

# SpatialBench: Benchmarking Multimodal Large Language Models for Spatial Cognition

Peiran Xu  
Sun Yat-Sen University

Sudong Wang  
HKUST (GZ)

Yao Zhu  
Zhejiang University

Jianing Li  
Peking University

Yunjian Zhang  
UCAS

## Abstract

*Spatial cognition is fundamental to real-world multimodal intelligence, allowing models to effectively interact with the physical environment. While multimodal large language models (MLLMs) have made significant strides, existing benchmarks often oversimplify spatial cognition, reducing it to a single-dimensional metric, which fails to capture the hierarchical structure and interdependence of spatial abilities. To address this gap, we propose a hierarchical spatial cognition framework that decomposes spatial intelligence into five progressively complex levels from basic observation to high-level planning. Building upon this taxonomy, we construct SpatialBench, a large-scale, fine-grained benchmark covering 15 tasks aligned with these cognitive levels. To provide a unified evaluation across heterogeneous tasks, we further introduce a high-level capability-oriented metric that reliably assesses a model’s overall spatial reasoning ability. Extensive experiments over massive MLLMs reveal distinct performance stratification across cognitive levels: models exhibit strong perceptual grounding yet remain limited in symbolic reasoning, causal inference, and planning. Additional human tests demonstrate that humans perform selective, goal-directed abstraction, while MLLMs tend to over-attend to surface details without coherent spatial intent. Our work establishes the first systematic framework for measuring hierarchical spatial cognition in MLLMs, laying the foundation for future spatially intelligent systems. Our codes are available at <https://github.com/XPR2004/SpatialBench>.*

## 1. Introduction

In daily life, human can effortlessly integrate spatial information from their surroundings, with a capability known as spatial cognition. This ability extends beyond mere ob-



Figure 1. In a parking lot scenario, the vehicle must understand the relationships of surrounding objects, reason about possible events, and plan the optimal route to reach the exit.

ject recognition, serving as a cognitive bridge between perceptual inputs and higher-level functions such as reasoning and navigation. With the rapid advancements of large language models (LLMs) [9, 18, 21, 24, 55, 60], multimodal large language models (MLLMs) have recently emerged as a major step toward general-purpose visual–linguistic intelligence [4, 6, 8, 11, 16, 19, 20, 27, 35, 40, 44, 45, 57, 66, 68, 70–72, 76]. By jointly aligning visual and textual modalities within a shared semantic space, MLLMs have moved beyond abstract visual representations, integrating linguistic context to interpret scenes in a more structured and human-like manner. Recent advances show that MLLMs have exhibited spatial reasoning abilities [10, 12, 17, 26, 34, 36, 46, 53, 54, 61, 63, 64, 75], and several bench-

marks have been introduced to quantify these capabilities [3, 37, 38, 47, 59, 65, 67, 69]. However, they remain fragmented and task-oriented, often emphasizing performance on specific vision–language tasks rather than assessing spatial cognition as a structured capability. In addition, most benchmarks rely on synthetic or narrowly defined datasets, lacking the visual diversity and real-world complexity necessary to probe genuine spatial cognition. Consequently, these evaluations provide only a partial view of spatial intelligence, making it difficult to analyze cognitive processes and reveal systematic deficiencies in models.

To overcome these limitations, we propose a cognitively grounded evaluation framework for spatial intelligence in MLLMs. Inspired by the cognitive map theory [5, 49, 58] in neuroscience, we conceptualize spatial cognition as a hierarchical process that evolves from low-level perception to high-level reasoning and decision making. Specifically, our framework decomposes spatial understanding into five progressive levels, including **observation (L1)**, **topology and relation (L2)**, **symbolic reasoning (L3)**, **causality (L4)**, and **planning (L5)**, each corresponding to distinct cognitive functions involved in human intelligence. For the example in Figure 1, consider a scenario where a car leaves a parking lot. The model recognizes relevant entities and their spatial configurations (L1), then it understands topological relations such as lane connectivity and obstructions (L2) and maps visual symbols to semantic meanings and evaluates potential detour options (L3). After that, it infers causal outcomes of possible maneuvers (L4), and finally, it integrates prior reasoning to generate a coherent plan (L5). This hierarchical design provides a structured lens through which to interpret model behavior, enabling ability-oriented rather than task-oriented evaluation of spatial intelligence.

Building upon this framework, we construct a large-scale spatial video dataset named SpatialBench that grounds spatial cognition evaluation in realistic multimodal scenarios. Unlike previous synthetic or narrowly scoped datasets, our collection is captured from diverse indoor and outdoor environments, encompassing both static spatial layouts and dynamic scene evolutions that reflect the multi-level cognitive demands of spatial intelligence. To realize the five-level cognitive hierarchy, we design 15 categories of spatial reasoning tasks, each aligned with a distinct stage of spatial cognition. Each video is paired with carefully designed questions and annotations aligned with these cognitive dimensions, enabling systematic, fine-grained, and cognitively interpretable assessment for MLLMs.

Our experiments show that although current MLLMs perform well on perceptual and relational reasoning tasks, their competence declines sharply in high-level tasks. Insights from the one-shot and human benchmarks suggest that humans rely on selective, goal-oriented reasoning, while MLLMs exhibit diffuse attention to scene details,

lacking a unified spatial cognition.

Our contributions are summarized as follows:

- We establish the first comprehensive and cognitively grounded framework for assessing spatial intelligence of MLLMs. Drawing inspiration from cognitive map theory, our framework hierarchically decomposes spatial cognition into five progressive levels, shifting evaluation from task-driven to ability-oriented assessment.
- We construct SpatialBench, a large-scale multimodal dataset specifically designed for evaluating spatial cognition in MLLMs. It features 15 distinct categories of spatial reasoning tasks aligned with five hierarchical cognitive levels, providing a robust foundation for systematic and scalable evaluation.
- We introduce a high-level ability-driven evaluation metric to assess spatial cognition in MLLMs. Through extensive experiments on a wide range of state-of-the-art open-source and commercial models, we uncover their strengths and limitations in spatial reasoning. We further conduct controlled human evaluations to compare human and model reasoning, offering new insights into the gap between artificial and human spatial intelligence.

## 2. Related Works

### 2.1. Multimodal Large Language Models

Recent progress in Large Language Models (LLMs) [9, 18, 21, 24, 55, 60] has catalyzed the evolution of MLLMs, which integrate visual and linguistic modalities within a unified semantic space. By aligning image representations with textual instructions, MLLMs demonstrate remarkable capability in understanding and generating multimodal content across a wide spectrum of real-world tasks [2, 13, 15, 21, 29]. In general, an MLLM consists of three core components: a modality encoder [39, 52], a language backbone (LLM), and a modality interface that connects the two. Recently, the capability of MLLMs has expanded beyond static images to encompass video understanding [14, 23, 50]. This advancement has led to the incorporation of video–language alignment during pre-training, allowing MLLMs to jointly model temporal semantics and motion dynamics within a unified multimodal framework [42].

### 2.2. Visual-based Spatial Cognition

Visual-based spatial cognition seeks to endow MLLMs with the ability to perceive and reason about three-dimensional spatial relationships directly from visual inputs [12, 17, 26, 36, 48, 63]. Several benchmarks have emerged to evaluate this capability from different perspectives. Video-MME [22] provides a comprehensive assessment across a range of video-related tasks involving recognition and perception. VLM4D [74] emphasizes dynamic motion analysis, and STI-Bench [41] examines physical reasoning by test-

ing models' ability to predict and estimate object motions and displacements. VSI-Bench [65] introduces a structured benchmark comprising eight question types, each designed to probe different dimensions of spatial understanding in MLLMs. Ego-ST Bench [62] extends this to self-centered navigation in egocentric environments. SpatialLadder [38] presents a comprehensive dataset spanning multiple categories, covering spatial reasoning tasks from single-image understanding to video-based inference. In contrast, MindCube [67] assesses MLLMs' ability to infer complete spatial structures from limited visual observations.

Existing spatial understanding benchmarks lack a unified, ability-oriented framework, focusing instead on isolated tasks such as recognition, grounding, or motion prediction. Their evaluation design remains fragmented, failing to capture the hierarchical progression of spatial cognition from perception to reasoning. Moreover, most datasets are constructed by extending simple indoor scenes [61, 65], offering limited diversity in environmental dynamics and spatial complexity. As a result, they provide only partial insights into MLLMs' spatial reasoning capabilities, leaving the broader question of systematic, cognitively grounded evaluation largely unexplored.

### 3. Spatial Cognition Ability Framework

#### 3.1. Conceptual Foundations of Spatial Cognition

Spatial cognition encompasses the processes that enable intelligent systems to perceive, represent, and reason about spatial relationships within their environment. It involves acquiring spatial knowledge from sensory inputs, forming internal representations of space, and utilizing these representations for high-level tasks [31, 58]. The cognitive map theory [49, 58] provides a foundational opinion of how such representations are organized. Originating from behavioral studies of animals and later supported by discoveries of place cells in the hippocampal system, this theory proposes that intelligent agents construct internal, map-like structures that encode both metric and topological relations. These cognitive maps allow flexible navigation, route planning, and generalization beyond direct sensory experience. More importantly, they reveal that spatial knowledge is not a flat or static representation, but a hierarchically organized system, where low-level perceptual and motor cues are progressively abstracted into higher-order representations that integrate semantic, relational, and causal information. Building on this foundation, modern computational perspectives extend spatial cognition toward causal and multimodal understanding [5, 7, 32]. This hierarchical and integrative view provides the theoretical basis for our proposed framework, which conceptualizes spatial cognition in MLLMs as a progressive process evolving from perception to reasoning and ultimately to planning.

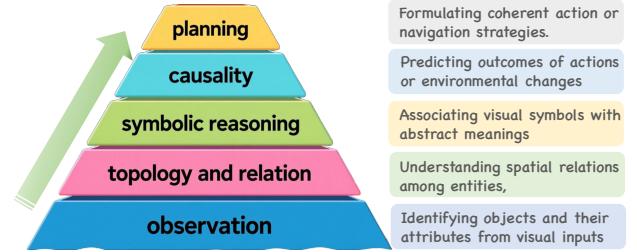


Figure 2. The proposed hierarchical spatial cognitive taxonomy.

#### 3.2. Hierarchical Spatial Cognitive Taxonomy

Based on the cognitive map theory, we propose the first systematic and hierarchical framework for spatial cognition evaluation, capturing the progressive development of spatial understanding from perception to high-level reasoning. Unlike prior benchmarks that focus on isolated visual tasks, our taxonomy is ability-driven, which means that each level represents a distinct, measurable cognitive capacity that reflects a specific stage of spatial intelligence. This hierarchical framework delineates spatial cognition into five progressive levels, each corresponding to a fundamental stage in the transition from sensory perception to deliberative reasoning: **observation** (L1), **topology and relation** (L2), **symbolic reasoning** (L3), **causality** (L4), and **planning** (L5). Together, these levels illustrate the progressive process by which intelligent systems transform raw perceptual information into organized spatial reasoning.

**Observation.** The foundational level of spatial cognition is observation, where the model identifies objects and their attributes from visual inputs. This stage corresponds to the extraction of basic perceptual elements such as object category, color, shape, and size.

**Topology and relation.** This level focuses on spatial relations among entities, such as adjacency, containment, orientation, and connectivity. Rather than perceiving isolated objects, it concerns the relational structure of the environment, describing how different elements are spatially arranged and interact within a coherent scene configuration.

**Symbolic reasoning.** Spatial understanding is extended beyond geometry into semantic interpretation. The agent is expected to associate visual symbols or spatial cues (e.g., arrows, pathways) with their abstract meanings and apply rule-based reasoning to infer spatial intent or constraints.

**Causality.** The causality level reflects the ability to infer spatiotemporal dependencies and predict outcomes of actions. It involves reasoning about how object movements, physical interactions, or agent behaviors lead to specific spatial consequences, integrating physical intuition and causal understanding into spatial reasoning.

**Planning.** This level represents the highest stage of spatial

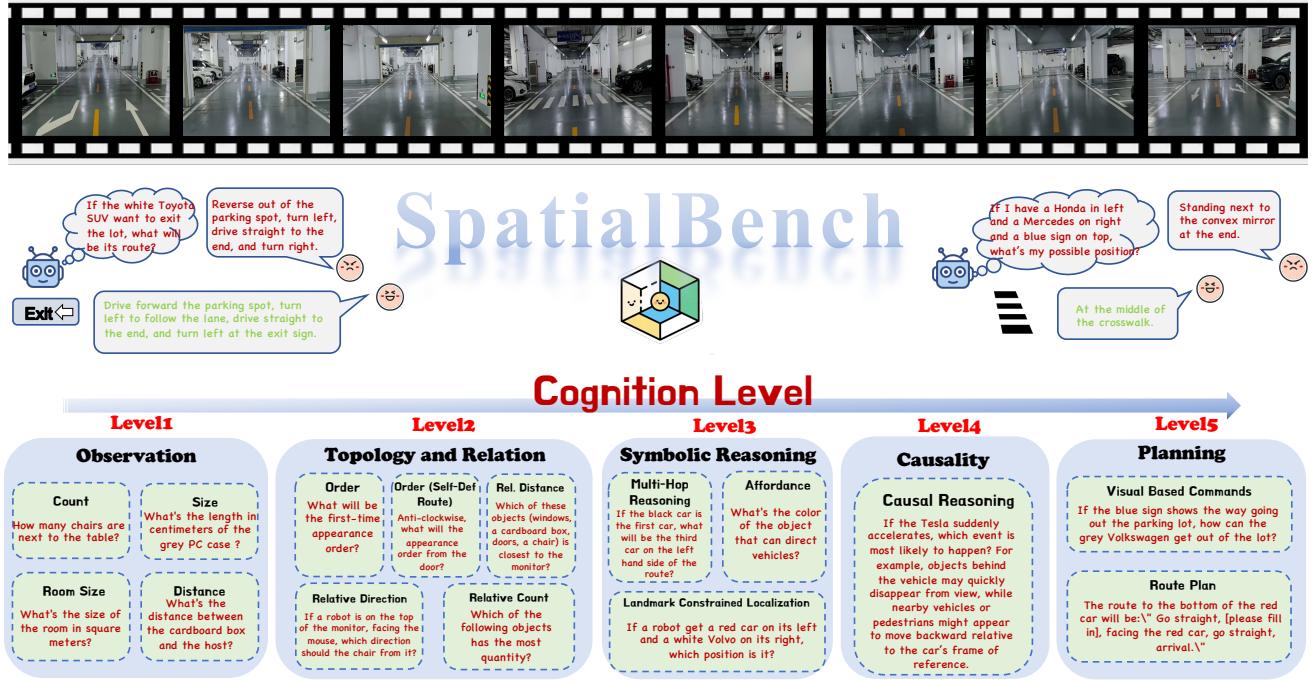


Figure 3. An overview of SpatialBench.

cognition, where perception, relational understanding, and causal reasoning are integrated to enable deliberate, goal-oriented decision making. At this level, an agent should synthesize its spatial representations and predictive reasoning to formulate coherent action sequences or navigation strategies that adapt to dynamic environmental contexts.

Collectively, these five levels delineate a progressive hierarchy that mirrors the cognitive evolution from sensory perception to reasoning and planning. By organizing spatial understanding into an ability-driven, hierarchically structured framework, our taxonomy provides the first systematic and cognitively grounded paradigm for assessing spatial intelligence in MLLMs, offering a unified foundation for interpreting and benchmarking spatial cognition across models of varying architectures and scales.

## 4. SpatialBench

### 4.1. Overview

We introduce SpatialBench, a large-scale benchmark for assessing evaluate the hierarchical spatial cognition of MLLMs using first-person videos. The dataset comprises 15 question types, each carefully mapped to one of the five cognitive levels introduced above:

- Observation (L1): object counting, object size, room size, absolute distance;
- Topology and relation (L2): appearance order, relative distance, relative direction, appearance order on self-

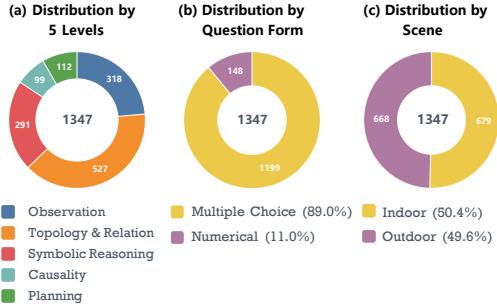


Figure 4. The statistics of SpatialBench.

defined route, relative counting;

- Symbolic reasoning (L3): multi-hop spatial reasoning, affordance, landmark-constrained pose localization;
- Causality (L4): spatial causal reasoning;
- Planning (L5): visual-based commands, route planning.

The SpatialBench dataset comprises 1,347 question-answer pairs sourced from 50 videos captured from an egocentric perspective. The collection spans both indoor and outdoor settings and includes static as well as dynamic scenes, covering challenging real-world contexts such as city roads, forest trails, residential areas, and underground environments. Together, these recordings provide diverse temporal and spatial complexity for evaluating MLLMs' spatial cognition ability in realistic scenarios. The detailed statistics of SpatialBench are shown in Figure 4.

## 4.2. Benchmark Construction

**Data Collection.** In contrast to prior benchmarks that adapt existing open-source datasets, SpatialBench is built from scratch through real-world recordings using our custom-designed sensing platform. The platform integrates a calibrated RGB camera and a 3D LiDAR sensor, which are spatially and temporally synchronized to ensure precise correspondence between visual and geometric modalities. The RGB camera continuously captures high-resolution visual streams, which serves as the basis for video question generation, while the LiDAR sensor synchronously records 3D point clouds that provide precise geometric information for size and distance related measurements. To ensure the accuracy and density of spatial data, we apply a filtering procedure to remove overly sparse or noisy point clouds. Data are collected from diverse environments, covering both indoor and outdoor settings such as offices, residential areas, city streets, and wooded regions, including both dynamic and static scenes. Each recording is conducted from a first-person perspective to preserve the egocentric characteristics essential for spatial understanding. For every video, we record standardized metadata including timestamps, scene categories, LiDAR frames, and synchronization parameters.

**Question-Answer Generation.** To ensure high-quality and semantically diverse annotations, human annotators work in pairs throughout the QA generation process. Within each pair, one annotator proposes candidate questions while the other independently reviews and validates them, checking for duplicates, ambiguous wording, and alignment with the intended cognitive level. All annotators are trained on the fifteen predefined task types, and they carefully review each video segment and propose candidate questions grounded in observed spatial relationships and scene dynamics. This collaborative review process ensures both the accuracy and relevance of the human-generated questions before they proceed to AI-assisted answer generation. For non-metric question types, we design specialized prompting templates tailored to each question category. These templates are then provided to state-of-the-art commercial models to generate corresponding answers. Along with each generated response, the model is required to output an evidence summary, including key frames and brief reasoning traces. For metric-related questions, we directly compute ground-truth answers using the LiDAR point cloud data. Precise 3D measurements are extracted through geometric fitting and spatial projection, thereby providing physically accurate ground-truth answers for all size and distance related questions. This hybrid design combines the semantic richness of human understanding, the efficiency of large model reasoning, and the geometric accuracy of sensor-derived measurements, resulting in a balanced and cognitively interpretable QA corpus.

**Annotation Verification.** We implement a multi-step ver-

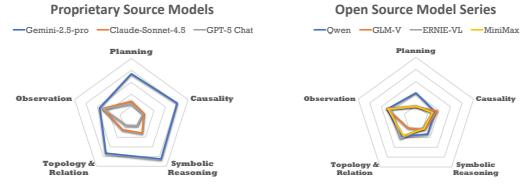


Figure 5. Comparison of models across spatial cognitive levels.

ification protocol to ensure the reliability of the generated annotations. For L1 and L2 questions, multiple leading models independently generate answers, and the consistency of these responses is evaluated. Questions with fully consistent model outputs are provisionally approved, but a subset of these automatically approved answers is further subject to human spot-checking to guarantee overall quality. Any discrepancies detected, whether during consistency checks or spot audits, trigger full human review. For question beyond L3, all annotations undergo mandatory human verification due to their higher cognitive complexity. The human review process follows a fixed checklist including: whether the evidence frames display key entities, whether multi-model outputs have been correctly interpreted, and whether answers conform to the pre-defined question schema. Annotations that are modified or rejected during the first review are subsequently evaluated by an additional annotator. This annotator examines the evidence summaries, the explanations and supporting screenshots submitted by the human reviewers, and also inspects the video and the question directly. Based on this comprehensive review, this annotator determines the final annotation. The verification protocol used ensures that all QA annotations are accurate, consistent, auditable, and reproducible, providing a trustworthy foundation for evaluating MLLMs’ spatial cognition.

## 5. Evaluation on SpatialBench

### 5.1. Setup

**Models.** We evaluate a diverse set of MLLMs to comprehensively assess their spatial cognitive abilities. Our benchmark covers both proprietary and open-source models with support for video understanding. For proprietary models, we include leading systems such as Gemini [56], GPT [30], and Claude-Sonnet. For open-source models, we evaluate representative families including Qwen [4], GLM [25], MiniMax [33], and ERNIE [73], encompassing variants across a broad range of architectures, parameter sizes (7B–235B), and training paradigms, enabling a systematic comparison of how different modeling strategies influence spatial cognition performance in SpatialBench.

**Metrics.** Following [65], we adopt evaluation metrics tailored to the answer type. Specifically, tasks in SpatialBench

					<b>Obj_count</b>	<b>Obj_Size</b>	<b>Room Size</b>	<b>Abs. Distance</b>	<b>App. Order</b>	<b>App. Order (Self-Def Route)</b>	<b>Rel. Distance</b>	<b>Rel. Direction</b>	<b>Rel. Count</b>	<b>Multi-Hop Reasoning</b>	<b>Affordance</b>	<b>Landmark Constrained Location</b>	<b>Causal Reasoning</b>	<b>Visual Based Commands</b>	<b>Route Plan</b>
Model	Ave.	Score	Rank		<b>Observation</b>				<b>Topology &amp; Relation</b>				<b>Symbolic Reasoning</b>			<b>Causality</b>	<b>Planning</b>		
Random	25.06	25.32	-		-				24.95	24.84	25.76	24.62	24.67	25.24	24.63	24.98	25.57	24.60	25.64
<i>Proprietary Models</i>																			
Gemini-2.5-pro	71.86	75.79	1	57.13	72.78	48.26	38.66	69.39	62.24	78.79	62.24	91.00	73.20	94.90	86.46	80.81	73.33	74.23	
GPT-4o-mini	33.44	30.92	9	31.49	54.17	28.26	24.95	73.47	49.48	26.53	34.69	20.00	31.96	9.18	23.96	33.33	0.00	31.96	
Claude-sonnet-4-5	32.10	28.56	12	41.59	70.51	40.87	41.67	21.79	13.58	32.10	33.33	20.99	36.25	30.00	26.58	22.50	20.00	28.75	
GPT-5-chat-latest	23.93	22.45	17	38.33	72.15	42.61	36.48	5.32	3.19	29.47	17.02	19.79	16.13	20.43	10.87	20.00	13.33	25.81	
<i>Open Source Models</i>																			
Qwen3-VL-235B-A22B-Instruct	38.84	37.79	2	43.76	70.10	40.43	34.33	68.37	24.49	30.30	32.99	31.00	36.08	25.51	37.50	33.33	26.67	43.30	
Qwen3-VL-235B-A22B-Thinking	36.60	36.27	3	45.36	71.03	43.18	27.73	45.92	31.58	31.63	20.41	37.00	44.33	17.35	27.08	28.28	40.00	45.36	
Qwen2.5-VL-72B-Instruct	32.73	33.82	4	26.14	57.94	43.48	30.21	37.76	13.27	27.27	36.73	30.00	40.21	16.33	43.75	36.36	6.67	37.11	
Qwen3-235B-A22B-Instruct-2507	34.88	33.65	5	37.03	54.85	46.52	29.48	74.49	63.27	24.24	23.47	19.00	35.05	7.14	18.75	33.33	0.00	41.24	
Qwen3-235B-A22B-Thinking-2507	34.99	33.25	6	32.00	60.90	24.29	32.13	58.24	28.26	29.67	39.13	23.91	28.09	40.91	21.84	29.55	40.00	33.70	
Qwen3-VL-30B-A3B-Instruct	37.07	32.83	7	48.02	59.28	63.04	34.85	74.49	58.16	30.30	25.51	21.00	35.05	6.12	22.92	29.29	0.00	38.14	
Qwen3-VL-30B-A3B-Thinking	36.83	31.22	8	48.79	63.40	27.83	30.72	71.43	45.36	31.31	27.55	32.00	35.05	11.22	32.29	26.26	13.33	30.93	
Qwen3-Next-80B-A3B-Instruct	32.48	30.89	10	30.20	48.35	22.17	25.77	73.47	9.18	34.34	30.61	23.00	50.52	12.24	30.21	25.25	13.33	36.08	
GLM-4V-Plus-0111	36.89	30.40	11	45.70	65.98	40.45	39.90	60.20	55.67	27.84	36.73	24.24	34.02	14.29	22.92	36.73	13.33	18.56	
ERNIE-4.5-Turbo-VL-32K	34.72	28.05	13	44.06	64.85	24.78	28.45	72.45	60.20	25.25	27.55	23.00	24.74	10.20	27.08	30.30	0.00	21.65	
GLM-4.5V	35.34	27.72	14	48.53	66.60	23.04	41.75	74.49	28.87	32.32	29.59	26.00	33.68	8.16	33.68	28.12	26.67	15.46	
MiniMax-Text-01	33.49	27.59	15	50.79	61.65	32.61	35.57	45.92	40.82	24.24	44.90	24.00	39.18	3.06	25.00	28.28	6.67	21.65	
QwQ-32B	29.76	23.81	16	37.82	54.54	30.87	24.33	72.45	35.71	17.17	31.63	18.00	34.02	9.18	11.46	28.28	13.33	13.40	
Qwen2.5-VL-7B-Instruct	17.84	17.45	18	19.21	28.04	17.39	12.37	18.37	11.34	11.11	16.33	20.20	15.05	14.74	23.96	18.95	33.33	13.54	
					19.69				15.48					17.96		18.95		16.22	

Table 1. Evaluation results on SpatialBench.

					<b>Obj_count</b>	<b>Obj_Size</b>	<b>Room Size</b>	<b>Abs. Distance</b>	<b>App. Order</b>	<b>App. Order (Self-Def Route)</b>	<b>Rel. Distance</b>	<b>Rel. Direction</b>	<b>Rel. Count</b>	<b>Multi-Hop Reasoning</b>	<b>Affordance</b>	<b>Landmark Constrained Location</b>	<b>Causal Reasoning</b>	<b>Visual Based Commands</b>	<b>Route Plan</b>
Model	Ave.	Score			<b>Observation</b>				<b>Topology &amp; Relation</b>				<b>Symbolic Reasoning</b>			<b>Causality</b>	<b>Planning</b>		
Gemini-2.5-pro	71.86	75.79		57.13	72.78	48.26	38.66	69.39	62.24	78.79	62.24	91.00	73.20	94.90	86.46	80.81	73.33	74.23	
Gemini-2.5-pro (one-shot)	67.43	68.12		53.10	76.29	43.91	42.06	52.04	54.08	74.75	61.22	91.00	66.67	91.84	81.25	66.57	66.67	67.01	
GPT-5-chat-latest	23.93	22.45		38.33	72.15	42.61	36.48	5.32	3.19	29.47	17.02	19.79	16.13	20.43	10.87	20.00	13.33	25.81	
GPT-5-chat-latest (one-shot)	59.03	61.08		40.10	71.34	53.04	38.66	58.16	48.98	58.59	39.80	74.00	53.12	85.71	78.12	56.57	53.33	65.98	
Qwen3-VL-235B-A22B-Instruct	38.84	37.79		43.76	70.10	40.43	34.33	68.37	24.49	30.30	32.99	31.00	36.08	25.51	37.50	33.33	26.67	43.30	
Qwen3-VL-235B-A22B-Instruct (one-shot)	61.26	63.20		44.80	69.38	46.52	27.94	65.31	43.88	74.75	59.18	73.00	52.08	86.73	79.17	53.54	53.33	73.20	
					47.29				63.29					72.76		53.54		70.54	

Table 2. Evaluation results under the one-shot setting.

are categorized into Multiple-Choice Answer (MCA) and Numerical Answer (NA) formats. For MCA tasks, we employ accuracy (ACC) as the primary evaluation metric [22, 28], which measures the proportion of exactly matched answers between the model’s predictions and the ground-truth labels. While for NA tasks, where answers involve continuous numerical values, we adopt the mean relative

accuracy (MRA) metric [43, 65], defined as:

$$\text{MRA} = \frac{1}{10} \sum_{\theta \in \Omega} \mathbb{I}\left(\frac{|y' - y|}{y} < 1 - \theta\right), \quad (1)$$

where  $y'$  and  $y$  are the model’s prediction and ground truth, respectively, and  $\theta$  is the confidence threshold from a thresholds set  $\Omega = \{0.5, 0.55, \dots, 0.95\}$ .

**Overall Score.** We propose a high-level capabil-

		<i>Obj. count</i>	<i>Obj. Size</i>	<i>Room Size</i>	<i>Abs. Distance</i>	<i>App. Order</i>	<i>App. Order (Self-Def Route)</i>	<i>Rel. Distance</i>	<i>Rel. Direction</i>	<i>Rel. Count</i>	<i>Multi-Hop Reasoning</i>	<i>Affordance</i>	<i>Landmark Constrained Location</i>	<i>Causal Reasoning</i>	<i>Visual Based Commands</i>	<i>Route Plan</i>		
Model	Ave.	Score	Observation					Topology & Relation					Symbolic Reasoning			Causality	Planning	
Human Level	93.89	96.40	100.00	53.75	67.50	60.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	
Gemini-2.5-pro	71.86	75.79	57.13	72.78	48.26	38.66	69.39	62.24	78.79	62.24	91.00	73.20	94.90	86.46	80.81	73.33	74.23	

Table 3. Human benchmark test results.

ity-oriented overall score to integrate performance across the five hierarchical cognitive levels. To emphasize higher-level reasoning while preserving balance, we assign adaptive weights to each level:

$$C_i = F_i S_i, \quad F_i = \alpha D_i + 0.1(1 - \alpha)E_i, \quad i = 1, \dots, 5, \quad (2)$$

where  $S_i$  is the standard deviation of model scores,  $D_i$  is the category’s question proportion, and  $\alpha$  controls the trade-off between baseline and adaptive weights and  $E_i$  is non-negative variables to be optimized with  $\sum_{i=1}^5 E_i = 10$ . Under the monotonicity constraint  $C_{i+1} > C_i$  the target is to minimize

$$\text{Var}(C_{i+1} - C_i) - k \sum_{j=1}^5 C_j^2, \quad (3)$$

yielding complexity-aware overall scores.  $\alpha$  and  $k$  will be fixed manually. Finally, the overall score is computed as  $\sum_{i=1}^5 C_i M_i$ , where  $M_i$  is the average rating on each level.

## 5.2. Benchmarking MLLM Performance

Table 1 presents the model performance on SpatialBench across five hierarchical levels. Proprietary models (e.g., Gemini-2.5-pro) occupy the top of the ranking and show substantially higher overall scores than most open-source counterparts. This gap is most pronounced on high-level tasks such as symbolic reasoning, causality, and planning. Nevertheless, several open-source series (e.g., Qwen variants) reach competitive performance on lower and mid levels, indicating that open models can achieve strong perceptual and relational understanding with carefully designed training paradigms. For open-source models, a clear correlation appears between model scale and average performance: larger models generally achieve higher overall scores, showing that scale remains an important factor. However, size alone does not guarantee better performance. Within the same model family, different versions (such as instruction-tuned vs. thinking mode) can show clear performance gaps. This suggests that architectural design and the integration strategy for visual and linguistic information plays a crucial role in spatial cognitive ability.

Figure 5 illustrates the average performance of several representative models across the five cognitive levels. It can

be observed that across the board, observation and topology tasks are relatively easier for MLLMs. Most models achieve substantially higher scores on tasks such as object counting, size and distance estimation, and simple topological queries. By contrast, higher-level abilities remain challenging for models, including symbolic reasoning, causality, and visual planning. On these tasks, average performance drops noticeably and variance between models increases. This pattern suggests that while current MLLMs can reliably extract visual evidence and reason about basic relations, they struggle to (a) convert perceptual inputs into robust symbolic rules, (b) infer causal or dynamic consequences accurately, and (c) generate multi-step and convincing plans for a given objective.

## 5.3. One-Shot Evaluation

In-context learning is a critical capability for MLLMs, reflecting their ability to leverage minimal examples for reasoning. To evaluate this, we perform a one-shot assessment: for each task, a single annotated example comprising a QA pair, reasoning explanation, and key frames is provided. Models are then asked to answer a test question from the same task, enabling examinations on how effectively they can generalize spatial reasoning from limited guidance.

We select several representative models from the one-shot evaluation, including two proprietary models: Gemini-2.5-pro (the best-performing) and GPT-5-chat-latest (the weakest), as well as a strong open-source model Qwen3-VL-235B-A22B-Instruct. Gemini-2.5-pro shows a decline in performance, primarily across the intermediate cognitive levels. In contrast, GPT-5-chat-latest and Qwen3-VL demonstrate substantial improvements. Although these models still trail Gemini in absolute scores, their performance under prompting approaches that of the previously strongest model. This result indicates that even lower-performing models can rapidly enhance specific spatial reasoning abilities when given minimal in-context guidance, highlighting the potential of one-shot prompting to boost MLLMs’ higher-level cognitive capabilities. The results suggest that some powerful models, such as Gemini 2.5 Pro, can reason efficiently and accurately from context on their own, performing better than when given human-guided prompts. This indicates that their built-in contextual understanding is particularly well-suited for spatial intelligence

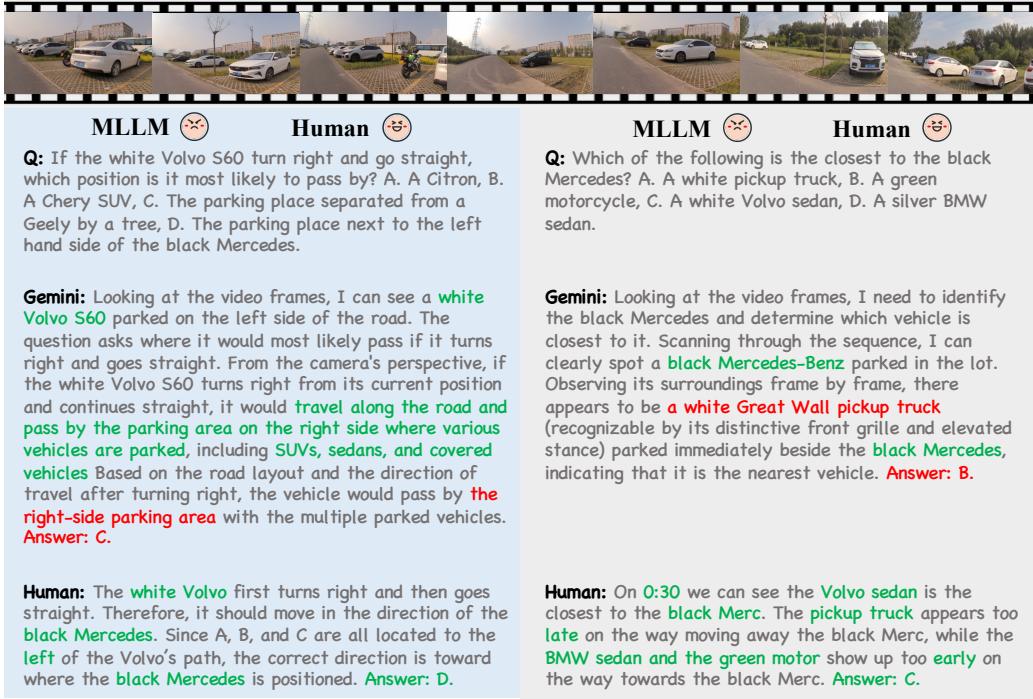


Figure 6. The differences of thinking processes between MLLM and human.

tasks. By contrast, GPT-5 appears comparatively weaker in intrinsic spatial logic, yet achieves more advanced spatial intelligence by leveraging the GPT series’ steadily improving linguistic capabilities [1]. Similarly, Qwen3-VL, which demonstrates strong text-based reasoning skills [51], benefits from one-shot prompting in a way comparable to GPT-5, highlighting the role of linguistic reasoning in enhancing spatial cognition under minimal guidance.

#### 5.4. Benchmarking Human Performance

To quantify the gap between current MLLMs and human-level intelligence across different cognitive dimensions, we conduct a human benchmark experiment. In this experiment, 33 human participants are presented with SpatialBench, and they provide answers directly without guidance, allowing us to measure their unaided spatial reasoning and planning abilities. The results shown in Table 3 indicate that human achieve near-perfect performance across nearly all tasks with an overall score of 96.40. Humans perform particularly well on higher-level tasks: symbolic reasoning, causality, and planning, and all achieve essentially 100% accuracy, while even lower-level observation and topology tasks maintain strong performance. Across all levels, humans exhibit not only higher absolute accuracy but also more consistent performance, reflecting robust generalization and contextual understanding that MLLMs have yet to fully achieve. These findings underscore the substantial gap

that still exists between human and machine spatial intelligence. While MLLMs show promising abilities in extracting visual information and reasoning about simple topological relations, they remain far from matching human performance in tasks that require integrating high-level reasoning. This test thus provides a clear target for future model development, highlighting the need to enhance multi-step reasoning and context-sensitive planning in MLLMs.

Figure 6 shows an example of the differences between Gemini and human. The human focuses on the key directional cue: the turning path of the white Volvo, and quickly eliminates irrelevant options based on spatial orientation, demonstrating strong goal-directed and spatially grounded reasoning. In contrast, Gemini describes the entire scene in a more exhaustive but unfocused manner, mentioning many vehicles and areas without identifying the crucial spatial relationship. This suggests that while MLLMs can recognize objects and describe scenes accurately, they often lack selective attention and directional understanding, leading them to infer by association rather than by reasoning about movement and geometry.

## 6. Conclusion

In this work, we introduce SpatialBench, a comprehensive benchmark built upon a five-level hierarchical spatial cognition framework that progressively evaluates MLLMs from low-level observation to high-level planning. This layered

design reflects the cognitive progression from perception to decision-making, enabling a more interpretable and fine-grained diagnosis of multimodal spatial intelligence. Experimental results show that while modern MLLMs demonstrate strong perception and relational reasoning, their abilities in symbolic abstraction, causal inference, and spatial planning remain limited. SpatialBench establishes a principled foundation for hierarchical evaluation and future development of spatially grounded intelligence in MLLMs.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [8](#)
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [2](#)
- [3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19107–19117, 2022. [2](#)
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. [1](#), [5](#)
- [5] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332, 2013. [2](#), [3](#)
- [6] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. [1](#)
- [7] Matthew Botvinick, David GT Barrett, Peter Battaglia, Nando de Freitas, Darshan Kumaran, Joel Z Leibo, Timothy Lillicrap, Joseph Modayil, Mohamed Shakir, Neil C Rabinowitz, et al. Building machines that learn and think for themselves. *Behavioral and Brain Sciences*, 40, 2017. [3](#)
- [8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. [1](#)
- [9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [1](#), [2](#)
- [10] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE, 2025. [1](#)
- [11] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13817–13827, 2023. [1](#)
- [12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. [1](#), [2](#)
- [13] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. [2](#)
- [14] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26418–26428, 2024. [2](#)
- [15] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. Pali: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations (ICLR)*, 2023. [2](#)
- [16] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2023. [1](#)
- [17] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [1](#), [2](#)
- [18] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhang-hao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yong-hao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, 2023. [1](#), [2](#)
- [19] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [1](#)
- [20] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangwen Zhang, Kaisheng

- Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations (ICLR)*, 2024. 1
- [21] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*, pages 8469–8488, 2023. 1, 2
- [22] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, 2025. 2, 6
- [23] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhao Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 2
- [24] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chat-gpt outperforms crowd workers for text-annotation tasks. *National Academy of Sciences of the United States of America*, 2023. 1, 2
- [25] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 5
- [26] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26181–26191, 2025. 1, 2
- [27] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13504–13514, 2024. 1
- [28] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. 6
- [29] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10878–10887, 2023. 2
- [30] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5
- [31] Maria Kozhevnikov and Jyotika Puri. Different types of survey-based environmental representations: egocentric vs. allocentric cognitive maps. *Brain Sciences*, 13(5):834, 2023. 3
- [32] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 3
- [33] Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025. 5
- [34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research (TMLR)*, 2025, 2025. 1
- [35] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1
- [36] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulic. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1786–1807, 2024. 1, 2
- [37] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqiao Zhang, Yongliang Shen, Weiming Lu, and Yuetong Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025. 2
- [38] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yuetong Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025. 2, 3
- [39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022. 2
- [40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742, 2023. 1
- [41] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025. 2
- [42] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual repre-

- sentation by alignment before projection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5971–5984, 2024. 2
- [43] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6
- [44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1
- [45] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296, 2024. 1
- [46] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: on-demand spatial-temporal understanding at arbitrary resolution. In *International Conference on Learning Representations (ICLR)*, 2025. 1
- [47] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 2
- [48] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2
- [49] John O’keefe and Lynn Nadel. The hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4):487–494, 1979. 2, 3
- [50] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025. 2
- [51] Qwen. Qwen3-vl-235b-a22b-instruct. <https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct>, 2025. Accessed: 2025-11-13. 8
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2
- [53] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Krähenbühl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations (ICLR)*, 2025. 1
- [54] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning. *arXiv preprint arXiv:2410.16162*, 2024. 1
- [55] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023. 1, 2
- [56] Gemini Team. Gemini: A family of highly capable multi-modal models. *arXiv preprint arXiv:2312.11805*, 2023. 5
- [57] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, 2025. 1
- [58] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948. 2, 3
- [59] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2
- [60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2
- [61] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting MLLM capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 1, 3
- [62] Peiran Wu, Yunze Liu, Chonghan Liu, Miao Liu, and Junxiao Shen. St-think: How multimodal large language models reason about 4d worlds from ego-centric videos. *arXiv preprint arXiv:2503.12542*, 2025. 3
- [63] Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research (TMLR)*, 2024, 2024. 1, 2
- [64] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023. 1
- [65] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643, 2025. 2, 3, 5, 6
- [66] Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingen Zhou. mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051, 2023. 1
- [67] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025. 2, 3
- [68] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28202–28211, 2023. 1

- [69] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025. 2
- [70] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 1
- [71] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.
- [72] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozhi Gao, and Joyce Chai. Groundhog grounding large language models to holistic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14227–14238, 2024. 1
- [73] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019. 5
- [74] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: towards spatiotemporal awareness in vision language models. *arXiv preprint arXiv:2508.02095*, 2025. 2
- [75] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 1
- [76] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*, 2024. 1

## A. Mathematical Modeling of Overall Score

To provide a unified measure of a model’s spatial cognitive competence, we introduce an overall score that integrates performance across all five hierarchical cognitive levels. The goal of this metric is not merely to average accuracy, but to construct a complexity-aware evaluation that reflects the progressively demanding nature of spatial cognition. Lower levels mainly involve perceptual and geometric understanding, while higher levels require abstract reasoning, causal inference, and planning. A meaningful overall score must therefore (1) preserve the relative importance of each cognitive level, (2) emphasize higher-level reasoning without overwhelming lower-level contributions, and (3) maintain fairness across categories with different question counts and variances.

To achieve these goals, we design a weighting mechanism that adaptively adjusts each level’s contribution based on its intrinsic difficulty and score distribution. Instead of manually assigning fixed weights, we formulate an optimization-driven approach that learns monotonic, complexity-aligned weights while controlling the imbalance between levels. This results in an overall metric that is interpretable, robust to distributional differences across task categories, and sensitive to a model’s true cognitive progression rather than raw accuracy alone.

To formalize the construction of our overall score, we assign adaptive weights to the five cognitive levels ( $i = 1, 2, 3, 4, 5$  corresponding to Observation through Planning). Our goal is to encourage higher performance on more complex cognitive abilities while preserving a balanced contribution across levels.

We begin by computing, for each level, the empirical standard deviation

$$S_i, i = 1, 2, \dots, 5, \quad (4)$$

which reflects the intrinsic difficulty and discriminative range of that category. This quantity is paired with the category’s question proportion (i.e., its initial weight)

$$D_i, i = 1, 2, \dots, 5, \quad (5)$$

serving as the baseline weighting factor. To introduce controlled adaptivity, we adjust these baseline weights through

$$F_i = \alpha D_i + 0.1(1 - \alpha)E_i, \quad (6)$$

where  $\alpha$  is a manually chosen hyperparameter governing the balance between the baseline distribution  $D_i$  and the optimized adjustment  $E_i$ . The resulting effective weight for each category is then

$$C_i = F_i S_i, i = 1, 2, \dots, 5, \quad (7)$$

which we require to increase monotonically with cognitive complexity.

To obtain a smooth hierarchy of difficulty, we further aim to make the increments between adjacent levels as uniform as possible. This leads to the following constrained nonlinear optimization problem:

$$\begin{aligned} \min f &= \text{Var}(C_{i+1} - C_i) - k \sum_{j=1}^5 C_j^2, \quad i = 1, 2, 3, 4 \\ (\mathbf{P}) \quad &\left\{ \begin{array}{l} C_{i+1} - C_i > 0, \quad i = 1, 2, 3, 4 \\ \sum_{j=1}^5 E_j = 10 \\ E_i \geq 0, \quad i = 1, 2, \dots, 5 \end{array} \right. , \end{aligned} \quad (8)$$

where the parameter  $k$  expresses the preference for maintaining the original baseline separation prescribed by the

standard deviations  $S_i$ . Solving Problem (P) yields a set of complexity-aware weights  $C_i$ , which are subsequently combined with the per-level average model ratings to produce the final overall score.

Mathematically, the original formulation introduces strict inequality constraints

$$C_{j+1} - C_j > 0, \quad j = 1, 2, 3, 4, \quad (9)$$

which renders the feasible region

$$\{(E_i) \mid C_{j+1} - C_j > 0\} \quad (10)$$

an open set in Euclidean space. As a result, the constraint set is not compact and is incompatible with the standard form required by many numerical optimization solvers; specifically, it does not contain the closed and bounded simplex

$$\{(E_i) \mid \sum_{j=1}^5 E_j = 10, \quad E_i \geq 0\}. \quad (11)$$

Rather than transforming the problem into a fully standardized form, which would introduce unnecessary complications and yield no practical benefit, we relax the strict inequalities to non-strict ones and solve the following modified program:

$$\begin{aligned} \min f &= \text{Var}(C_{i+1} - C_i) - k \sum_{j=1}^5 C_j^2, \quad i = 1, 2, 3, 4 \\ (P)' \quad &\left\{ \begin{array}{l} C_{i+1} - C_i \geq 0, \quad i = 1, 2, 3, 4 \\ \sum_{j=1}^5 E_j = 10 \\ E_i \geq 0, \quad i = 1, 2, \dots, 5 \end{array} \right. . \end{aligned} \quad (12)$$

The relaxation makes the feasible set closed and compatible with standard numerical solvers, while preserving the essential structural constraint that higher-level categories should not receive smaller weights than lower ones. After solving P)', we simply discard degenerate solutions in which

$$C_i = C_{i+1}, \quad \exists i \in \{1, 2, 3, 4\}, \quad (13)$$

since such solutions violate the intended strictly increasing hierarchy of cognitive complexity. This procedure is computationally effective and fully adequate for our application, as the optimization landscape naturally favors non-degenerate solutions when  $k$  is chosen appropriately.

The new constraint set is now in standard form. Since both

$$\begin{aligned} &\{(E_i) \mid C_{j+1} - C_j > 0, \quad j = 1, 2, 3, 4\} \text{ and} \\ &\sum_{j=1}^5 E_j = 10, \quad E_i \geq 0 \} \end{aligned} \quad (14)$$

are closed subsets of the Euclidean space, and their intersection (our new feasible region) is also closed.

Recall that  $F_i = \alpha D_i + 0.1(1-\alpha)E_i$  which is continuous in each  $E_i$ , Consequently,  $C_i = F_i S_i$  is also continuous since  $S_i$  is constant. Therefore,

$$f_{sum} = \sum_{j=1}^5 C_j^2 \quad (15)$$

is continuous as well. Let  $\mathbf{E} = (E_1, E_2, \dots, E_5)$ , since each  $C_i$  is a continuous function of  $\mathbf{E}$ , define

$$g_i(\mathbf{E}) = C_{i+1}(\mathbf{E}) - C_i(\mathbf{E}), \quad i = 1, 2, 3, 4, \quad (16)$$

which is continuous in  $\mathbf{E}$ . Their linear combination

$$\bar{g}(\mathbf{E}) = \frac{1}{4} \sum_{i=1}^4 g_i(\mathbf{E}) \quad (17)$$

remains continuous. Thus,

$$f_{var} = \text{Var}(g_i(\mathbf{E})) = \frac{1}{4} \sum_{i=1}^4 (g_i(\mathbf{E}) - \bar{g}(\mathbf{E}))^2 \quad (18)$$

is also continuous because it is constructed from addition and squaring of continuous functions. Hence,

$$f = f_{var} - f_{sum} \quad (19)$$

is a continuous function of  $\mathbf{E}$ .

By the Extreme Value Theorem, any continuous function on a closed and bounded feasible region must attain its minimum. Therefore, problem P)' admits at least one global minimizer under the new constraint formulation.

To numerically solve the nonlinear programming problem, we employ the `scipy.optimize.minimize` function in Python. We evaluate multiple combinations of  $(\alpha, k)$ , and observe that  $(\alpha, k) = (0.4, 0.01)$  yields a particularly favorable optimum. Under this setting, the optimizer returns

$$(E_i) = (0, 0, 1.4911, 3.8347, 4.6742), \quad (20)$$

achieving

$$\text{Var}(C_{i+1} - C_i) = 0.0264, \quad (21)$$

which is substantially lower than the variance obtained when  $\alpha = 0$ , where the optimal solution becomes

$$(E_i) = \frac{2}{3}(1, 2, 3, 4, 5) \quad (22)$$

with

$$\text{Var}(C_{i+1} - C_i) = 0.0968. \quad (23)$$

Moreover, the chosen parameter setting effectively resolves the initial undesirable ordering issue observed when  $\alpha = 1$ , where the resulting sequence satisfies

$$C_4 < C_5 < C_1 < C_3 < C_2, \quad (24)$$

violating the monotonicity condition. In contrast,  $(\alpha, k) = (0.4, 0.01)$  produces a monotone and well-behaved solution consistent with our design constraints.

## B. Case Study

To further illustrate the challenges of video-based spatial reasoning, we analyze a representative failure case from the causal reasoning category, and the results are shown in Figure 7. The question asks where a white Volvo S60 would most likely pass if it turns right and continues straight. While the ground truth is option D, the two evaluated MLLMs exhibit markedly different behaviors. MLLM1 demonstrates strong perceptual grounding: it accurately identifies the Volvo, reconstructs the forward scene layout, and reasons about nearby landmarks such as a black Mercedes and empty parking spots. However, its reasoning deteriorates once the camera performs a U-turn, causing the model to implicitly assume that the camera’s motion still reflects the Volvo’s hypothetical trajectory; this misalignment leads it to break the continuity of the reconstructed scene and misinterpret the spatial ordering. In contrast, MLLM2 fails much earlier and provides only a superficial description. It implicitly treats the camera’s viewing direction as the Volvo’s movement direction, confusing left/right relations and ultimately selecting an incorrect parking area. Human annotators, however, easily recognize that options A–C lie in the opposite direction of a right turn, and that the Volvo would naturally drive toward the black Mercedes, making D the only plausible answer. This case reveals a critical limitation: weaker models struggle to form any coherent scene representation, while stronger models can reconstruct static layouts but fail to maintain spatial consistency across continuous camera motion. Even when stable landmarks (e.g., the covered car) exist to support scene continuity, current models do not reliably leverage them. Overall, this example demonstrates that successful causal reasoning in dynamic scenes fundamentally depends on robust 3D scene reconstruction and continuity tracking—capabilities that remain insufficiently developed in existing MLLMs.

## C. Egocentric Reasoning Breakdown

We further examine two representative failure cases involving egocentric direction reasoning in both indoor and outdoor environments. For the indoor case shown in Figure 8, the task requires determining the relative direction of an AC control panel from the perspective of a robot standing

on top of the projector and facing the whiteboard. While the ground truth is Backward, the model incorrectly predicts Left. A close inspection of the reasoning trace reveals that although the model correctly identifies the projector, the whiteboard, and the AC control panel within the room, it fails at the final spatial transformation: converting absolute room layout into the camera’s egocentric frame. The model implicitly adopts the recorder’s viewing direction as the reference frame, causing a systematic rotation of its inferred directions. As a result, the AC control panel which lies behind the camera when facing the whiteboard is erroneously mapped to its left. This case highlights a broader weakness: even when object localization is accurate, current MLLMs often conflate scene-centric, camera-centric, and agent-centric coordinate systems. Such confusion leads to consistent directional inversion or orthogonal errors, especially in indoor scenes where multiple frames provide shifting viewpoints. Strengthening explicit frame-of-reference reasoning remains essential for achieving reliable spatial understanding.

For the outdoor case in Figure 9, the model again fails to correctly align the robot’s egocentric perspective with the scene layout observed from the recorder’s view. Although the model successfully identifies the motorcycle and the black Mercedes, it misinterprets the orientation of the robot after moving “onto the road.” Because it interprets the forward direction based on the camera’s viewpoint rather than the robot’s own heading, the model incorrectly assumes that the robot should turn left to move down the road. This perspective confusion leads to a reversed decision in the route-planning step, causing the model to select “turn left” instead of the correct “turn right.” The error highlights the model’s difficulty in maintaining consistent egocentric orientation during multi-step spatial navigation.



Q: If the white Volvo S60 turn right and go straight, which position is it most likely to pass by?

'A. A Citroen.', 'B. A Chery SUV',

'C. The parking place separated from a Geely by a tree.', 'D. The parking place next to the left hand side of the black.'

Q-type: Causal Reasoning, Ground Truth: D

**MLLM1:** \*\*Locate the White Volvo S60:\*\* ...

\*\*Determine the Starting Point and Direction of Travel:\*\* The camera is moving forward along the road, passing the Volvo. ... its path would be the same as the camera's path from that point forward.

\*\*Analyze the Path of Travel:\*\*

- \* In frames 27-32, the camera is passing the Volvo S60.
- \* Immediately after the Volvo, in frames 33-36, the camera passes a few empty parking spots and then a black Mercedes-Benz B-Class, also parked on the right.

\*\*Evaluate the Options based on the Path:\*\*

- \* **A.** ... This occurs **\*before\*** reaching the Volvo S60. Therefore, if the Volvo moves forward from its spot, it will not pass the Citroën.

- \* **B.** ... The Chery SUV is not part of the same continuous scene as the Volvo.

- \* **C.** ... In frame 27, ... This describes the parking spot **\*next to\*** the Volvo's starting position ... it moves **\*away\*** from this spot, not **\*past\*** it along the road.

- \* **D.** ... a black Mercedes is parked a short distance down the road from the Volvo... (following the camera's path), it will inevitably pass the black Mercedes. The "[option D]" refers to the empty spot between the Volvo's original position and the Mercedes. The Volvo **would drive right past** this spot.

**MLLM2:** Looking at the video frames, I can see a white Volvo S60 parked on the left side of the road. The question asks where it would most likely pass if it turns right and goes straight. From the camera's perspective, if the white Volvo S60 turns right from its current position and continues straight, it would travel along the road and pass by the parking area on the right side where various vehicles are parked, **including SUVs, sedans, and covered vehicles**. Based on the road layout and the direction of travel after turning right, the vehicle would pass by the right-side parking area with the multiple parked vehicles. The answer is C.

**Human:** The white Volvo turn left and go straight. It should drive towards the black Merc. ABC are all on the opposite way.

**Result:** MLLM2 probably believe the camera shows the trace "the white Volvo S60 turn right and go straight". MLLM1 did a splendid job identifying the scene and infer like a human. But it get lost when the camera holder make the U-turn.

**Conclusion:** Poor MLLMs **cannot reconstruct** the scenes from video while great ones **have decent reconstruction capability**, resulting whether the question is properly solved. But great MLLMs still need to **improve in continuous scenes identification**. In this scene the covered car can support continuity as a landmark.

Figure 7. Case study on causal reasoning in dynamic scenes. This example shows that weaker MLLMs fail to construct a coherent scene, while stronger ones reconstruct the static layout but lose continuity when the camera changes direction.



Question: If a robot is standing on the top of the projector and facing the white board, what's the direction of the AC control panel from the robot?

- Options:
- Right.
  - Forward.
  - Left.
  - Backward.

**Ground Truth: D**

**Model Answer: C**

#### Model Reasoning:

- Establish the Robot's Position and Orientation:\*\* The question states that a hypothetical robot is standing on top of the projector and facing the whiteboard.
- Locate the Projector:\*\* The video frames clearly show a white projector placed on the long conference table. For example, frames 3, 5, and 9 show the projector on the table.
- Locate the Whiteboard:\*\* The video frames show a large, mobile whiteboard on a stand at one end of the room. Frames 5 and 9 show the whiteboard's position relative to the projector and the table. The projector is aimed towards a blank wall, but the question specifies the robot is facing the \*whiteboard\*, which is at the far end of the room.
- Establish the Robot's "Forward" Direction:\*\* Since the robot is on the projector and facing the whiteboard, its "forward" direction is towards that end of the room.

5. \*\*Locate the AC Control Panel:\*\* By carefully examining the walls in the video, a small, square, beige-colored AC control panel can be seen. It is most clearly visible in frames like 1, 4, 7, and 8. It is mounted on the long, solid wall, the one with a row of chairs lined up against it.

6. \*\*Determine the Relative Direction:\*\*

- \* The robot is on the table, facing the whiteboard (Forward).
- \* The wall with the windows is on the **opposite side** of the table from the AC control panel.
- \* The AC control panel is on the wall to the **left** of the conference table, from the perspective of someone facing the whiteboard.
- \* Therefore, from the robot's position on the projector, looking forward at the whiteboard, the AC control panel is on the wall to its left.

#### Error Analysis:

The model is confused by relative directions and recorder-view direction, resulting in a complete opposite consequence. This indicates that the model fails to construct an agent-centric coordinate system and instead defaults to the camera's egocentric frame.

Figure 8. Failure case illustrating egocentric direction misinterpretation. Although the camera is defined as facing the whiteboard, the model confuses recorder-view and agent-view directions, leading to the incorrect prediction ("Left").



Question: You are a robot. If you are at the position of motorcycle and facing the road and your target position is the black Mercedes, your route will be: "Go straight, onto the road, [please fill in], go straight, until the black Mercedes is on my right, turn right, go straight." Fill in the "[]".

Options: A. turn right B. turn left C. make a U-turn D. stop and wait

**Ground Truth: A**

**Model Answer: B**

#### Model Reasoning:

- Identify the Starting Point:\*\* The question states the starting position is that of the "motorcycle". By examining the video frames, we can locate a green and black motorcycle parked on the right side of the road.
  - Identify the Target:\*\* The target position is the "black Mercedes". Scanning through the video frames, we find a black Mercedes B-Class parked further down the same road, also on the right side.
  - Analyze the Initial Instructions:\*\* The route begins with "Go straight, onto the road...". The motorcycle is parked perpendicularly to the road. To get "onto the road", the robot must move forward, out of the parking space. After this maneuver, the robot will be on the paved road.
  - Determine the Next Necessary Action:\*\* The next instruction is "...go straight, until the black Mercedes is on my right...". To follow this, the robot must first orient itself to travel **\*along\*** the road in the direction of the Mercedes.
- \* The robot is on the road, facing the left side.

- \* The black Mercedes is located further down the road.
- \* To face this direction, the robot must make a 90-degree turn to its left.

5. \*\*Visualize the Path:\*\*

"Go straight, onto the road". The robot is now on the road, **facing the cars on the left**. The robot must turn **\*left\*** to face down the road towards the Mercedes. "go straight...". Now correctly oriented, the robot drives along the road. "...until the black Mercedes is on my right...". The robot continues until it is parallel to the target vehicle. "...turn right, go straight." The robot turns into the parking space to reach the Mercedes.

6. \*\*Evaluate the Options:\*\*

Based on this step-by-step analysis, the only logical action to fill the blank is "**turn left**".

#### Error Analysis:

MLM misjudge the direction from "robot" view and from recorder view, thinking the robot should go down the road with opposite initial turn.

Figure 9. Failure case illustrating route-planning direction misinterpretation. Although the robot's forward orientation is clearly defined after moving onto the road, the model confuses recorder-view and agent-view directions, leading to the incorrect prediction ("Left").