the novel tasks and instructions introduced by OST-Bench. Perception Errors are the dominant failure mode for the Agent Visible Info category. In contrast, for tasks requiring more complex spatio-temporal reasoning, such as Agent–Object Spatial Relationships and Agent State, Reasoning Errors constitute a substantial portion of the failures. Based on the number of errors per task category and their distribution across the three error types, we estimate that Reasoning Errors account for over 60% of all errors, making them the primary bottleneck limiting current MLLM performance on OST-Bench.

Spatio-temporal Reasoning Shortcut of MLLMs. OST-Bench requires models to reason online over space and time, leveraging past observations to build spatial connections between the current state and prior states or previously seen objects. Within our in-depth error analysis, the model's Reasoning Error reflects a lack of this ability and reveals a common phenomenon as follows: The model tends to take shortcuts in reasoning, performing shallow and unsupported inference based on minimal information, and is reluctant to retrieve and utilize key information from long-term memory that could aid in answering the question. We name this phenomenon as *Spatio-temporal Reasoning Shortcut*. As shown in Fig. 5 example, the model correctly identifies that a television appeared in earlier frames and recognizes its own positional change over time. However, it makes an unfounded inference that the TV must now be behind it based solely on the fact that the TV is currently not visible, without using available spatial anchors such as the locations of a table or chair that could help establish a grounded reference frame. Additional examples of such shortcut behaviors are provided in Appendix C.2 to further illustrate their prevalence.

4.3.2 Cross-View Analysis

While most models struggle with complex spatio-temporal reasoning over sequentially growing memory, we introduce a targeted subset of OST-Bench to better delineate the capability boundaries of

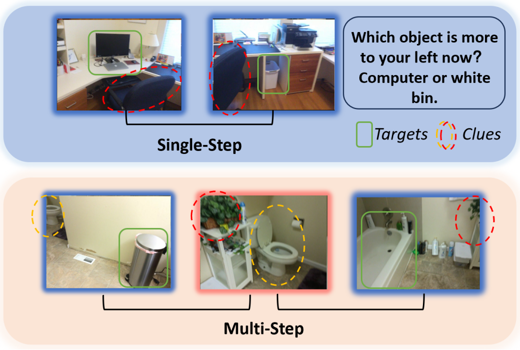

Which object is more to your left now? Computer or white bin.
Targets
Clues
Single-Step
Multi-Step

Figure 6: Single- vs. multi-step spatial connection settings. Target objects and spatial clues are highlighted.

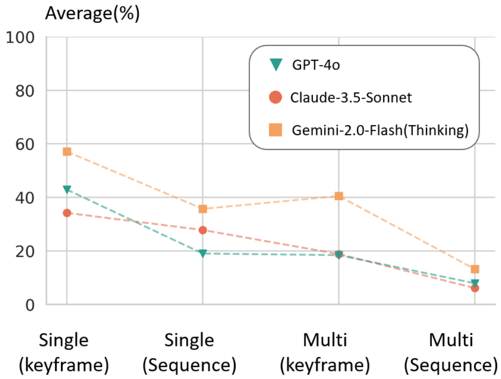


Figure 7: Model performance across four task settings: keyframe- vs. sequence-based context, and single- vs. multi-step spatial connection.

models. Questions of this subset focus on the spatial relationship between the agent and two objects that appear in different frames (e.g., *"Which object is more to your left?"*). It requires the model to construct cross-view spatial connections to answer correctly. It evaluates performance across two dimensions:

(1) **Single- vs. Multi-step Spatial Connection.** In the single-step setting, the spatial connection between two target objects can be directly inferred by analyzing the pair of the two frames that contain them. In contrast, the multi-step setting demands higher-level reasoning capabilities, where single-step pairwise frame analysis proves insufficient. This scenario requires the model to integrate spatial cues across multiple keyframes (typically more than two), iteratively establishing pairwise relationships between frames to enable chained reasoning through intermediate steps. As illustrated in Fig. 6, in the single-step case, the spatial connection between the computer and the white bin can be directly inferred through the shared objects (chair and table) in the single pair of images. While in the multi-step case, establishing the spatial connection between the bathtub and the gray trash bin necessitates an anchor image to bridge intermediate objects (trash bin $\rightarrow$ toilet $\rightarrow$ potted plant $\rightarrow$ bathtub), forming a multi-step spatial reasoning chain.

(2) **Keyframe- vs. Sequence-based Context.** In the keyframe-based setting, all keyframes that contain target objects or spatial cues sufficient to solve the problem are directly provided as input. In contrast, the sequence-based setting embeds these keyframes within a longer memory sequence that includes many irrelevant frames. The model must identify and leverage the relevant ones, thereby testing its capacity for long-term memory retrieval and reasoning.

This subset provides an opportunity to examine the model's performance across different levels of difficulty. We construct this dataset using a hybrid approach of rule-based generation and manual filtering. We curate 200 questions and evaluate three advanced MLLMs: Gemini-2.0-Flash (Thinking), GPT-4o, and Claude-3.5-Sonnet. For each question, models are required to provide both the answer and its reason. Only when both the final answer and the reason are correct is the response counted as correct. Based on our evaluation, as the results shown in Fig.7, we report the following key findings: (1) As tasks change from single to multi-step spatial connecting setting, which requires more complex reasoning, all models experience a substantial drop in accuracy; (2) Long-memory challenges further degrade performance. When models are required to locate relevant frames from sequence-based input rather than keyframe provided directly, accuracy drops significantly. In the most challenging tasks, which need to establish multi-step spatial connection in sequence-based context, all models fall to around 10% accuracy. The results show that the model's performance drops significantly when faced with either complex clue-based spatial reasoning requirements or long-term memory retrieval demands. OST-Bench exemplifies this dual challenge, as it requires models to retrieve information from a long, temporally extended memory while simultaneously constructing spatial relationships by integrating cues from multiple images to perform multi-step reasoning. These two factors jointly contribute to the poor performance observed on OST-Bench, highlighting the need to advance both capabilities in future model development.

4.3.3 Fine-tuning Analysis

To dig deeper into the upper bound of current models' capabilities and to better understand how much of the performance gap can be recovered through training with in-domain data, we conducted fine-tuning experiments on several representative models. Following a procedure similar to that used for constructing the benchmark samples, we generated training data from 7k training scenes across ScanNet, Matterport3D, and ARKitScenes, yielding a total of 50k annotated samples. All models were fine-tuned for a single epoch, and the results are summarized in Tab. 4. Overall, all evaluated models achieved performance gains exceeding 10% after fine-tuning. However, a closer analysis reveals several important insights:

Method	Setting	Overall	JUD.	EST.	CNT.	Temp-Loc.	A State	A Info	AO
QwenVL2.5-7B	Zero-Shot	41.2	50.1	22.7	62.1	36.6	40.4	52.6	31.5
	Fine-Tuned	54.0	59.0	41.2	74.6	50.2	48.3	69.8	43.5
InternVL2.5-8B	Zero-Shot	44.6	52.4	29.3	56.4	49.2	41.9	58.6	34.6
	Fine-Tuned	57.4	64.1	38.5	74.9	57.5	44.0	79.3	46.3
InternVL2.5-38B	Zero-Shot	50.8	61.4	31.5	61.1	45.9	45.4	73.9	34.0
	Fine-Tuned	60.2	68.4	44.1	73.1	56.1	50.8	81.7	47.5

Table 4: **Performance comparison of models under zero-shot and fine-tuned settings.** "A." abbrev for "Agent" and "AO." abbrev for "Agent-Object Spatial Relationship".

- Among the three major task categories, the largest improvements emerged in Agent Visible Info tasks — particularly for models with smaller parameter sizes. In contrast, the other two task categories, despite showing some gains, remained at or below 50% accuracy. This indicates that even with in-domain adaptation, models still struggle with tasks that demand complex spatio-temporal reasoning. Simple supervised fine-tuning on OST-Bench is therefore insufficient to resolve its core challenges.
- All four question formats benefited from fine-tuning, yet deeper inspection of predictions reveals more nuanced observations. Although fine-tuning improves the scores of Estimation (EST) and Judgement (JUD) tasks, closer inspection reveals that these gains do not reflect genuine reasoning improvements: models frequently output nearly identical values or default to the same option across samples, indicating reliance on dataset-specific shortcuts or memorization rather than true understanding. Moreover, their instruction-following ability degrades post-finetuning, with many responses failing to provide both the final answer and the required reasoning.

While fine-tuning significantly improves raw performance, a considerable gap remains compared to human-level accuracy. This highlights two key points: (1) data-only supervised fine-tuning is insufficient to solve the challenges posed by OST-Bench — improvements may also be required on the model architecture or training methodology side; and (2) the benchmark itself is both challenging and robust. Despite being constructed using templates, OST-Bench resists shortcut learning and cannot be easily exploited through superficial patterns in the training distribution.

5 Limitations and Conclusion

In this work, we propose OST-Bench, a novel benchmark for evaluating the online spatio-temporal reasoning capabilities of MLLMs. By emphasizing both online processing and spatio-temporal understanding, OST-Bench more accurately reflects the complexities of real-world perception and reasoning. Our extensive evaluation of leading MLLMs shows that OST-Bench poses significant challenges for models, particularly in tasks requiring complex spatio-temporal reasoning and maintaining answer accuracy as input accumulates over time in an online setting. We hope the public release of OST-Bench will serve as a catalyst for future research in online embodied understanding. We assume that the environment remains static. However, in real-world scenarios, object states and positions often change due to interactions with humans or agents. Additionally, our benchmark focuses solely on the agent's online perception and reasoning abilities, capturing only one aspect of real-world embodied tasks. Other crucial capabilities, such as interactive behaviors and active manipulation, are not considered in our current setting. These limitations highlight promising directions for future research and benchmark development.

6 Acknowledgement

This work is funded in part by the National Key R&D Program of China, and Shanghai Artificial Intelligence Laboratory.

References

[1] Arkitscenes license. `https://github.com/apple/ARKitScenes/blob/main/LICENSE`.

[2] Embodiedscan and mmscan access. `https://docs.google.com/forms/d/e/1FAIpQLScUXEDTksGiqHZp31j7Zp7zlCNV7p_08uViwP_Nbzfn3g6hhw/viewform`.

[3] Matterport3d license. `https://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf`.

[4] Scannet license. `https://kaldir.vc.in.tum.de/scannet/ScanNet_TOS.pdf`.

[5] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020.

[6] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir. On evaluation of embodied navigation agents, 2018.

[7] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024.

[8] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.

[9] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. Qwen technical report, 2023.

[10] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report, 2025.

[11] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, and E. Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[12] R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara. Out of the box: Embodied navigation in the real world. In N. Tsapatsoulis, A. Panayides, T. Theocharides, A. Lanitis, C. Pattichis, and M. Vento, editors, *Computer Analysis of Images and Patterns*, pages 47–57, Cham, 2021. Springer International Publishing.

[13] K. Chandrasegaran, A. Gupta, L. M. Hadzic, T. Kota, J. He, C. Eyzaguirre, Z. Durante, M. Li, J. Wu, and L. Fei-Fei. Hourvideo: 1-hour video-language understanding, 2024.

[14] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[15] D. Z. Chen, A. X. Chang, and M. Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[16] J. Chen, Z. Lv, S. Wu, K. Q. Lin, C. Song, D. Gao, J.-W. Liu, Z. Gao, D. Mao, and M. Z. Shou. Videollm-online: Online video large language model for streaming video, 2024.

[17] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, L. Gu, X. Wang, Q. Li, Y. Ren, Z. Chen, J. Luo, J. Wang, T. Jiang, B. Wang, C. He, B. Shi, X. Zhang, H. Lv, Y. Wang, W. Shao, P. Chu, Z. Tu, T. He, Z. Wu, H. Deng, J. Ge, K. Chen, K. Zhang, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.

[18] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024.

[19] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu. Spatialrgpt: Grounded spatial reasoning in vision language models, 2024.

[20] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[21] Z. Fan, J. Zhang, R. Li, J. Zhang, R. Chen, H. Hu, K. Wang, H. Qu, D. Wang, Z. Yan, H. Xu, J. Theiss, T. Chen, J. Li, Z. Tu, Z. Wang, and R. Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction, 2025.

[22] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, R. Ji, and X. Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024.

[23] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022.

[24] B. Jia, Y. Chen, H. Yu, Y. Wang, X. Niu, T. Liu, Q. Li, and S. Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024.

[25] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer, 2024.

[26] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024.

[27] Y. Li, J. Niu, Z. Miao, C. Ge, Y. Zhou, Q. He, X. Dong, H. Duan, S. Ding, R. Qian, P. Zhang, Y. Zang, Y. Cao, C. He, and J. Wang. Ovo-bench: How far is your video-llms from real-world online video understanding?, 2025.

[28] J. Lin, H. Gao, X. Feng, R. Xu, C. Wang, M. Zhang, L. Guo, and S. Xu. The development of llms for embodied navigation. *CoRR*, abs/2311.00530, 2023.

[29] J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han. Vila: On pre-training for visual language models, 2024.

[30] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2024.

[31] R. Lyu, J. Lin, T. Wang, S. Yang, X. Mao, Y. Chen, R. Xu, H. Huang, C. Zhu, D. Lin, and J. Pang. Mmscan: A multi-modal 3d scene dataset with hierarchical grounded language annotations, 2025.

[32] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[33] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud, K. Yadav, Q. Li, B. Newman, M. Sharma, V. Berges, S. Zhang, P. Agrawal, Y. Bisk, D. Batra, M. Kalakrishnan, F. Meier, C. Paxton, A. Sax, and A. Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498, 2024.

[34] OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mądry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko,

A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O'Connell, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, and Y. Malkov. Gpt-4o system card, 2024.

[35] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli,

T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. Gpt-4 technical report, 2024.

[36] R. C. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957.

[37] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, S. Bohez, K. Bousmalis, A. Brohan, T. Buschmann, A. Byravan, S. Cabi, K. Caluwaerts, F. Casarini, O. Chang, J. E. Chen, X. Chen, H.-T. L. Chiang, K. Choromanski, D. D'Ambrosio, S. Dasari, T. Davchev, C. Devin, N. D. Palo, T. Ding, A. Dostmohamed, D. Driess, Y. Du, D. Dwibedi, M. Elabd, C. Fantacci, C. Fong, E. Frey, C. Fu, M. Giustina, K. Gopalakrishnan, L. Graesser, L. Hasenclever, N. Heess, B. Hernaez, A. Herzog, R. A. Hofer, J. Humplik, A. Iscen, M. G. Jacob, D. Jain, R. Julian, D. Kalashnikov, M. E. Karagozler, S. Karp, C. Kew, J. Kirkland, S. Kirmani, Y. Kuang, T. Lampe, A. Laurens, I. Leal, A. X. Lee, T.-W. E. Lee, J. Liang, Y. Lin, S. Maddineni, A. Majumdar, A. H. Michaely, R. Moreno, M. Neunert, F. Nori, C. Parada, E. Parisotto, P. Pastor, A. Pooley, K. Rao, K. Reymann, D. Sadigh, S. Saliceti, P. Sanketi, P. Sermanet, D. Shah, M. Sharma, K. Shea, C. Shu, V. Sindhwani, S. Singh, R. Soricut, J. T. Springenberg, R. Sterneck, R. Surdulescu, J. Tan, J. Tompson, V. Vanhoucke, J. Varley, G. Vesom, G. Vezzani, O. Vinyals, A. Wahid, S. Welker, P. Wohlhart, F. Xia, T. Xiao, A. Xie, J. Xie, P. Xu, S. Xu, Y. Xu, Z. Xu, Y. Yang, R. Yao, S. Yaroshenko, W. Yu, W. Yuan, J. Zhang, T. Zhang, A. Zhou, and Y. Zhou. Gemini robotics: Bringing ai into the physical world, 2025.

[38] S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, Z. Wang, R. Fergus, Y. LeCun, and S. Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024.

[39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[40] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer, 2025.

[41] T. Wang, X. Mao, C. Zhu, R. Xu, R. Lyu, P. Li, X. Chen, W. Zhang, K. Chen, T. Xue, X. Liu, C. Lu, D. Lin, and J. Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[42] D. Wu, F. Liu, Y.-H. Hung, and Y. Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence, 2025.

[43] H. Wu, D. Li, B. Chen, and J. Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024.

[44] Y. Wu, P. Zhang, M. Gu, J. Zheng, and X. Bai. Embodied navigation with multi-modal information: A survey from tasks to methodology. *Information Fusion*, 112:102532, 2024.

[45] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[46] J. Yang, S. Yang, A. W. Gupta, R. Han, L. Fei-Fei, and S. Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2024.

[47] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[48] H. Zhang, Y. Wang, Y. Tang, Y. Liu, J. Feng, J. Dai, and X. Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024.

[49] P. Zhang, K. Zhang, B. Li, G. Zeng, J. Yang, Y. Zhang, Z. Wang, H. Tan, C. Li, and Z. Liu. Long context transfer from language to vision, 2024.

[50] Y. Zhang, Z. Gong, and A. X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023.

[51] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li. Video instruction tuning with synthetic data, 2024.

[52] D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13624–13634, June 2024.

[53] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness, 2024.

[54] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction accurately and clearly reflect the contributions and scope of our paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: You can find our discussion of the limitations in Sec.5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer:[NA]

Justification: Our work does not contain any theoretical results; all conclusions are based on rigorous analysis of experimental results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the Sec.4, we clearly disclose all the information needed to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, our paper provides open access to the data and code in the abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: As our benchmark is intended solely for evaluation purposes, it includes only a test split. We clearly describe the evaluation settings for each model in Sec.3 and 4, with additional evaluation details provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: See in the Sec. 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide details about the computer resources required for the open-source models used in the experiments in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research fully adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both potential positive and negative societal impacts are discussed in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not involve the release of data or models that pose a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide detailed attributions of all external assets used in our work in the appendix, including their licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, our dataset introduces new assets that are well documented, with the documentation provided alongside the assets through the dataset link we provide.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve such experiments and therefore does not include related content.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve such experiments and therefore does not include related content.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Benchmark Details **23**
A.1 Exploration Route Generation 23
A.2 Visible Information Processing 24
A.3 Rule-based Generation 25
A.4 Statistics 28
A.5 Benchmark Examples 28

B Implementation Details **28**

C Experiment Analysis Details **29**
C.1 Spatially-Grounded Model Evaluation 29
C.2 More Findings in Tables 30
C.3 Cases of Three Error Types 30
C.4 Cases of Spatio-temporal Reasoning Shortcut 32
C.5 Subset Construction Process for Cross-View Analysis 32

D Inference Time of the Models **32**

E Social Impact **34**

F License and Acess **35**
F.1 License and Acess for Existing Assets 35
F.2 License and Acess for OST-Bench 35

A Benchmark Details

This section provides additional details on the construction of our benchmark, including the algorithm used for route generation, the method for determining object visibility, the rules for benchmark sample generation, and summary statistics of the generated data.

A.1 Exploration Route Generation

While ScanNet and ARKitScenes offer egocentric video sequences with associated per-frame camera parameters, Matterport3D provides, for each scene, n camera positions distributed throughout the environment. From each position, k images are captured at different viewing angles, as illustrated in Fig. 8. We aim to leverage this information to construct a simulated trajectory of an agent exploring the scene from a first-person perspective. As mentioned in the main paper, the trajectory must satisfy two key requirements: (a) **Path continuity**, the movement between adjacent frames should be smooth, avoiding abrupt spatial jumps over short time intervals. (b) **Observation continuity**, adjacent frames in the video must have a certain degree of visual overlap, which is crucial for providing the cross-frame visual continuity necessary for constructing a coherent 3D understanding of the scene. The videos provided by ScanNet and ARKitScenes naturally satisfy both of these requirements.

The video we aim to generate is a sequence of tuples $\{(n_i, k_i, c_i)\}$, where n_i denotes the camera position index among the n predefined locations, k_i indicates the viewing angle index among the k available viewing angles at that position, and c_i is the corresponding captured image. Based on the two aforementioned requirements(Fig.8), (a) We first construct a minimum spanning tree(MST) $T(N, E)$ over all camera positions using Prim's algorithm, where edge weights are defined by the

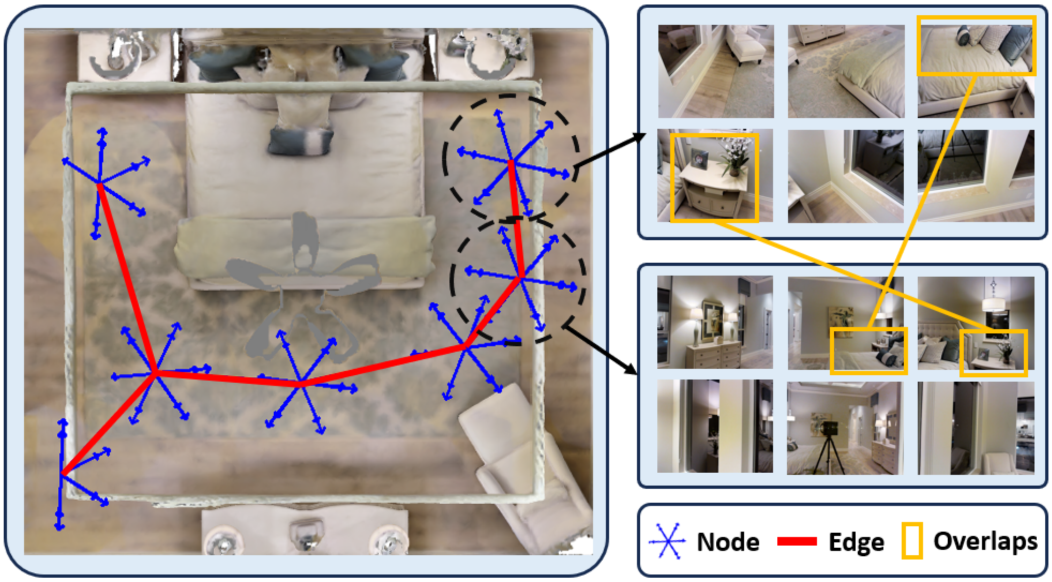


Figure 8: **Illustration of the route generation process.** The radial arrows represent multiple different viewing angles at a position, and the red edges denote connections generated by the MST algorithm. The right part shows the captured images for each viewing angle of two adjacent nodes. The agent can only move along the edges of the tree, and adjacent frames are required to have a certain amount of overlap.

Euclidean distances between positions. We constrain the agent's movement to either transitions between neighboring positions connected by edges in the MST ($E_{n_i,n_j} \in E$), or changes in viewing angles at the same position. This design ensures path continuity throughout the simulated trajectory. (b) We enforce that adjacent images in the sequence must have sufficient visual overlap. That is, for any $i \geq 0$, the overlap between images c_i, c_{i+1} must satisfy $Overlap(c_i, c_{i+1}) > threshold$. This constraint preserves observation continuity across frames. Based on these two rules, we perform a random walk over the nodes to generate the sequence. Starting from a randomly selected initial state with a random tuple (n_0, k_0, c_0), at each step, we randomly select a valid and previously unseen tuple representing the next state and append it to the sequence. This process continues until no valid tuples remain or the sequence reaches a predefined length.

It is important to note that the generated videos ensure continuity in terms of paths and observations, but do not guarantee temporal continuity (i.e., they only provide discrete frame ordering without information on the time intervals between frames). However, since our benchmark setting uses rounds as discrete timestamps, such temporal information is not required, and the provided data is sufficient for our purposes.

A.2 Visible Information Processing

Attribute Visibility. For the attribute visibility of objects, to reduce computational complexity, we first apply a necessary condition: if an object is visible, then at least one of its 3D points must be projectable onto the 2D image plane within the image boundaries and without occlusion. This condition allows for the rapid elimination of most invisible objects in each image. For objects satisfying this condition, we project the surface points of their bounding boxes onto the image's 2D plane. We first compute the projected area A_2 without considering occlusion or image boundaries. Then, we calculate the visible area A_1 by accounting for occlusions and restricting projections to within the image bounds. An object is deemed visible if either (1) the ratio of visible to total projected area, A_1/A_2, exceeds a predefined threshold, or (2) the absolute visible area A_1 is sufficiently large.

Spatial Visibility. For the spatial visibility of objects, building on attribute visibility, we further check whether at least five vertices of the object's 9-DoF bounding box are visible in the frames observed so far. If this condition is met, we assume the object's center position, size (length, width, height), and related spatial information are all available, thus satisfying the criteria for spatial visibility.

A.3 Rule-based Generation

Our OST-Bench comprises three major categories: *Agent State*, *Agent Visible Info*, and *Agent-object Spatial Relationship*. Within these categories, we define a total of 15 question subtypes. Data samples are generated through a rule-based approach, guided by a set of principles outlined below.

(1) **Multi-round Dialogue Format.** OST-Bench adopts a multi-round dialogue setup. In each round, 4–5 new frames from the video are selected sequentially in chronological order as new observations and appended to the historical observation sequence. Each question is asked at the timestamp of the last frame in the current round. All information after this timestamp is considered unavailable, and we ensure that the question is answerable based solely on the observations up to that timestamp.

(2) **Sample Pool Construction and Selection.** For each question subtype, we exhaustively generate all possible data samples to form a candidate pool. We ensure that no identical question-answer pair appears across different dialogue rounds (although the same question might occur, the answers must differ). In each round, we first randomly select a question subtype and then randomly select a data sample from its corresponding candidate pool as the question for that round.

(3) **Object Reference.** Object references in questions are divided into two types. The first is category-level reference, where a category word is used to refer to all instances of that category (e.g., *"How many books are there in the room?"*). The second is instance-level reference, where a specific grounded description is used to uniquely identify a single object.(e.g., *"Where is the yellow-covered book labeled with the word 'atomic'?"*). These descriptions are sourced from MMScan's object-level annotations. To eliminate ambiguity, we ensure that this referred object is the only instance of its category within historical observations.

(4) **Memory-based Reasoning Requirement.** To rigorously test a model's ability to reason over long-term memory and avoid overly simple questions, we ensure that no question can be answered using only the newly added observations in the current round. Each question requires integrating information from both the current and previous dialogue rounds. For example, we ensure that at least one relevant object is absent from the observations in the current round, thereby requiring the model to recall it from prior rounds.

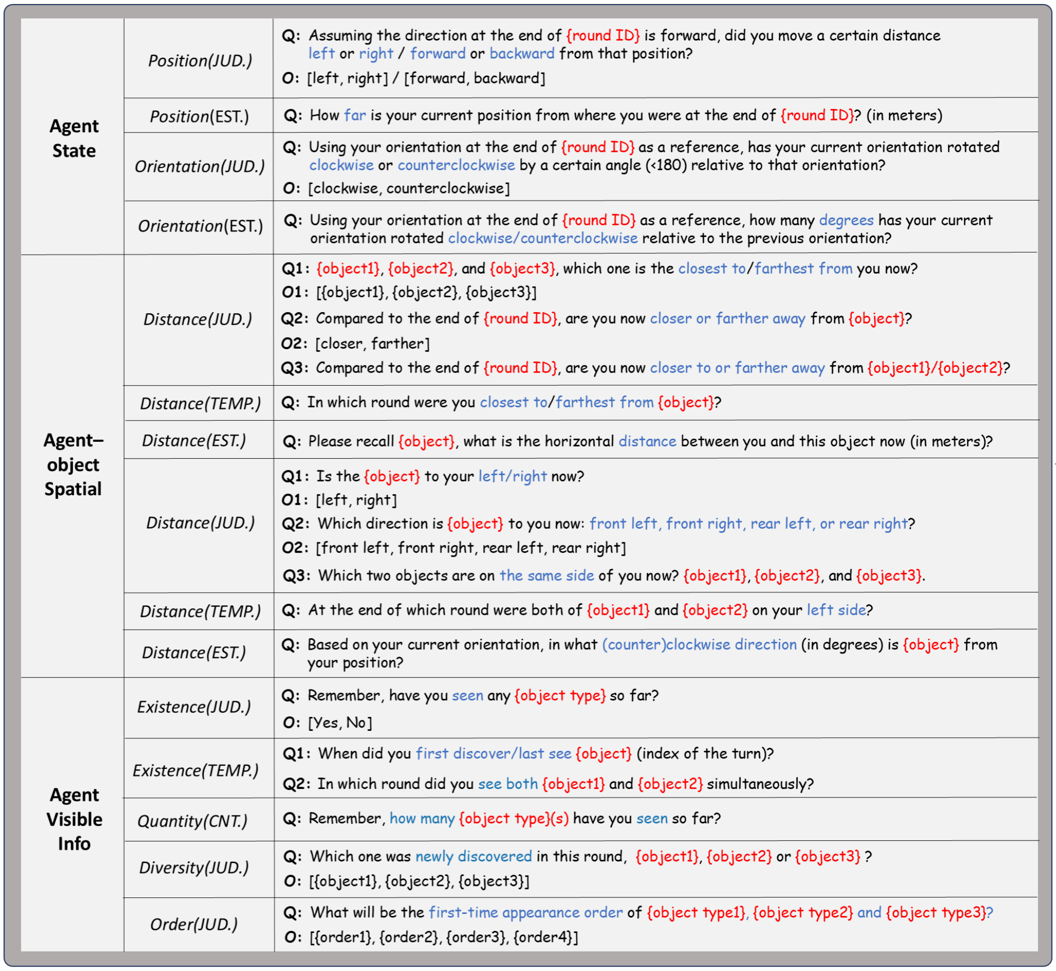
(5) **Ensuring Clarity and Avoiding Ambiguity.** To ensure the validity and clarity of the questions and to avoid controversial or ambiguous cases, we impose specific thresholds during sample generation so that the answers are unambiguous and clearly inferable. For example, when a question involves comparing two distances, we require the difference between the distances to exceed a predefined threshold to ensure a significant contrast. Similarly, for questions such as determining whether an object is on the left or right, we require the object to be clearly positioned on one side. Objects located near the decision boundary (e.g., close to the center) are excluded to prevent ambiguity in interpretation.

Fig.9 presents the predefined templates used for generating questions across different subtypes. The specific generation strategies for each subtype are detailed below:

Agent State. This category encompasses tasks that require the agent to judge or estimate its own spatial state, including its position and orientation. Since there is no globally defined coordinate system in OST-Bench, all measurements are made relative to a specific historical time point.

- *Position (Judgement)*: In this type, the task is to determine whether the agent has moved to the left or right (forward or backward), relative to its position and orientation at the end of a previous round T_1. The question is formulated as a binary choice, with the correct answer being either *left* or *right*(*forward* or *backward*). Let P_1 and O_1 denote the position and orientation at the end of round T_1, and P_2 denote the current position. We compute the parallel and perpendicular components of the vector $P_2 - P_1$ with respect to O_1. A question is generated only if the absolute value of either component exceeds a predefined threshold (1 meter). The correct answer is determined by the sign of the respective component: a positive value indicates forward or right, while a negative value indicates backward or left.

- *Position (Estimation)*: In this subtype, the task is to estimate how far the agent has moved from its position at the end of a previous round T_1. The ground-truth answer is defined as the Euclidean distance between the agent's current position and its position at T_1.



Agent State	*Position(JUD.)*	**Q:** Assuming the direction at the end of {round ID} is forward, did you move a certain distance left or right / forward or backward from that position? **O:** [left, right] / [forward, backward]
	Position(EST.)	**Q:** How far is your current position from where you were at the end of {round ID}? (in meters)
	Orientation(JUD.)	**Q:** Using your orientation at the end of {round ID} as a reference, has your current orientation rotated clockwise or counterclockwise by a certain angle (<180) relative to that orientation? **O:** [clockwise, counterclockwise]
	Orientation(EST.)	**Q:** Using your orientation at the end of {round ID} as a reference, how many degrees has your current orientation rotated clockwise/counterclockwise relative to the previous orientation?
Agent–object Spatial	*Distance(JUD.)*	**Q1:** {object1}, {object2}, and {object3}, which one is the closest to/farthest from you now? **O1:** [{object1}, {object2}, {object3}] **Q2:** Compared to the end of {round ID}, are you now closer or farther away from {object}? **O2:** [closer, farther] **Q3:** Compared to the end of {round ID}, are you now closer to or farther away from {object1}/{object2}?
	Distance(TEMP.)	**Q:** In which round were you closest to/farthest from {object}?
	Distance(EST.)	**Q:** Please recall {object}, what is the horizontal distance between you and this object now (in meters)?
	Distance(JUD.)	**Q1:** Is the {object} to your left/right now? **O1:** [left, right] **Q2:** Which direction is {object} to you now: front left, front right, rear left, or rear right? **O2:** [front left, front right, rear left, rear right] **Q3:** Which two objects are on the same side of you now? {object1}, {object2}, and {object3}.
	Distance(TEMP.)	**Q:** At the end of which round were both of {object1} and {object2} on your left side?
	Distance(EST.)	**Q:** Based on your current orientation, in what (counter)clockwise direction (in degrees) is {object} from your position?
Agent Visible Info	*Existence(JUD.)*	**Q:** Remember, have you seen any {object type} so far? **O:** [Yes, No]
	Existence(TEMP.)	**Q1:** When did you first discover/last see {object} (index of the turn)? **Q2:** In which round did you see both {object1} and {object2} simultaneously?
	Quantity(CNT.)	**Q:** Remember, how many {object type}(s) have you seen so far?
	Diversity(JUD.)	**Q:** Which one was newly discovered in this round, {object1}, {object2} or {object3} ? **O:** [{object1}, {object2}, {object3}]
	Order(JUD.)	**Q:** What will be the first-time appearance order of {object type1}, {object type2} and {object type3}? **O:** [{order1}, {order2}, {order3}, {order4}]

Figure 9: **Rule-based generation templates for all subtypes in OST-Bench.** Placeholders to be filled with specific content are marked in red, and question focal points are highlighted in blue. "JUD."/ "CNT." / "TEMP." / "EST." are abbreviations for "judgement","counting","temporal-localization", and "estimation"; "Q" and "O" denote "Question" and "Options"

- *Orientation (Judgement)*: This binary-choice question asks whether the agent has rotated clockwise or counterclockwise by an angle(less than 180 degrees) relative to its orientation at the end of round T_1. We compute the angle between the current orientation vector and the one at the end of T_1. To exclude ambiguous borderline cases, questions are generated only if the angle lies within the intervals $[\theta, 180 - \theta]$ or $[180 + \theta, 360 - \theta]$, where θ is a threshold used to exclude borderline cases. Angles within the first interval indicate clockwise rotation, while those within the second indicate counterclockwise rotation.

- *Orientation (Estimation)*: In this question type, the task is to estimate how many degrees the agent has rotated, clockwise or counterclockwise, relative to its orientation at the end of a previous round T_1. The answer is given as the angle between the current orientation and the orientation at the end of round T_1.

Agent Visible Info. All objects involved in this category of questions must satisfy the attribute visibility constraint, meaning that their existence must be identifiable from past observations. This category evaluates the model's understanding of agent visible information, including subtasks such as object existence, quantity, diversity, and the order of appearances.

- *Existence (Judgement)*: This type asks whether a certain category was visible in any of the previous observations. The answer is binary: *yes* or *no*. To balance positive and negative samples, we generate questions for object categories that do not appear in prior observations with a 50% probability.

- *Existence (Temporal Localization)*: This type includes two forms of queries: (1) Identifying the earliest/latest round in which a specific object was visible; (2) Identifying the round in which two specific objects were simultaneously visible. For both forms of queries, we ensure the answer is unique—i.e., there is exactly one round that satisfies the condition.
- *Quantity (Counting)*: This task requires counting how many objects of a specified category were visible in past observations. To avoid trivial cases, we exclude questions where the correct answer is one. Additionally, to balance the distribution, negative samples—where the target category does not appear at all—are introduced to constitute 25% of the total samples.
- *Diversity (Judgement)*: This question type asks which object is newly observed in the current round. The agent must choose one object from three candidates, all of which are visible in the current observation. Among them, only one has not appeared in any previous round, while the other two have been seen before.
- *Order (Judgement)*: This question type involves determining the appearance order of three different object categories. The agent must select the correct sequence from four given permutations. We ensure that the first appearance round of each object category is distinct to avoid ambiguity in ordering.

Agent-Object Spatial Relationship. This category focuses on constructing spatial metric relationships between the agent and a specific object O at a specific time T. The distance between the agent and object O at time T is defined as the shortest distance from the camera coordinate to any point in the object's point cloud. The angle of object O relative to the agent at time T is computed as the angle between the camera's horizontal orientation vector and the vector pointing from the camera to the center of object O. All objects involved in this category must satisfy the spatial visibility constraint, which means that their center coordinates, dimensions (length, width, height), and other spatial properties must be reliably obtainable from previous observations.

- *Distance (Judgement)*: This question type includes three forms of queries: (1) determining which of the three objects is currently farthest from or closest to the agent; (2) judging whether the current distance between the agent and a specific object is greater or smaller than the distance at the end of a previous round; (3) judging whether the current distances between the agent and two specific objects are greater or smaller than those at the end of a previous round, with four possible answer choices. For the first form, at least one object must be invisible in the current round, and the distance to the correct answer object must differ significantly (i.e., by more than a predefined threshold) from the distances to the other two objects. For the second and third forms, the change in distance between the two time points must also exceed the threshold to ensure a meaningful distinction.
- *Distance (Temporal Localization)*: This task asks the agent to identify the round in which it was closest to or farthest from a specific object. The distance in the correct round must be significantly smaller (for closest) or larger (for farthest) than in all other rounds.
- *Distance (Estimation)*: This query requires estimating the current distance between the agent and a specific object, which is invisible in the current round and thus requires recalling information from previous rounds.
- *Direction (Judgement)*: This question type includes three forms of queries: (1) judging whether a specific object is currently on the agent's left or right side; (2) judging whether a specific object currently lies in the left-front, left-back, right-front, or right-back quadrant relative to the agent; (3) identifying which two out of three objects are currently on the same side of the agent. For the first two forms, we enforce angular thresholds by excluding objects whose relative angles fall within 10 degrees of the decision boundaries between sides or quadrants, thereby avoiding ambiguity. For the third form, at least two of the three objects are invisible in the current round, forcing the model to rely on memory.
- *Direction (Temporal Localization)*: This query asks the agent to identify the round in which both objects A and B were located on the same side (left or right) relative to the agent. We ensure that in each round, both objects are clearly on either the left or right side (at least 10

degrees away from the decision boundary), and that there is exactly one round satisfying this condition.

- *Direction (Estimation)*: This query requires estimating the angle, clockwise or counterclockwise, of a specific object relative to the agent's current orientation. The object is not visible in the current round, requiring retrieval from prior observations.

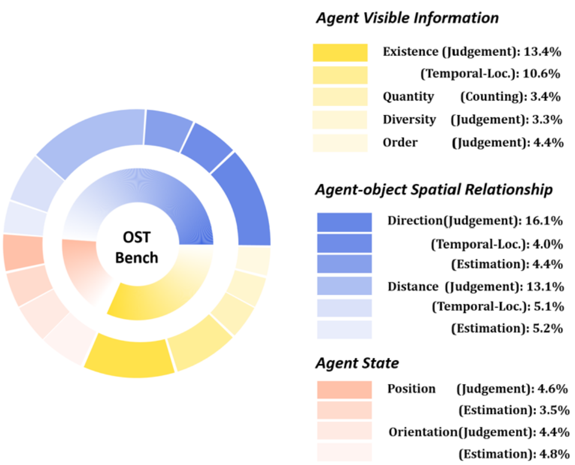


Figure 10: Distribution of sample counts across different subtypes in OST-Bench.

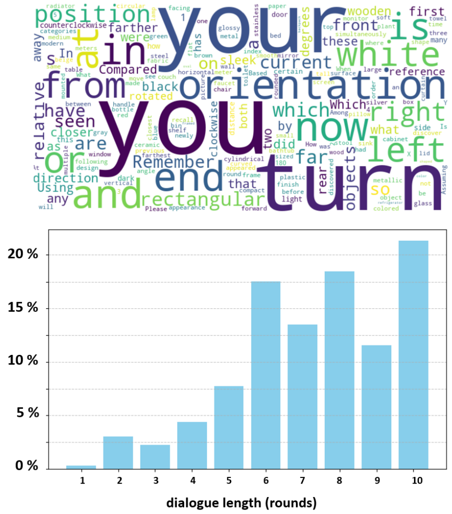


Figure 11: Word cloud (top) and dialogue length distribution (bottom) of OST-Bench.

A.4 Statistics

Based on the generation methods described above, OST-Bench totally consists of 1.4k trajectories(a trajectory per scene) and 10k data samples. The distribution of sample counts across different subtypes is shown in Fig. 10. We also present in Fig. 11 the word frequency distribution in OST-Bench (visualized as a word cloud), as well as the distribution of dialogue lengths.

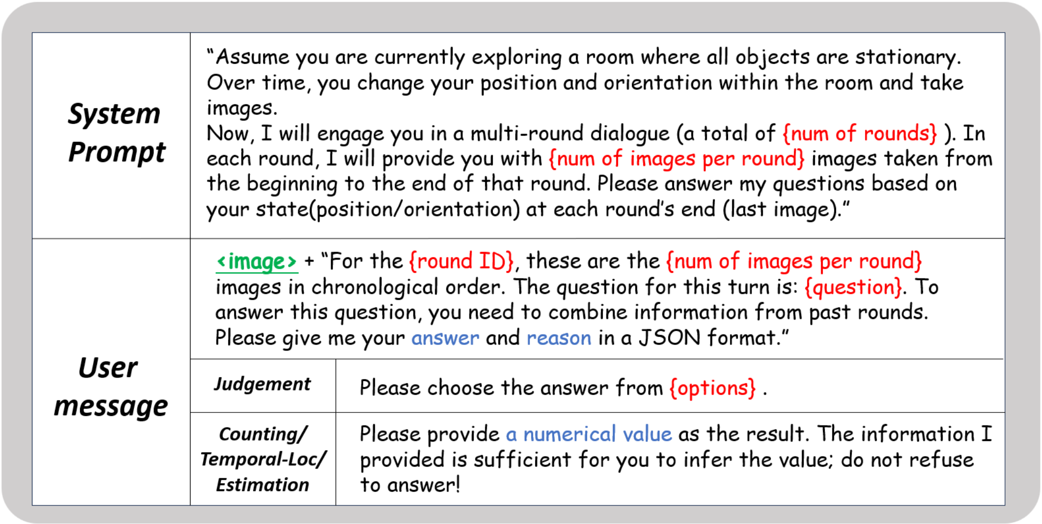
A.5 Benchmark Examples

In Fig. 16 and 17 we provide more examples from our benchmark, including a total of 12 data samples from two scenes (exploration trajectories).

B Implementation Details

For the multi-round dialogue, we first provide a system prompt to inform the models of the task setup. In each round, we sequentially input a set of images representing new video frames, along with a prompt containing a question, as illustrated in Fig.12. For judgment questions, we include the options in the prompt. For the other three qusetion formats (estimation, counting, and temporal-localization), we prompt the model to output a specific numerical value and explicitly instruct it to answer the question. This instruction is necessary, as we observed during experiments that models may otherwise refuse to respond, claiming insufficient information.

For proprietary models, we interact with the OpenAI and Anthropic APIs, both of which support multi-round dialogue with image inputs. In these APIs, models are invoked by explicitly specifying their model names. For the OpenAI API, we use *gpt-4o* for GPT-4o, *gpt-4.1* for GPT-4.1, *gemini-2.0-flash* for Gemini-2.0-Flash, and *gemini-2.0-flash-thinking-exp* for its thinking variant. For the Anthropic API, we use *claude-3-5-sonnet-latest* to access Claude-3.5-Sonnet. The system prompt is set to the task description, and each round's input includes newly added images and questions. For open-source models (InternVL, QwenVL, LLaVA-Onevision, and LLaVA-Video), we manually construct the multi-round context by concatenating the dialogue history, new images, and the current prompt as the input at each round. To avoid out-of-memory errors, input images are resized accordingly. For



System Prompt	"Assume you are currently exploring a room where all objects are stationary. Over time, you change your position and orientation within the room and take images. Now, I will engage you in a multi-round dialogue (a total of {num of rounds}). In each round, I will provide you with {num of images per round} images taken from the beginning to the end of that round. Please answer my questions based on your state(position/orientation) at each round's end (last image)."	
User message	<image> + "For the {round ID}, these are the {num of images per round} images in chronological order. The question for this turn is: {question}. To answer this question, you need to combine information from past rounds. Please give me your answer and reason in a JSON format."	
	Judgement	Please choose the answer from {options} .
	Counting/ Temporal-Loc/ Estimation	Please provide a numerical value as the result. The information I provided is sufficient for you to infer the value; do not refuse to answer!

Figure 12: Model input content, including the system prompt and inputs for each round. Text placeholders to be filled are highlighted in red, while the green <image> token represent image placeholders to be filled.

models with up to 8 billion parameters, inference is run on a single NVIDIA A100 GPU. For models with 32 billion parameters or more, we perform multi-GPU inference using 8 NVIDIA A100 GPUs via model and data parallelism. Additionally, we implement multithreaded processing to accelerate the inference of open-source models.

C Experiment Analysis Details

C.1 Spatially-Grounded Model Evaluation

In addition to general-purpose VLMs, we further evaluate several representative models that incorporate spatial grounding and memory mechanisms to varying degrees, including Spatial-MLLM[42], VLM-3R[21], and LLaVA-3D[53].

- Spatial-MLLM and VLM-3R follow a VGGT[40] + VLM architecture and take RGB image sequences as input, where VGGT provides geometry-aware scene representations.
- LLaVA-3D leverages RGB-D image sequences to encode 3D spatial information into 2D token embeddings.

All three models were trained on spatial reasoning datasets and achieved strong results on their respective benchmarks. We evaluate their performance on OST-Bench and compare them to their corresponding base models — Spatial-MLLM vs. QwenVL2.5-3B, and VLM-3R / LLaVA-3D vs. LLaVA-Video-7B.(Tab 5) The key findings are summarized below:

- **Only VLM-3R delivers consistent gains over its base model.** Spatial-MLLM and LLaVA-3D exhibit substantial performance drops, whereas VLM-3R shows steady improvements, particularly in Agent State, Agent-Object Spatial Relationship, and Estimation tasks.
- **Instruction-following ability degrades noticeably.** Compared to their base models—which reliably follow prompts and output both answers and reasoning—all three grounded models struggle to adhere to the required response format. Spatial-MLLM is constrained to producing only floating-point values or multiple-choice options, while all three models frequently omit reasoning or generate incoherent explanations.
- **Limited generalization beyond training-aligned distributions.** Despite excelling on spatial reasoning datasets such as VSI and MMScan, Spatial-MLLM and LLaVA-3D fail to generalize effectively to OST-Bench, which features more diverse and temporally grounded prompts. VLM-3R demonstrates partial transferability, yet its gains remain modest.

These observations suggest that while memory-enhanced spatial grounding can improve performance on tasks aligned with model pretraining objectives, it often comes at the expense of generalization. Such models may lose part of the base LLM's robustness — underperforming on previously simple tasks (e.g., Agent Visible Info or Counting questions in OST-Bench) and struggling with instruction-following in out-of-distribution settings. They tend to excel only on in-domain tasks and transfer poorly to broader benchmarks like OST-Bench, which involve more diverse and complex reasoning demands.

Method	Overall	JUD.	EST.	CNT.	TEMP.	A. State	A. Info	AO.
(base) QwenVL2.5-3B	34.8	47.9	18.7	59.4	19.8	34.2	47.5	25.7
Spatial-MLLM	26.8	37.3	21.9	29.5	15.3	25.5	39.4	20.9
(base) LLaVA-Video-7B	39.3	52.8	16.1	63.1	33.8	33.5	58.3	28.8
VLM-3R	42.9	55.1	28.3	49.6	36.0	39.9	58.1	34.4
LLaVA-3D	30.1	46.1	5.9	13.5	36.3	29.7	38.4	26.3

Table 5: **Performance comparison between specially designed models and their corresponding base models.** "A." abbrev for "Agent" and "AO." abbrev for "Agent-Object Spatial Relationship".

C.2 More Findings in Tables

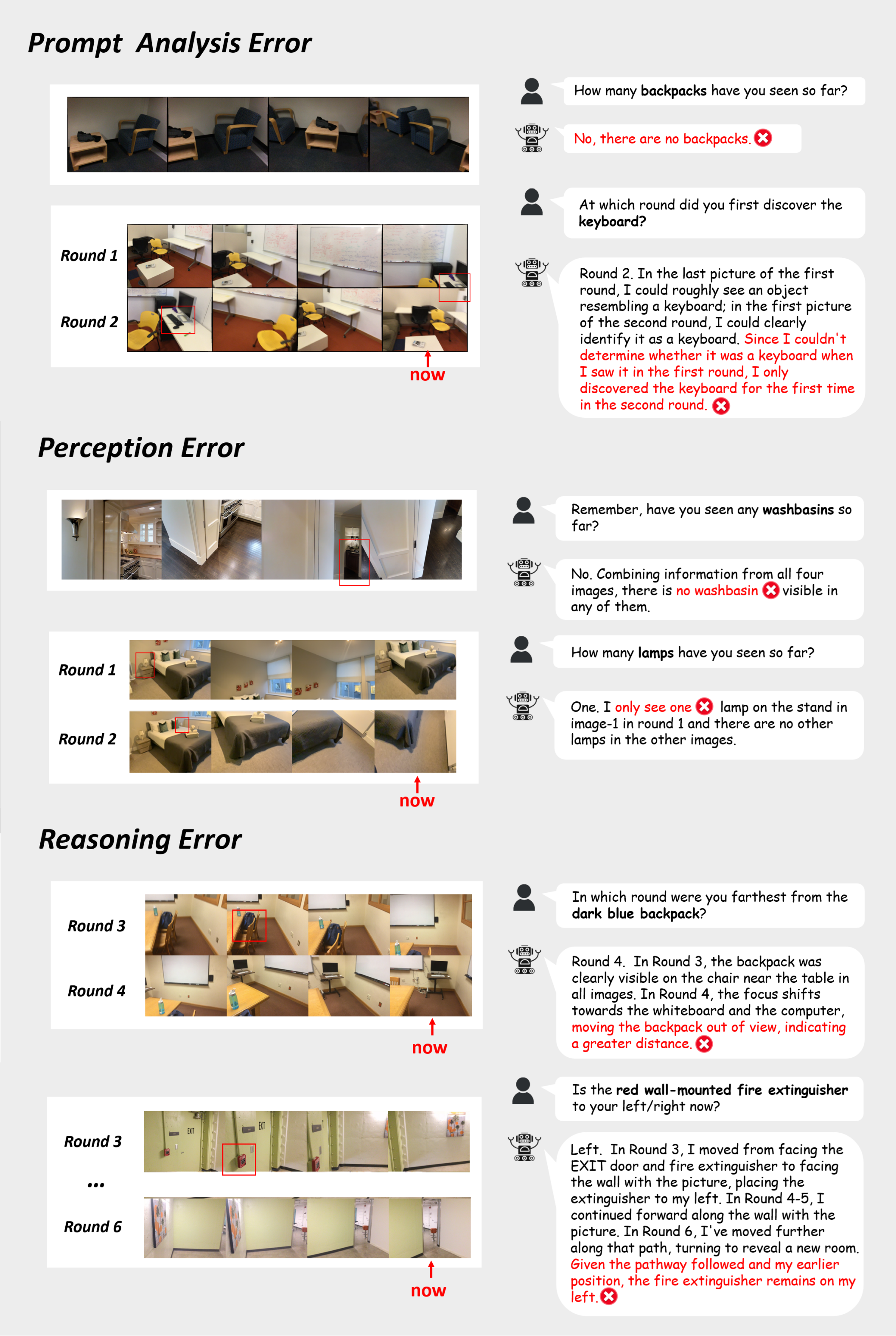
Difficulty of Estimation Tasks. As shown in Table 2 in the main paper, models perform particularly poorly on estimation tasks, achieving scores well below the chance-level baseline. Humans also struggle with these questions, obtaining significantly lower scores compared to other task categories. This is because estimation questions go beyond innate human perceptual abilities. Humans are better at perceiving spatial relationships approximately than estimating spatial measurements precisely, requiring not only spatial reasoning but also extensive empirical knowledge accumulated from experience.

Detection Success vs. Counting Failure. As shown in Table 2 in the main paper, models achieve notably high scores on object-existence questions, demonstrating a strong ability to identify whether and when objects appear. However, their performance drops significantly for object-quantity tasks, which require counting. Upon examining specific cases, we found that models frequently confuse whether objects across frames are the same or distinct, mistaking two different objects as identical or failing to track the same object across frames. This suggests that the task demands not just detection capabilities but also cross-frame reasoning.

The Illusion of Better Distance Understanding. As shown in Table 2 in the main paper, models appear to perform slightly better on Agent-object distance questions compared to Agent-object direction, but this advantage is superficial. This is primarily due to the *Spatio-temporal Reasoning Shortcut* phenomenon: models tend to assume that objects currently visible are closer, while those out of view are farther away, without engaging in genuine spatial reasoning. Although this heuristic can occasionally lead to correct answers, since such patterns do occur in a small portion of our benchmark, it fails to generalize. As a result, models still perform poorly on Agent-object distance questions overall.

C.3 Cases of Three Error Types

In Fig.13, we present examples of the three types of errors: *Prompt Analysis Error*, *Perception Error*, and *Reasoning Error*. In the first example of Prompt Analysis Error, the prompt explicitly requires the model to output a specific quantity. However, the model fails to interpret this requirement correctly and responds with "no" instead of providing a numerical answer such as "0". In the second example, the model misunderstands the meaning of the word "discover" in the prompt. It assumes that partially seeing the keyboard in Round 1 does not count as a discovery and that only fully observing it in Round 2 qualifies as such. This misinterpretation leads to an incorrect answer. In the two Perception Error examples, the model fails to correctly identify a washbin located in the corner of the room and only detects one of the two lamps in the bedroom, missing the other. In the Reasoning Error examples, although the model correctly understands the prompt and accurately perceives the location of the target objects, it makes an error in reasoning about their spatial relation with the agent, leading to incorrect conclusions.



Prompt Analysis Error

How many **backpacks** have you seen so far?

No, there are no backpacks. ✖

At which round did you first discover the **keyboard?**

Round 2. In the last picture of the first round, I could roughly see an object resembling a keyboard; in the first picture of the second round, I could clearly identify it as a keyboard. Since I couldn't determine whether it was a keyboard when I saw it in the first round, I only discovered the keyboard for the first time in the second round. ✖

Perception Error

Remember, have you seen any **washbasins** so far?

No. Combining information from all four images, there is no washbasin ✖ visible in any of them.

How many **lamps** have you seen so far?

One. I only see one ✖ lamp on the stand in image-1 in round 1 and there are no other lamps in the other images.

Reasoning Error

In which round were you farthest from the **dark blue backpack?**

Round 4. In Round 3, the backpack was clearly visible on the chair near the table in all images. In Round 4, the focus shifts towards the whiteboard and the computer, moving the backpack out of view, indicating a greater distance. ✖

Is the **red wall-mounted fire extinguisher** to your left/right now?

Left. In Round 3, I moved from facing the EXIT door and fire extinguisher to facing the wall with the picture, placing the extinguisher to my left. In Round 4-5, I continued forward along the wall with the picture. In Round 6, I've moved further along that path, turning to reveal a new room. Given the pathway followed and my earlier position, the fire extinguisher remains on my left. ✖

Figure 13: Illustrative Examples of the Three Error Types: Prompt Analysis Error, Perception Error, and Reasoning Error.

C.4 Cases of Spatio-temporal Reasoning Shortcut

In the main paper, we have discussed the *Spatio-temporal Reasoning Shortcut* phenomenon exhibited by the models. In Fig.14, we provide additional examples to further demonstrate the prevalence of this behavior. For clarity, we display only the key video frames relevant to each question. Temporal expressions in the questions and model responses are replaced with t1, t2, and t3, and marked above the corresponding frames. All of these examples demonstrate the model's tendency to rely on shortcuts in spatio-temporal reasoning.

In the first example, GPT-4o incorrectly infers that the blackboard has moved closer simply based on its transition from being invisible to visible, ignoring spatial cues such as the chairs and the decorations on the wall. In the second example, Gemini-2.0-Flash performs a seemingly correct inference using only two frames (the current and target frames), concluding that the wall currently in front of the agent is adjacent and perpendicular to the wall in t1, while disregarding intermediate frames that contain crucial contradictory evidence. In the third example, InternVL-2.5-78B observes that the TV was on the right side of the room in earlier frames and then directly assumes it remains there when it becomes invisible. In the fourth and fifth examples, the models make incorrect judgments due to the target object being invisible in the specific frames. In the sixth example, the model only focuses on the frames where the stand appears and the current frame, while skipping over intermediate frames that indicate the agent turned around, wrongly assuming that the current orientation is aligned with the previous one.

C.5 Subset Construction Process for Cross-View Analysis

As mentioned in the main paper, when constructing the dataset for the Cross-View subset, we first generate an initial batch of data using a rule-based method and then manually filter the data to obtain the final set of 200 samples. Our rule-based construction method for generating the Cross-view subset with different levels of difficulty is described as follows:

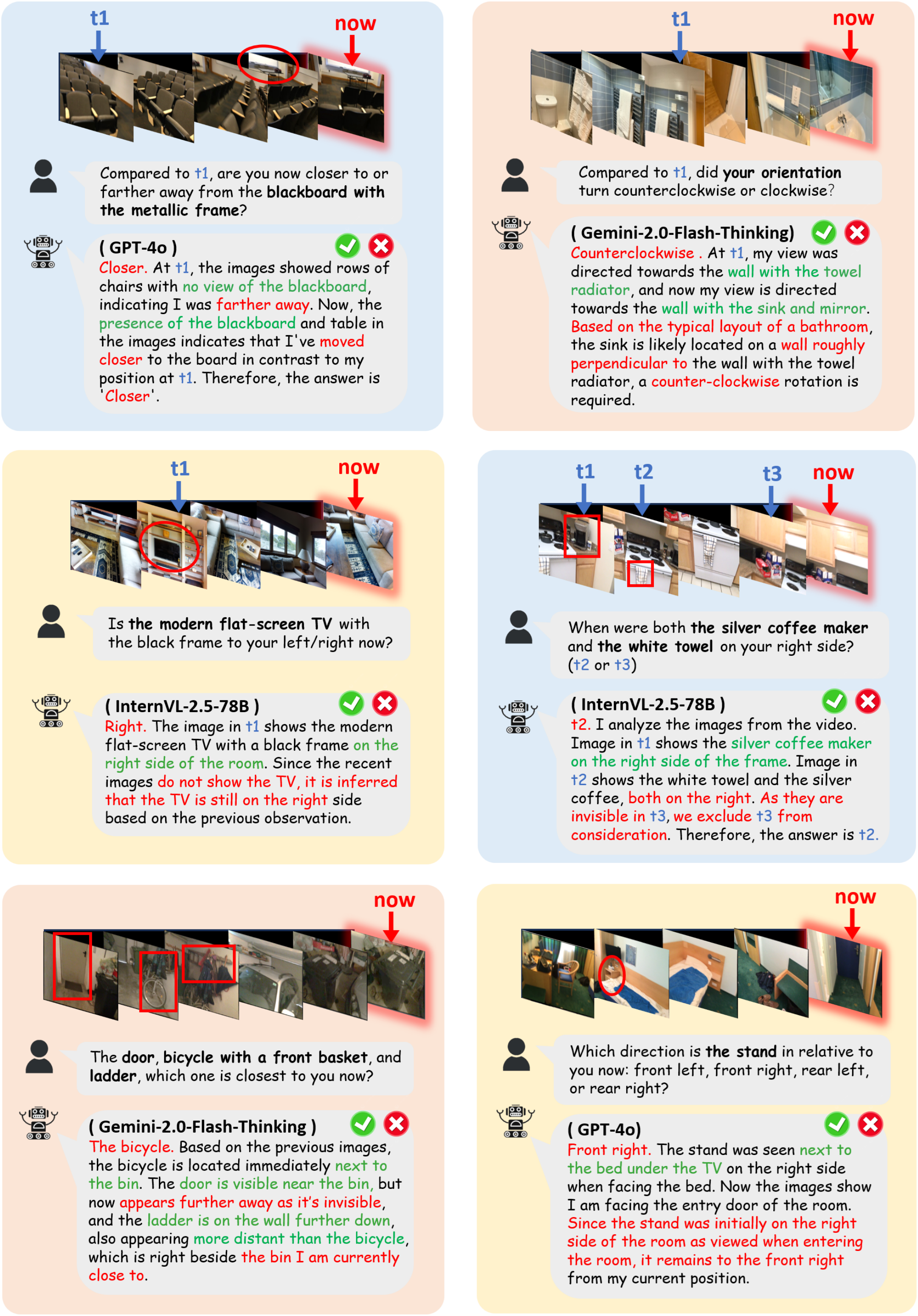
(a) **Single-Step Spatial Connection.** We first iterate over all possible object pairs (O_1, O_2) in the scene. For each object pair, we traverse all possible frame pairs (F_1, F_2) within the video sequence. A frame pair is selected if it satisfies the following conditions: (1) O_1 is visible in F_1 but not in F_2; (2) O_2 is visible in F_2 but not in F_1; (3) F_1 and F_2 share at least one overlapping object. This setup ensures that the spatial relationship between O_1 and O_2 can be inferred via single-step reasoning. All tuples (O_1, O_2, F_1, F_2) satisfying these constraints are collected as initial data for the **keyframe-based context**. To construct the **sequence-based context**, we embed F_1 and F_2 into a video sequence V that includes frames not containing O_1 or O_2, resulting in tuples of the form (O_1, O_2, V).

(b) **Multi-Step Spatial Connection.** Similarly, we iterate over all object pairs (O_1, O_2) and traverse all frame triplets (F_1, F_2, F_3) from the video sequence. A triplet is selected if it meets the following conditions: (1) O_1 is visible in F_1 but not in F_2 or F_3; (2)O_2 is visible in F_2 but not in F_1 or F_3;(3) F_1 or F_3 share at least one overlapping object;(4) F_2 or F_3 share at least one overlapping object;(5) F_1 and F_2 have no overlapping objects. This configuration ensures that solving the problem requires multi-step reasoning. All valid tuples $(O_1, O_2, F_1, F_2, F_3)$ satisfying these constraints are collected as initial data for the **keyframe-based context**. Similarly, to construct the **sequence-based context**, we embed F_1, F_2 and F_3 into a video sequence V that includes frames not containing O_1 or O_2, resulting in tuples of the form (O_1, O_2, V).

D Inference Time of the Models

Although OST-Bench does not impose real-time constraints, we conducted a supplementary study on models' inference time, indirectly reflecting the delay in decision-making exhibited by the models in real-world embodied tasks. Since the inference time of proprietary models is also affected by network latency, we restrict our analysis to open-source models and report their inference time per question.

The Fig.15 illustrates how the model's inference time per question changes as the duration of exploration increases. The results reveal a clear trend: as exploration time increases and more historical context accumulates, inference latency grows rapidly. When the number of dialogue rounds becomes large (e.g., beyond 10), the inference time becomes prohibitively high, especially for large-scale models, making real-time interaction impractical. This latency surge stems from the fact that any frame in history may contain critical information, forcing the model to attend to a growing



Compared to t1, are you now closer to or farther away from the **blackboard with the metallic frame**?

(GPT-4o) ✓ ✗
Closer. At t1, the images showed rows of chairs with no view of the blackboard, indicating I was farther away. Now, the presence of the blackboard and table in the images indicates that I've moved closer to the board in contrast to my position at t1. Therefore, the answer is 'Closer'.

Compared to t1, did **your orientation** turn counterclockwise or clockwise?

(Gemini-2.0-Flash-Thinking) ✓ ✗
Counterclockwise . At t1, my view was directed towards the wall with the towel radiator, and now my view is directed towards the wall with the sink and mirror. Based on the typical layout of a bathroom, the sink is likely located on a wall roughly perpendicular to the wall with the towel radiator, a counter-clockwise rotation is required.

Is **the modern flat-screen TV** with the black frame to your left/right now?

(InternVL-2.5-78B) ✓ ✗
Right. The image in t1 shows the modern flat-screen TV with a black frame on the right side of the room. Since the recent images do not show the TV, it is inferred that the TV is still on the right side based on the previous observation.

When were both **the silver coffee maker** and **the white towel** on your right side? (t2 or t3)

(InternVL-2.5-78B) ✓ ✗
t2. I analyze the images from the video. Image in t1 shows the silver coffee maker on the right side of the frame. Image in t2 shows the white towel and the silver coffee, both on the right. As they are invisible in t3, we exclude t3 from consideration. Therefore, the answer is t2.

The **door**, **bicycle with a front basket**, and **ladder**, which one is closest to you now?

(Gemini-2.0-Flash-Thinking) ✓ ✗
The bicycle. Based on the previous images, the bicycle is located immediately next to the bin. The door is visible near the bin, but now appears further away as it's invisible, and the ladder is on the wall further down, also appearing more distant than the bicycle, which is right beside the bin I am currently close to.

Which direction is **the stand** in relative to you now: front left, front right, rear left, or rear right?

(GPT-4o) ✓ ✗
Front right. The stand was seen next to the bed under the TV on the right side when facing the bed. Now the images show I am facing the entry door of the room. Since the stand was initially on the right side of the room as viewed when entering the room, it remains to the front right from my current position.

Figure 14: **More examples of Spatio-temporal Reasoning Shortcuts.** Green text marks correct reasoning; red indicates errors. For clarity, only key video frames relevant to each question are shown, with temporal references replaced by t1, t2, and t3.

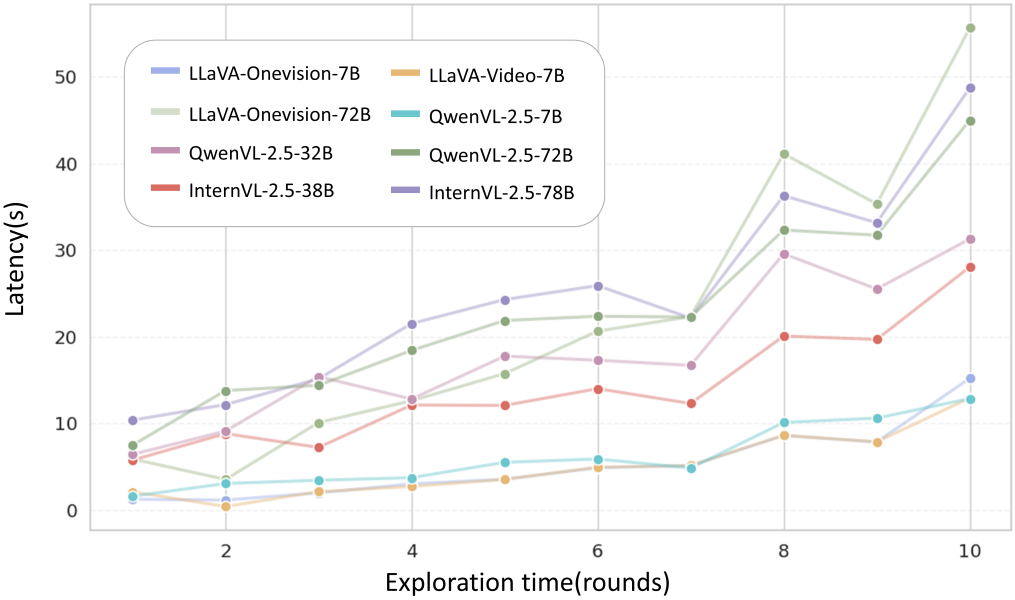


Figure 15: The trend of the model's inference time per question as the duration of exploration increases.

number of frames at every step. Thus, inference time scales approximately linearly with history length.

To provide context, we also measured human inference time. While average latency isn't directly comparable due to individual variation, we find that for all human testers, response time remained stable regardless of how long the exploration had lasted. This starkly contrasts with model behavior. The underlying reason is that humans can actively abstract and compress information throughout the exploration process, forming an internal knowledge base. Rather than treating each question as a fresh input, humans recall previously formed abstractions, enabling efficient reasoning without reprocessing all historical data.

This comparison highlights a critical need: for models to perform well in real-world embodied tasks, they must learn to dynamically distill and retain knowledge during exploration. Instead of passively accumulating history or answering questions in isolation, models should develop mechanisms to summarize and store essential information in an efficient, retrievable form, paving the way for scalable and real-time embodied reasoning.

E Social Impact

OST-Bench aims to advance the development of multimodal large language models (MLLMs) with stronger online spatio-temporal reasoning capabilities, which are critical for real-world embodied tasks such as assistive robotics, autonomous navigation, and human-robot interaction. By introducing a more realistic and challenging benchmark, we hope to drive progress toward more reliable and generalizable agents capable of perceiving and reasoning in real-world environments under online settings. However, as the benchmark assumes a static environment and focuses only on perception and reasoning, there is a risk of overestimating model readiness for real deployment. Caution is needed to avoid misuse or overreliance on models without broader capabilities like interaction or manipulation, which are essential for safe and responsible AI integration in the real world.

F License and Acess

F.1 License and Acess for Existing Assets

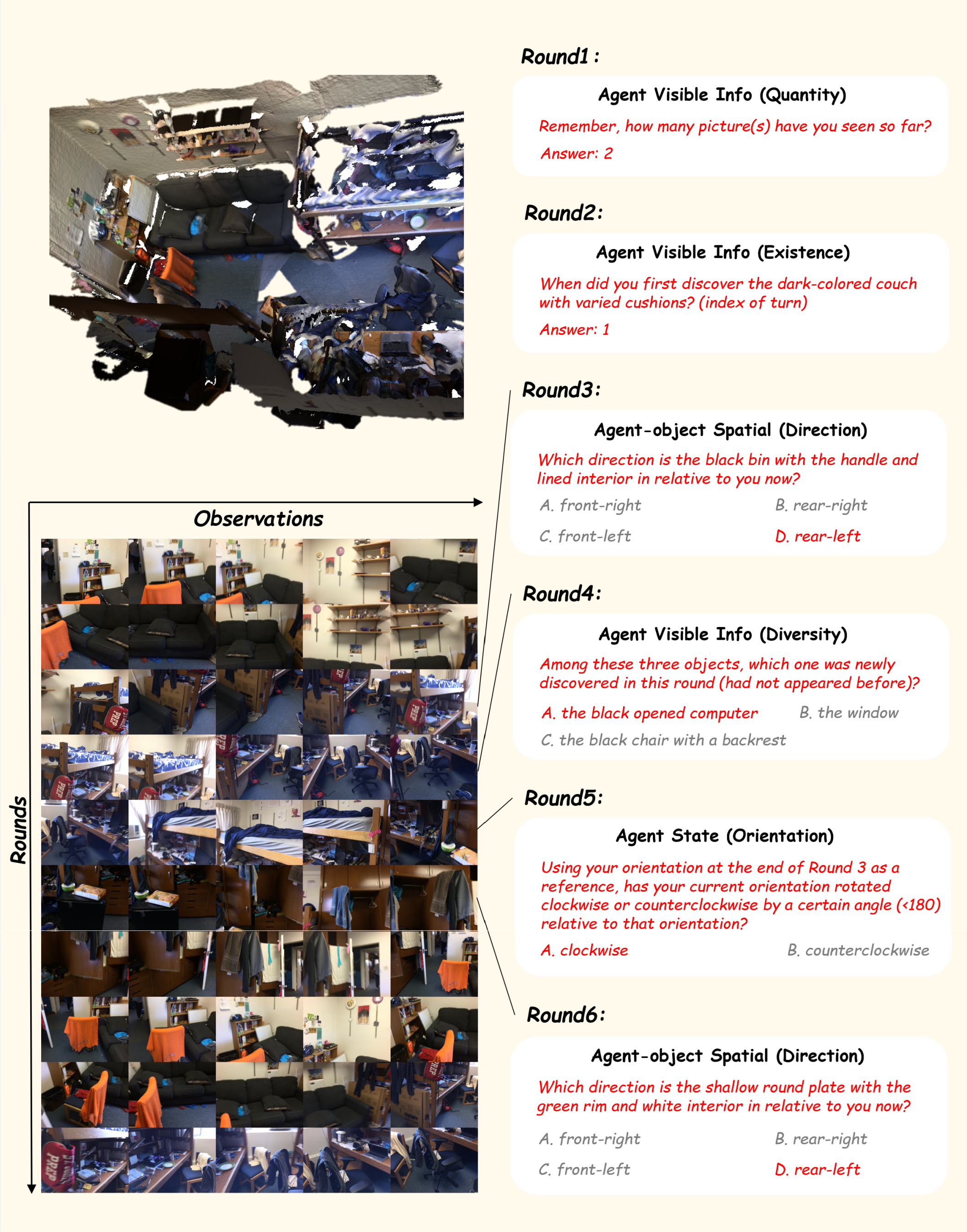
As mentioned in the main paper, our real-world scene data is sourced from ScanNet, Matterport3D, and ARKitScenes. To access and use these three datasets, users should follow their original licenses [4, 3, 1], and ask their official hosts for authorization. Additionally, our annotated data come from EmbodiedScan and MMScan, access to these datasets requires submitting a request via a Google Form [2] and following the license attached to the form.

We use ScanNet, Matterport3D, and ARKitScenes as the scene data and leverage the video information provided in them. We adopt the bounding box annotations and textual annotations from EmbodiedScan and MMScan as the base datasets for our benchmark. Throughout the usage of these datasets, their licenses and terms of use are properly respected.

F.2 License and Acess for OST-Bench

The OST-Bench dataset is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0) and available for direct download at `https://github.com/rbler1234/OST-Bench` or `https://www.kaggle.com/datasets/jinglilin/ost-bench/data`.

We release our benchmark under the CC-BY license and Terms of Use, and require that any use of the dataset for model evaluation be properly disclosed. This license supplements but does not override the original licenses of source materials; users must also comply with all relevant legal requirements concerning data subjects. This statement clarifies the obligations and liabilities associated with using this benchmark. While we strive to ensure the accuracy and legality of all samples, we do not guarantee their absolute completeness or correctness. We assume no responsibility for any legal or other issues that may arise from the use of OST-Bench, including but not limited to copyright infringement, privacy violations, or the misuse of sensitive information. By accessing, downloading, or using OST-Bench, you acknowledge that you accept this statement and agree to comply with the full terms of the CC-BY license. If you do not agree with these terms or the CC-BY license, you are not permitted to use this benchmark. OST-Bench will be hosted and maintained on GitHub and the Kaggle platforms.



Round1 :

Agent Visible Info (Quantity)

Remember, how many picture(s) have you seen so far?

Answer: 2

Round2:

Agent Visible Info (Existence)

When did you first discover the dark-colored couch with varied cushions? (index of turn)

Answer: 1

Round3:

Agent-object Spatial (Direction)

Which direction is the black bin with the handle and lined interior in relative to you now?

A. front-right B. rear-right

C. front-left D. rear-left

Round4:

Agent Visible Info (Diversity)

Among these three objects, which one was newly discovered in this round (had not appeared before)?

A. the black opened computer B. the window

C. the black chair with a backrest

Round5:

Agent State (Orientation)

Using your orientation at the end of Round 3 as a reference, has your current orientation rotated clockwise or counterclockwise by a certain angle (<180) relative to that orientation?

A. clockwise B. counterclockwise

Round6:

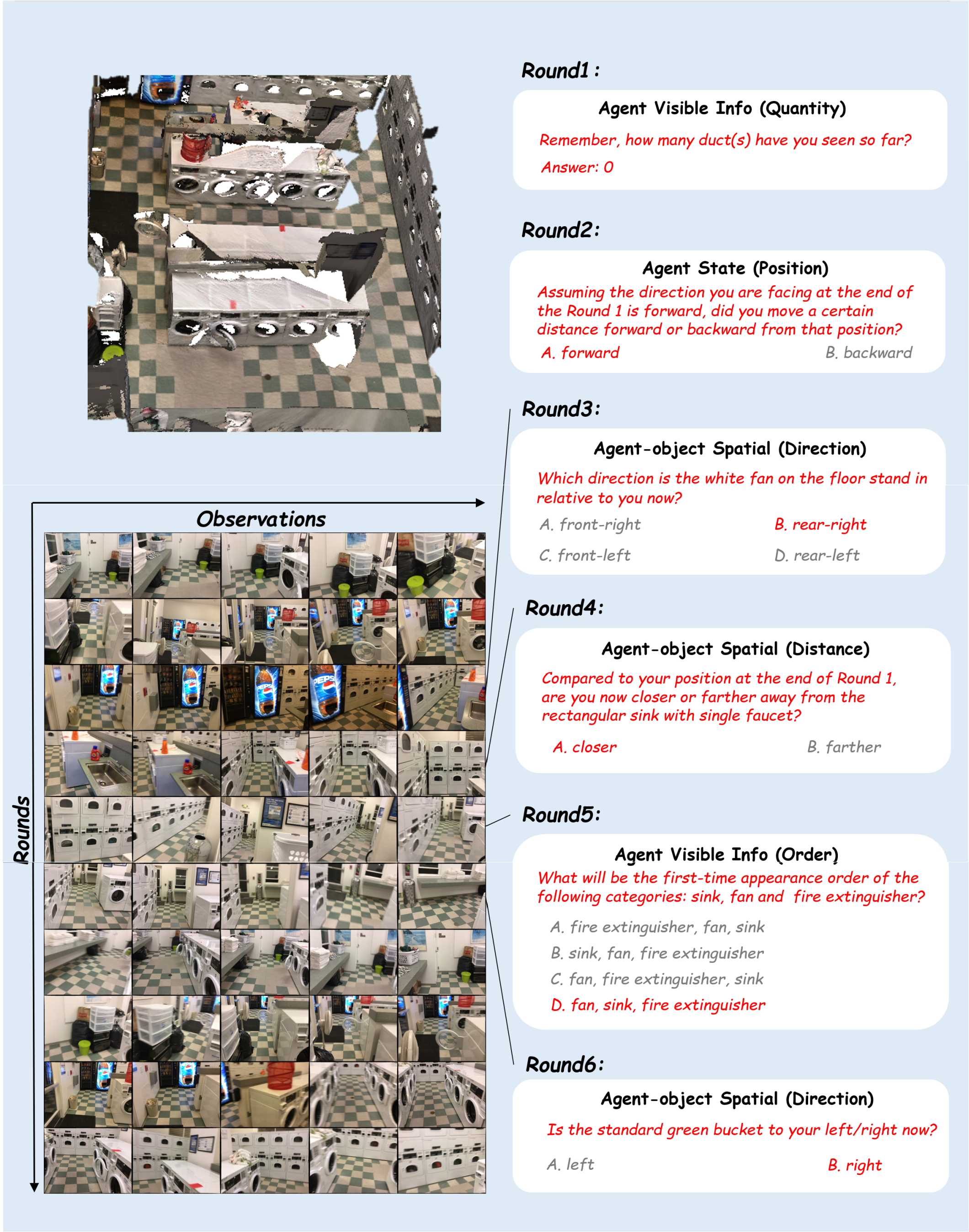
Agent-object Spatial (Direction)

Which direction is the shallow round plate with the green rim and white interior in relative to you now?

A. front-right B. rear-right

C. front-left D. rear-left

Figure 16: **Example 1 of OST-Bench data samples.** Each row represents the newly added observations in each round, with images input from left to right within each round. The example shows the question-answer pairs from the first six rounds.



Round1 :

Agent Visible Info (Quantity)

Remember, how many duct(s) have you seen so far?

Answer: 0

Round2:

Agent State (Position)

Assuming the direction you are facing at the end of the Round 1 is forward, did you move a certain distance forward or backward from that position?

A. forward *B. backward*

Round3:

Agent-object Spatial (Direction)

Which direction is the white fan on the floor stand in relative to you now?

A. front-right *B. rear-right*

C. front-left *D. rear-left*

Round4:

Agent-object Spatial (Distance)

Compared to your position at the end of Round 1, are you now closer or farther away from the rectangular sink with single faucet?

A. closer *B. farther*

Round5:

Agent Visible Info (Order)

What will be the first-time appearance order of the following categories: sink, fan and fire extinguisher?

A. fire extinguisher, fan, sink

B. sink, fan, fire extinguisher

C. fan, fire extinguisher, sink

D. fan, sink, fire extinguisher

Round6:

Agent-object Spatial (Direction)

Is the standard green bucket to your left/right now?

A. left *B. right*

Figure 17: **Example 2 of OST-Bench data samples.** Each row represents the newly added observations in each round, with images input from left to right within each round. The example shows the question-answer pairs from the first six rounds.