Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces

Jihan Yang[1]* Shusheng Yang[1]* Anjali W. Gupta[1]* Rilyn Han[2]* Li Fei-Fei[3] Saining Xie[1]
[1]New York University [2]Yale University [3]Stanford University

Project Page Evaluation Code VSI-Bench

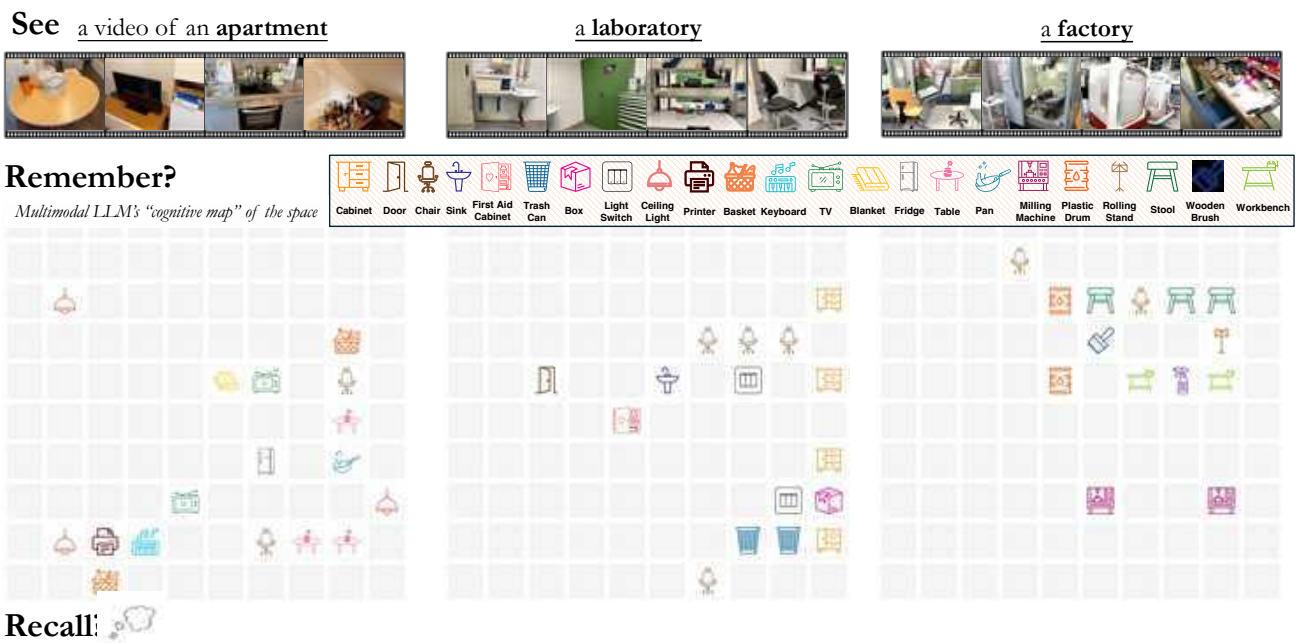


Figure 1. Whether at home, in the workplace, or elsewhere, the ability to perceive a space, remember its layout, and retrieve this spatial information to answer questions on demand is a key aspect of visual-spatial intelligence. Recent Multimodal LLMs can understand general videos, but can they "think spatially" when presented with a video recording of an environment? Can they build an accurate, implicit "*cognitive map*" that allows them to answer questions about a space? What are the strengths and limitations of using MLLMs to enhance spatial intelligence? We dig into these questions by setting up video data for MLLMs to watch, building a VQA benchmark to check their recall, and examining what the MLLMs actually remember and understand.

Abstract

*Humans possess the visual-spatial intelligence to remember spaces from sequential visual observations. However, can Multimodal Large Language Models (MLLMs) trained on million-scale video datasets also "think in space" from videos? We present a novel video-based visual-spatial intelligence benchmark (VSI-Bench) of over 5,000 question-answer pairs, and find that MLLMs exhibit competitive—though subhuman—visual-spatial intelligence. We probe models to express how they think in space both linguistically and visually and find that while spatial reasoning capabilities remain the primary bottleneck for MLLMs to reach higher benchmark performance, local world models and spatial awareness do emerge within these models. Notably, prevailing linguistic reasoning techniques (*e.g.*, chain-of-thought, self-consistency, tree-of-thoughts) fail to improve performance, whereas explicitly generating cognitive maps during question-answering enhances MLLMs' spatial distance ability.*

*Equal contribution.

1. Introduction

When shopping for furniture, we often try to recall our living room to imagine if a desired cabinet will fit. Estimating distances is difficult, yet after even a single viewing, humans can mentally reconstruct spaces, recalling objects in a room, their positions, and sizes. We live in a sensory-rich 3D world where visual signals surround and ground us, allowing us to perceive, understand, and interact with it.

Visual-spatial intelligence entails perceiving and mentally manipulating spatial relationships [26]; it requires myriad capabilities, including relational reasoning and the ability to transform between egocentric and allocentric perspectives (Sec. 2). While Large Language Models (LLMs) [3, 6, 9, 35, 59, 65, 66, 75, 79, 80, 85, 100] have advanced linguistic intelligence, visual-spatial intelligence remains under-explored, despite its relevance to robotics [7, 8, 21, 62], autonomous driving [77], and AR/VR [12, 27, 53].

Multimodal Large Language Models (MLLMs) [1, 4, 15, 33, 41, 47, 47, 76], which integrate language and vision, exhibit strong capacities to think and reason in open-ended dialog and practical tasks like web agents [21, 28, 32, 34]. To advance this intelligence in the visual-spatial realm, we introduce `VSI-Bench`, a video-based benchmark featuring over 5,000 question-answer pairs across nearly 290 real indoor-scene videos (Sec. 3). Video data, by capturing continuous, temporal input, both parallels how we observe the world and enables richer spatial understanding and reasoning than static images. Evaluating open- and closed-source models on `VSI-Bench` reveals that even though a large performance gap exists between models and humans, MLLMs exhibit emerging visual-spatial intelligence despite the challenges of video understanding, textual understanding, and spatial reasoning (Sec. 4).

To analyze model behavior and inspired by dual-coding theory [18], which posits that linguistic and visual processing are distinct yet complementary, we prompt selected models for self-explanations (linguistic) and cognitive maps (visual). Analyzing the self-explanations reveals that spatial reasoning, as compared to visual perception, linguistic intelligence, or temporal processing, is the main factor behind weak performance on `VSI-Bench` (Sec. 5). "*Cognitive maps*", which represent internal layouts of environments [60, 78], allow us to evaluate MLLMs' implicit spatial world models and find that MLLMs build strong local models but weak global ones (Sec. 6). Furthermore, standard linguistic reasoning techniques fail to enhance performance on our benchmark. However, explicitly generating and using cognitive maps improves spatial distance question-answering.

Expressing visual-spatial intelligence is difficult (and often piecemeal), even for humans [26]. With this work, we aim to encourage the community to explore grounding frontier models with visual-spatial intelligence and to pave and illuminate this direction.

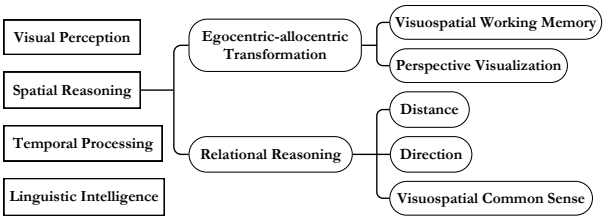


Figure 2. A taxonomy of **visual-spatial intelligence** capabilities.

2. Visual-Spatial Intelligence

We discuss preliminaries and scope visual-spatial intelligence to provide context and a framework for later analysis.

Term Use. We use "intelligence" rather than "cognition" as it is broader, and "spatial cognition" is a branch of cognitive psychology [81]. We prefix spatial intelligence in our work with "visual", as spatial intelligence exists irrespective of sensory modality (*e.g.*, a blind person can perceive space through other senses) [26]. Given our focus on video input, we discuss *visual*-spatial intelligence.

Investigation Scope. While classic spatial intelligence tests also include pen-paper tasks like the Mental Rotation Test [72], our focus is on visual-spatial intelligence as it applies to real-world environments, particularly in common spaces like homes, offices, and factories.

Taxonomy. We provide a taxonomy of capabilities potentially required for visual-spatial intelligence (Fig. 2), based on cognitive psychology [11, 26, 55, 60] and human experience with our benchmark tasks in Sec. 3. Visual perception, linguistic intelligence, temporal processing, and spatial reasoning are the four areas needed in `VSI-Bench`. For example, [11] shows that visual object and spatial processing are neurally distinct, which motivates "visual perception" and "spatial reasoning" as separate areas. We break spatial reasoning into two broad capabilities: relational reasoning and egocentric-allocentric transformation.

Relational reasoning is the ability to identify, via distance and direction, relationships between objects. It also encompasses reasoning about distance between objects by relying on visuospatial common sense about the sizes of other objects. For example, knowing a standard beverage can is approximately 12 cm tall, humans can estimate other object sizes by comparing visual proportions.

Egocentric-allocentric transformation involves shifting between a self-centered (egocentric) view and an environment-centered (allocentric) one. In our setting, each egocentric video frame maps to allocentric object positions and camera trajectory. When humans observe a space, they convert egocentric perceptions into an allocentric mental map, enabling perspective-taking from various viewpoints—essential for tasks like relative direction or route planning. This transformation relies on visualizing new perspectives and on visuospatial working memory [2], the abil-

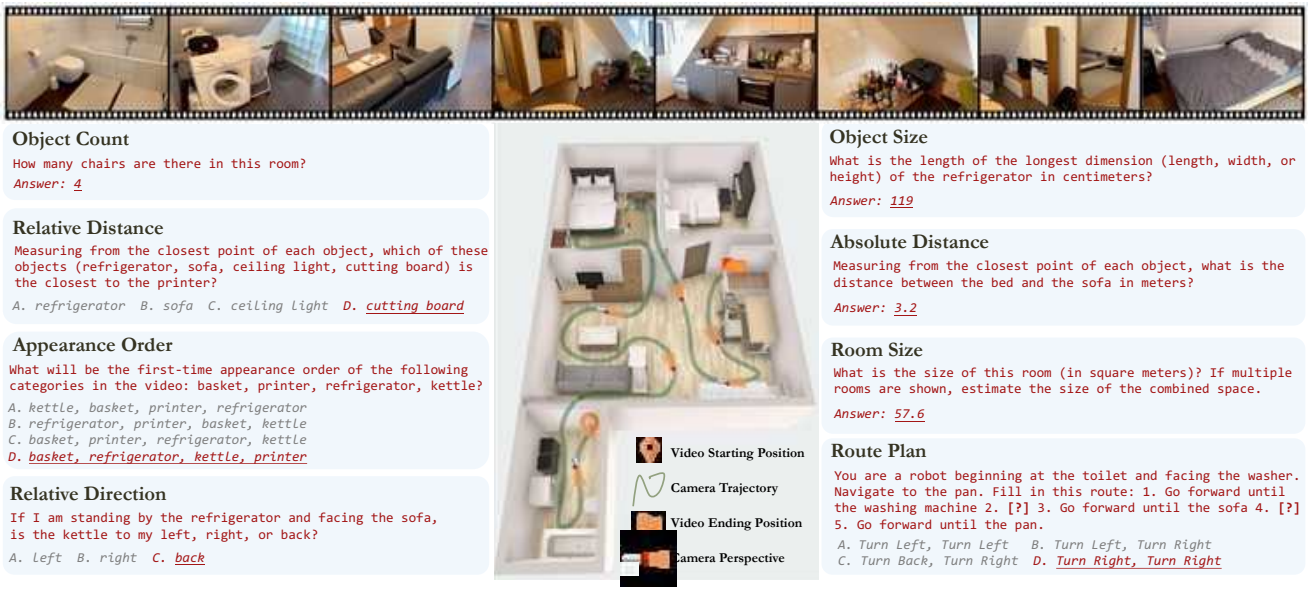


Figure 3. **Tasks demonstration of VSI-Bench**. Note: the questions above are simplified slightly for clarity and brevity.

ity to hold and manipulate spatial information, say by updating object positions from new egocentric input [20, 54].

Every task in VSI-Bench requires perceptual, linguistic, and temporal abilities and varying degrees of spatial reasoning. For example, egocentric-allocentric transformation is much more important for a task like route planning than object size estimation. These factors provide some context for the complexity of visual-spatial intelligence.

3. VSI-Bench

3.1. Overview

We introduce VSI-Bench to quantitatively evaluate the visual-spatial intelligence of MLLMs from egocentric video. VSI-Bench comprises over 5,000 question-answer pairs derived from 288 real videos. These videos are sourced from the validation sets of the public indoor 3D scene reconstruction datasets ScanNet [19], ScanNet++ [94], and ARKitScenes [5] and represent diverse environments—including residential spaces, professional settings (*e.g.*, offices, labs), and industrial spaces (*e.g.*, factories)—and multiple geographic regions. Repurposing these existing 3D reconstruction and understanding datasets offers accurate object-level annotations which we use in question generation and could enable future study into the connection between MLLMs and 3D reconstruction. VSI-Bench is high-quality, having been iteratively reviewed to minimize question ambiguity and to remove incorrect annotations propagated from the source datasets.

VSI-Bench includes eight tasks of three types: *configurational*, *measurement estimation*, and *spatiotemporal*. The configurational tasks (*object count, relative distance, relative direction, route plan*) test a model's understanding of the configuration of a space and are more intuitive for humans (see Sec. 4 for comparison between MLLM and human performance). Measurement estimation (of *object size, room size*, and *absolute distance*) is of value to any embodied agent. While predicting a measurement exactly is very difficult, for both humans and models, a better sense of distance and other measurements is intuitively correlated with better visual-spatial intelligence and underpins a wide range of tasks that require spatial awareness, like interaction with objects and navigation. Spatiotemporal tasks like *appearance order* test a model's memory of a space as seen in video. See Fig. 3 for an overview of VSI-Bench tasks and Fig. 5 for dataset statistics.

3.2. Benchmark Construction

We develop a sophisticated benchmark construction pipeline to effectively generate high-quality question-answer (QA) pairs at scale, as shown in Fig. 4.

Data Collection and Unification. We begin our dataset construction by standardizing various datasets into a unified meta-information structure, ensuring dataset-agnostic QA pair generation. Our benchmark aggregates existing 3D indoor scene understanding and reconstruction datasets: ScanNet [19], ScanNet++ [94], and ARKitScenes [5]. These datasets provide high-fidelity video scans capable of space reconstruction, ensuring MLLMs can answer space-level questions with only video input. Additionally, their object-level 3D annotations facilitated our question generation. We parse the datasets into a unified meta-information format including object categories, bounding boxes, video specifications (resolution and frame rate), and more.

Question-Answer Generation. QA pairs are primarily auto-annotated using the meta-information and question templates; the *route plan* task was human-annotated. We sophisticatedly design and refine the question template for each task and provide guidelines for human annotators. For

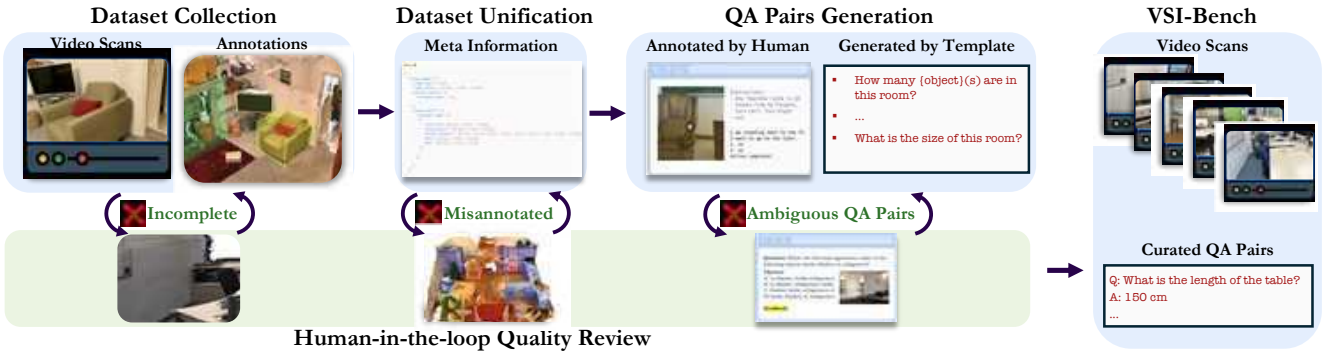


Figure 4. **Benchmark curation pipeline.** The pipeline first unifies diverse datasets into a standardized format and semantic space for consistent processing. QA pairs are then generated through both human annotation and question templates. To ensure quality, human verification is implemented at all key stages for filtering low-quality videos, annotations, and ambiguous QA pairs.

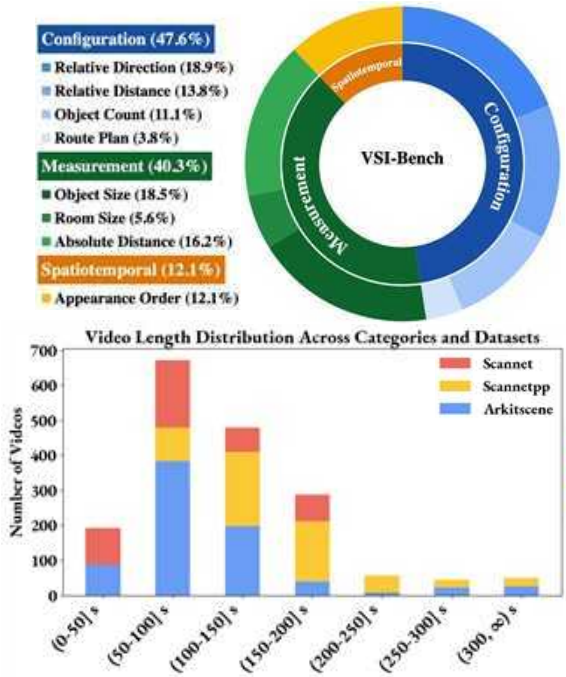


Figure 5. **Benchmark Statistics**. **Top**: The distribution of tasks across three main categories. **Bottom**: The video length statistic.

more detailed design, see Appendix B.1.

Human-in-the-loop Quality Review. Despite human-annotated data sources and a meticulously designed QA generation methodology, certain ambiguities and errors inevitably persist, primarily due to inherent annotation errors in the source datasets. We implement a human-in-the-loop verification protocol spanning benchmark construction. This iterative quality assurance is bidirectional: when evaluators flag ambiguous or erroneous questions, we trace the error source and remove the problematic data sample or modify the meta-information, question template, or QA generation rule accordingly to rectify other erroneous questions stemming from the same source. Following each human review cycle, we update and iterate the benchmark until it satisfies our quality standards.

4. Evaluation on `VSI-Bench`

4.1. Evaluation Setup

Benchmark Models. We comprehensively evaluate 15 video-supporting MLLMs across diverse model families, encompassing various parameter scales and training recipes. For proprietary models, we consider Gemini-1.5 [76] and GPT-4o [33]. For open-source models, we evaluate models from InternVL2 [14], ViLA [44], LongViLA [88], LongVA [98], LLaVA-OneVision [39], and LLaVA-NeXT-Video [99]. All evaluations are conducted under zero-shot settings and using each model's default prompts. To ensure reproducibility, we use greedy decoding for all models.

Metric Design. Based on whether the ground-truth answer is verbal or numerical, our tasks are suited to either a Multiple-Choice Answer (MCA) or Numerical Answer (NA) format (see Fig. 3). For MCA tasks, we follow standard practice [24, 30, 96] by using *Accuracy* ($\mathcal{ACC}$), based on exact matching (with possible fuzzy matching), as the primary metric. For NA tasks, where models predict continuous values, accuracy via exact matching fails to capture the degree of proximity between model predictions and ground-truth answers. Therefore, we introduce a new metric, *Mean Relative Accuracy* ($\mathcal{MRA}$), inspired by previous works [22, 45, 71]. Specifically, for a NA question, given a model's prediction $\hat{y}$, ground truth y, and a confidence threshold θ, relative accuracy is calculated by considering $\hat{y}$ correct if the relative error rate, defined as $|\hat{y}-y|/y$, is less than $1-\theta$. As single-confidence-threshold accuracy only considers relative error in a narrow scope, $\mathcal{MRA}$ averages the relative accuracy across a range of confidence thresholds $\mathcal{C} = \{0.5, 0.55, \ldots, 0.95\}$:

$$\mathcal{MRA} = \frac{1}{10}\sum_{\theta\in\mathcal{C}} \mathbb{1}\left(\frac{|\hat{y}-y|}{y} < 1-\theta\right). \quad (1)$$

$\mathcal{MRA}$ offers a more reliable and discriminative measurement for calculating the similarity between numerical predictions and ground truth values.

Chance Level Baselines. We provide two baselines:

Methods	Rank	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
			Numerical Answer				Multiple-Choice Answer			
Baseline										
Chance Level (Random)	-	-	-	-	-	-	25.0	36.1	28.3	25.0
Chance Level (Frequency)	-	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
VSI-Bench (tiny) Perf.										
†Human Level	-	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
†Gemini-1.5 Flash	-	45.7	50.8	33.6	56.5	45.2	48.0	39.8	32.7	59.2
†Gemini-1.5 Pro	-	48.8	49.6	28.8	58.6	49.4	46.0	48.1	42.0	68.0
†Gemini-2.0 Flash	-	45.4	52.4	30.6	66.7	31.8	56.0	46.3	24.5	55.1
Proprietary Models (API)										
GPT-4o	3	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	2	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	1	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-source Models										
InternVL2-2B	11	27.4	21.8	24.9	22.0	35.0	33.8	44.2	30.5	7.1
InternVL2-8B	5	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	3	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
LongVILA-8B	12	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5
VILA-1.5-8B	9	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8
VILA-1.5-40B	7	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
LongVA-7B	8	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7
LLaVA-NeXT-Video-7B	4	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B	1	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
LLaVA-OneVision-0.5B	10	28.0	46.1	28.4	15.4	28.3	28.9	36.9	34.5	5.8
LLaVA-OneVision-7B	6	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	2	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6

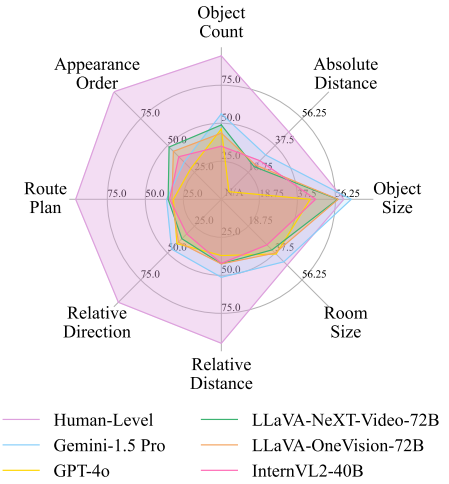


Table 1. **Evaluation on VSI-Bench. Left:** Dark gray indicates the best result among all models and light gray indicates the best result among open-source models. † indicates results on VSI-Bench (tiny) set. **Right:** Results including the top-3 open-source models.

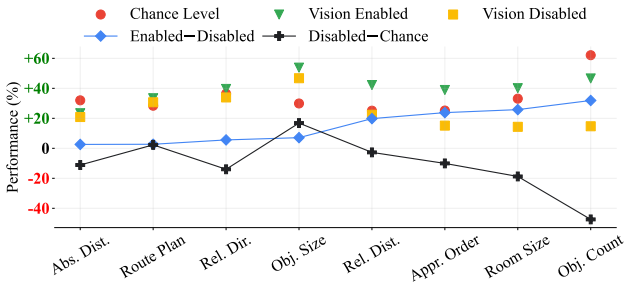


Figure 6. **Performance comparisons between Vision Enabled (w/ video), Vision Disabled (w/o video) and Chance Level (Freq.).** Enabled−Disabled indicates the gap between Vision Enabled and Vision Disabled, and Disabled−Chance betokens the gap between Vision Disabled and Chance Level (Freq.). Tasks are sorted by Enable−Disable for better understanding.

- *Chance Level (Random)* is the random selection accuracy for MCA tasks (and is inapplicable for NA tasks).
- *Chance Level (Frequency)* represents the highest performance MLLMs would achieve by always selecting the most frequent answer for each task. This identifies performance gains that may result from inherently long-tailed answers or imbalanced multiple-choice distributions.

Human Level Performance. We randomly sample a subset of 400 questions (50 per task), which we will refer to as VSI-Bench (tiny). Human evaluators independently answer each question, and their performance is evaluated using the above-mentioned metrics. For comparison, we also report Gemini-1.5 Pro's performance on VSI-Bench (tiny). See Appendix C for details on evaluation setups.

4.2. Main Results

Tab. 1 shows overall model performance on VSI-Bench. Our key observations are as follows:

Human Level Performance. Not surprisingly, human evaluators achieve 79% average accuracy on our benchmark, outperforming the best model by 33%. Notably, human performance on configuration and spatiotemporal tasks is remarkably high, ranging from 94% to 100%, indicating human intuitiveness. In contrast, the performance gap between humans and the best MLLM is much narrower on the three measurement tasks that require precise estimation of absolute distance or size, suggesting that MLLMs may have a relative strength in tasks requiring quantitative estimation.

Proprietary MLLMs. Despite a significant performance gap with humans, the leading proprietary model, Gemini-1.5 Pro, delivers competitive results. It surpasses the chance level baselines by a substantial margin and manages to approach human level performance in tasks such as absolute distance and room size estimation. It's worth noting that while human evaluators have years of experience in understanding the physical world spatially, MLLMs are only trained on 2D digital data like internet videos.

Open-source MLLMs. Top-tier open-source models like LLaVA-NeXT-Video-72B and LLaVA-OneVision-72B demonstrate highly competitive performance to closed-source models, trailing the leading Gemini-1.5 Pro by only 4% to 5%. However, the majority of open-source models (7/12) perform below the chance level baseline, indicating significant limitations in their visual-spatial intelligence.

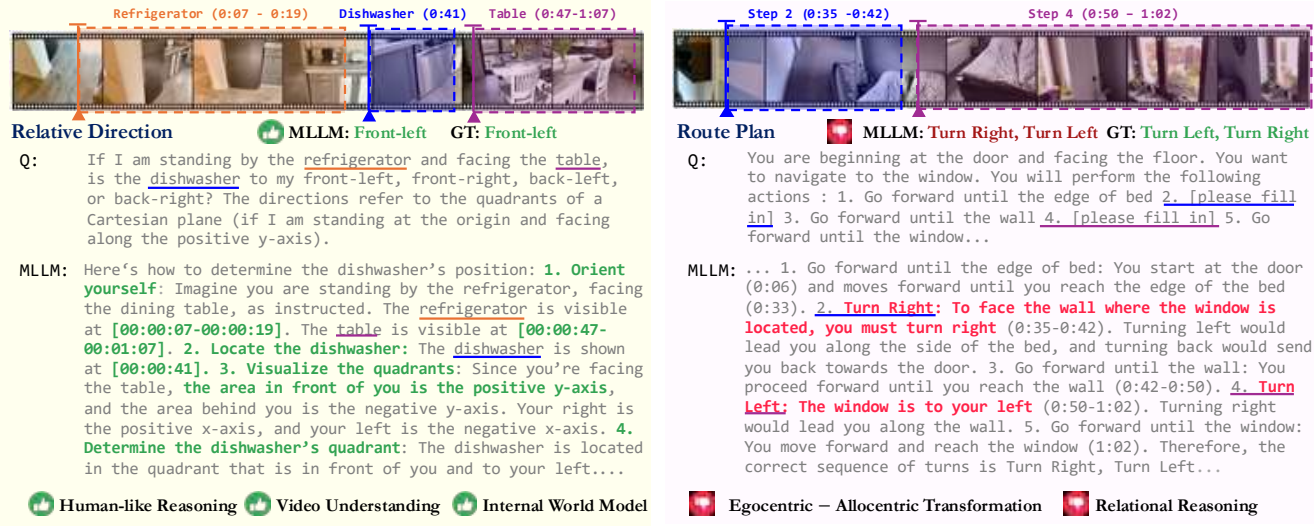


Figure 7. **Examples of how a MLLM thinks as seen in self-explanations**. While a MLLM exhibits strong video understanding and linguistic reasoning capabilities, its spatial reasoning capabilities are still developing.

Blind Evaluation. We compare MLLMs' performance against "Chance Level (frequency)" and "Vision Disabled" (blind) results, using averages across six of the strongest models (3 open-source and 3 closed-source). As shown in Fig. 6, the consistent improvements in "Enabled−Disabled" and general degradation in "Disabled−Chance" demonstrate that video is essential and beneficial for our VSI-Bench, with blind models performing below chance level. Meanwhile, MLLMs struggle to improve beyond chance level in the absolute distance estimation, route plan, and relative direction tasks, whether vision is enabled or not, underscoring the difficulty of these tasks. Note that on object size, "Vision Disabled" models already significantly outperform chance level, likely due to common-sense knowledge learned during language model training.

5. How MLLMs *Think in Space* Linguistically

To better understand when and why models succeed or fail and to elucidate the facets of visual-spatial intelligence they possess, we examine how MLLMs *think in space* linguistically here and visually in Sec. 6. We begin by prompting the best-performing MLLM in VSI-Bench, Gemini-1.5 Pro [76], to articulate its internal reasoning in language.

5.1. Probing via Self-Explanations

Self-explanations are a prevailing approach on par with traditional model explanations like LIME saliency maps [69] for understanding LLM-generated responses [25, 31, 51] and are widely used in analyzing language model behavior [64, 96]. We randomly sample a subset of 163 incorrect answers, prompt the MLLM to provide explanations for the predicted answers, and carefully review them by hand.

Case Studies. Fig. 7 presents self-explanations in both a success and an error case. In both examples, when thinking in space, the MLLM exhibits advanced video understanding, demonstrated by the impressive accuracy of its timestamped descriptions. The model also forms correct step-by-step reasoning processes, outlining steps such as "orient yourself", "locate the dishwasher" and "visualize the quadrants" for the relative direction task. Furthermore, the construction of a global coordinate system (Fig. 7, left) suggests that MLLMs may possess or build an implicit world model. Rather than using isolated frames, short clips, or random guesses, the MLLM used global spatial context and reasoning to infer correctly.

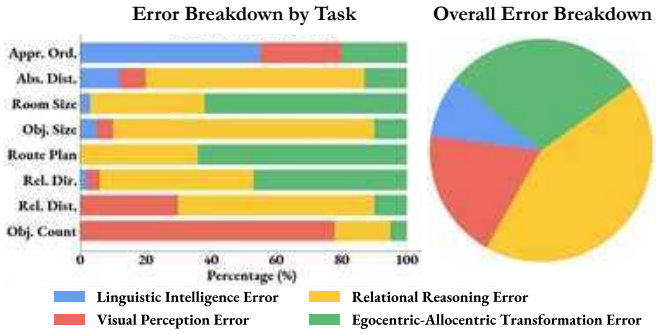


Figure 8. Human-conducted analysis of errors by type. Over 70% of errors stem from faulty spatial reasoning capabilities.

In the incorrect example (Fig. 7, right), we can identify faulty visual-spatial capabilities like egocentric-allocentric transformation and relational reasoning, as introduced in Fig. 2. In the video, the camera pans right to shift the view from the edge of the bed to the wall and window. The model obeys this egocentric view, responding that "to face the wall where the window is located, you must turn right" instead of creating an allocentric view reflecting the reality that the route from door to bed means turning left.

Error Analysis. To quantify and identify the main bottleneck for the best-performing MLLM on our benchmark, we analyze its errors on VSI-Bench (tiny), categorizing them

into four distinct types which arose from both our outlined visual-spatial capabilities (Fig. 2) and a clear four-way categorization of errors upon examination:

1. **Visual perception error**, stemming from unrecognized objects or misclassified object categories;
2. **Linguistic intelligence error**, caused by logical, mathematical reasoning, or language understanding defects;
3. **Relational reasoning error** includes errors in spatial relationship reasoning, *i.e.*, distance, direction, and size;
4. **Egocentric-allocentric transformation error**, resulting from an incorrect allocentric spatial layout or improper perspective-taking.

As shown in Fig. 8, around 71% of errors are attributed to spatial reasoning (as ontologically conceived in Fig. 2), which suggests that:

> *Spatial reasoning is the primary bottleneck for MLLM performance on* `VSI-Bench`.

Further analysis and case studies are in Appendix E.2.

5.2. Limits of CoT Methods in Visuospatial Tasks

Prompting techniques improve the reasoning and problem-solving abilities for large models across diverse tasks [32, 34, 73, 82]. Their successes motivate us to investigate whether these linguistic prompting methods could also improve the visual-spatial capabilities of MLLMs in `VSI-Bench`. We investigate three prevailing prompting techniques (see Appendix B.3 for more details):

- *Zero-Shot Chain-of-Thought (CoT).* Following [37, 86], we add "Let's think step by step" to the prompts.
- *Self-Consistency w/ CoT.* We follow [84] and set the MLLM's temperature to 1.0 to encourage diverse reasoning and then take the majority consensus from five runs (employed w/ Zero-Shot CoT) as the final prediction.
- *Tree-of-Thoughts (ToT).* Following the "Creative Writting" practice in [92], we divide reasoning into plan generation and answer prediction. The MLLM first drafts and selects a plan, then generates three candidate answers and selects the most confident one as prediction.

As shown in Fig. 9, surprisingly, all three linguistic reasoning techniques lead to performance degradation on `VSI-Bench`. Zero-Shot CoT and ToT reduce average performance by about 4%, and self-consistency, though slightly better, still falls 1.1% below the no-prompting baseline. The unilateral improvement in the appearance order and absolute distance estimation tasks is easily explained by their significant percentage of linguistic intelligence errors (see Fig. 8). In contrast, the room size and object size tasks suffer a large 8% to 21% decrease, showing that encouraging a model to think more can be not just unreliable but downright harmful. Meanwhile, as shown in Tab. 2, ZeroShot CoT achieves a 1.6% improvement on the general video understanding benchmark VideoMME [24]. Therefore, our results suggest that:

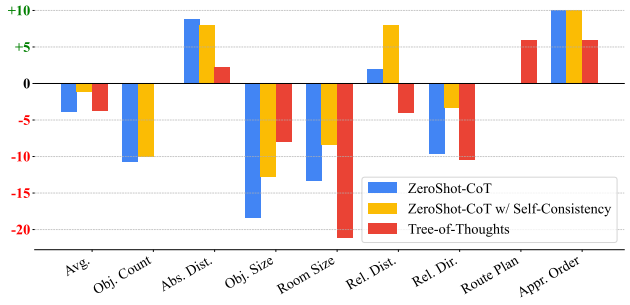


Figure 9. Relative improvements of *CoT*, *self-consistency* and *Tree-of-Thought* compared to the baseline. All three prevailing prompting techniques fail on average on our benchmark, and, in some cases, task performance becomes *much worse* after applying them. This implies that `VSI-Bench` cannot be solved by solely improving linguistic capabilities.

Case	Performance
Gemini-1.5 Pro (w/o CoT)	77.2
Gemini-1.5 Pro (w/ CoT)	79.8

Table 2. **Gemini-1.5 Pro CoT performance on a 500-questions subset in VideoMME.**

> *Linguistic prompting techniques, although effective in language reasoning and general visual tasks, are harmful for spatial reasoning.*

6. How MLLMs *Think in Space* Visually

Since humans subconsciously build mental representations [58, 78] of space when reasoning spatially, we explore how MLLMs remember spaces.

6.1. Probing via Cognitive Maps

We prompt MLLMs to express their internal representations of the spaces they see using cognitive maps, a well-established framework for remembering objects in a set environment [60, 78]. We prompt the best-performing MLLM, Gemini-1.5 Pro, to predict object center positions within a 10 $\times$ 10 grid based on video input (see Fig. 11b for grid size ablation and Appendix B.4 for prompt). We present examples of the resulting cognitive maps in Fig. 10.

To quantitatively assess these cognitive maps, we evaluate the Euclidean distance between all pairs of objects within each map. We consider the distance (on the grid) between two objects to be correct if it deviates by no more than one grid unit from the distance in the ground truth cognitive map. As shown in Fig. 11, we divide the map-distances into eight distinct bins for analysis. Interestingly, we find that the MLLM achieves a remarkable 64% accuracy in positioning adjacent objects within its cognitive map, indicating robust *local* spatial awareness. However, this accuracy significantly deteriorates as the distance between two objects

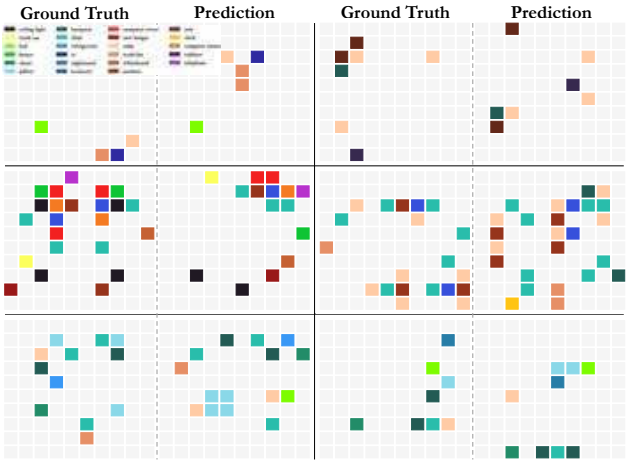


Figure 10. Visualization of cognitive maps from MLLM and GT.

increases, which suggests that:

> *When remembering spaces, a MLLM forms a series of local world models in its mind from a given video, rather than a unified global model.*

This observation aligns with the challenge of forming a global space representation from discrete video frames, which is inherently difficult for MLLMs. While this task may not be trivial for humans either, it is likely that they can build such global space representations more accurately.

6.2. Better Distance Reasoning via Cognitive Maps

Given the local awareness of MLLMs in remembering spaces (see Fig. 10 and Fig. 11) and the importance of mental imagery to how humans think in space, we investigate whether generating and using cognitive maps can help MLLMs' spatial reasoning in terms of `VSI-Bench`'s relative distance task. This tests if the local distance awareness emerged through cognitive maps transfers to improved distance recall and reasoning.

Case	Rel. Dist Acc.
w/o Cog. map	46.0
w/ Cog. map	56.0
w/ Cog. map (GT)	66.0

(a) Cognitive map prompting.

Cog. Map Src.	Size	Rel. Dist Acc.
MLLM	10×10	56.0
MLLM	20×20	54.0
GT	10×10	66.0
GT	20×20	78.0

(b) Cognitive map canvas size.

Table 3. **Relative distance task with cognitive map.**

We prompt Gemini-1.5 Pro to first generate a cognitive map based on the given video and question, and then to use the predicted map to answer the question. As shown in Tab. 3a, we find that using mental imagery improves a MLLM's relative distance accuracy by 10%. The 20% to 32% gain over baseline with the ground truth cognitive map underscores the importance of building accurate mental maps of a scene, which enforce a globally consistent topology, but indicates that such mental imagery is only one part of the puzzle, albeit a crucial one. These results point to building a mental spatial world model or cognitive map as a valuable pretext task or a promising solution for MLLMs to tackle visual-spatial reasoning.

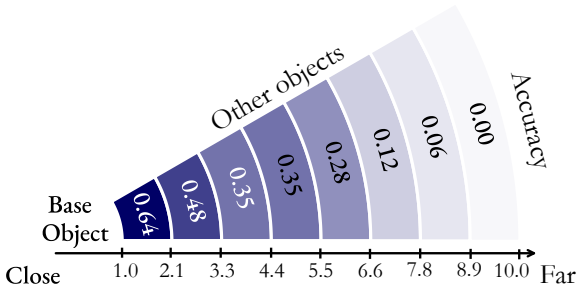


Figure 11. **Locality of the MLLM's predicted cognitive maps.** The MLLM's map-distance accuracy decreases dramatically with increasing object distance.

7. Related Works

Apart from visual-spatial intelligence in Sec. 2, we further ground our work in the following two related areas:

MLLMs with Visual-Spatial Awareness. Building on the powerful language and reasoning abilities of LLMs [3, 9, 65, 66, 75, 79, 80] and the feature extraction abilities of modern vision encoders [29, 63, 67], MLLMs, especially visual MLLMs, exhibit unprecedented visual understanding capabilities [33, 39, 76, 83, 88, 99], a promising direction toward developing world models [48] and embodied agents [17, 21, 36, 57]. However, grounding MLLMs in the real world presents significant challenges for models' visual-spatial intelligence, motivating recent efforts [10, 13, 16, 40, 46, 91, 102]. Unlike prior works, which primarily focus on understanding spatial information through 2D images [68, 74, 90] or solely language [56, 70, 87, 87, 89], our work assesses models' visual spatial intelligence using real-world videos, which more closely mirrors human understanding of the world and application scenarios for embodied agents.

Benchmarking MLLMs on Video. With MLLMs displaying impressive performance on still-images across perception, reasoning, and multi-disciplinary tasks [38, 50, 95, 96], there is increasing interest in evaluating MLLMs' video understanding capabilities [23, 24, 42, 43, 49, 52, 53, 61, 93]. For example, Video-MME [24] comprehensively evaluates MLLMs across various video-related tasks, including recognition and perception. EgoSchema [53] and OpenEQA [62] evaluate MLLMs' understanding abilities using egocentric videos. Despite their significance, most prior works focus on content-level understanding [24, 42, 53, 61], which primarily serves as a temporal extension of 2D image understanding without 3D spatial consideration. Extending beyond prior benchmarks, our work establishes a testbed evaluating models' 3D video-based visual-spatial intelligence, using video as an interface to understand the real world.

8. Discussion and Future Work

We study how models see, remember, and recall spaces by building `VSI-Bench` and investigating the performance and behavior of MLLMs on it. Our analysis of how MLLMs think in space linguistically and visually identifies existing strengths (*e.g.*, prominent perceptual, temporal, and linguistic abilities) and bottlenecks for visual-spatial intelligence (*e.g.*, egocentric-allocentric transformation and relational reasoning). While prevailing linguistic prompting methods fail to improve spatial reasoning, building explicit cognitive maps does enhance the spatial distance reasoning of MLLMs. Future avenues of improvement include task-specific fine-tuning, developing self-supervised learning objectives for spatial reasoning, or visuospatial-tailored prompting techniques for MLLMs.

Acknowledgements. We thank Ellis Brown, Ryan Inkook Chun, Youming Deng, Oscar Michel, Srivats Poddar, Xichen Pan, Austin Wang, Gavin Yang, and Boyang Zheng for their contributions as human annotators and evaluators. We also thank Fred Lu for proofreading our manuscript. We also thank Chen Feng, Richard Tucker, Noah Snavely, Leo Guibas, and Rob Fergus for their helpful discussions and feedback. This work was mainly supported by the Open Path AI Foundation, Google TPU Research Cloud (TRC) program, and the Google Cloud Research Credits program (GCP19980904). S.X. thanks the OpenAI Researcher Access program and an Amazon Research award for their support.

References

[1] 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. *NeurIPS*, 2022. 2

[2] Alan Baddeley. Working memory. *Science*, 255(5044): 556–559, 1992. 2

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2, 8

[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. 2

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS*, 2021. 3, 13

[6] Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *arXiv preprint arXiv:2305.00948*, 2023. 2

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 2

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *RSS*, 2023. 2

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 2020. 2, 8

[10] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 8

[11] Christopher F Chabris, Thomas E Jerde, Anita W Woolley, Margaret E Gerbasi, Jonathon P Schuldt, Sean L Bennett, J Richard Hackman, and Stephen M Kosslyn. Spatial and object visualization cognitive styles: Validation studies in 3800 individuals. *Group brain technical report*, 2:1–20, 2006. 2

[12] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. In *NeurIPS*, 2024. 2

[13] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 8

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

[15] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2

[16] 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 *NeurIPS*, 2024. 8

[17] Junmo Cho, Jaesik Yoon, and Sungjin Ahn. Spatially-aware transformers for embodied agents. In *ICLR*, 2023. 8

[18] James M. Clark and Allan Paivio. Dual coding theory and education. *Educational Psychology Review*, 3(3):149–210, 1991. 2

[19] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 3, 13

[20] Milton J. Dehn. *Working Memory and Academic Learning: Assessment and Intervention*. John Wiley & Sons, 2011. 3

[21] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *ICML*, 2023. 2, 8

[22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 4

[23] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. In *NeurIPS*, 2024. 8

[24] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 4, 7, 8

[25] Haoyu Gao, Ting-En Lin, Hangyu Li, Min Yang, Yuchuan Wu, Wentao Ma, Fei Huang, and Yongbin Li. Self-explanation prompting improves dialogue understanding in large language models. In *COLING*, 2024. 6

[26] Howard Gardner. *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, tenth-anniversary edition, second paperback edition edition, 1983. 2

[27] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2

[28] Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. In *ICLR*, 2024. 2

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 8

[30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. 4

[31] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*, 2023. 6

[32] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *ICML*, 2022. 2, 7

[33] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 4, 8

[34] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *ICLR*, 2024. 2, 7

[35] Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schütze, and Peter Clark. Language models with rationality. In *EMNLP*, 2023. 2

[36] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. In *CoRL*, 2024. 8

[37] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 7, 15

[38] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024. 8

[39] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4, 8

[40] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. *arXiv preprint arXiv:2406.02537*, 2024. 8

[41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2

[42] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 8

[43] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. *arXiv preprint arXiv:2311.17404*, 2023. 8

[44] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 4

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4

[46] Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*, 2024. 8

[47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 2, 16

[48] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 8

[49] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of ACL*, 2024. 8

[50] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2025. 8

[51] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *ACL*, 2023. 6

[52] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, 2024. 8

[53] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 2023. 2, 8

[54] Julia McAfoose and Bernhard T. Baune. Exploring visual–spatial working memory: A critical review of concepts and models. *Neuropsychology Review*, 2009. 3

[55] Chiara Meneghetti, Laura Miola, Tommaso Feraco, Veronica Muffato, and Tommaso Feraco Miola. Individual differences in navigation: an introductory overview. *Prime archives in psychology*, 2022. 2

[56] Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval. *NeurIPS*, 2024. 8

[57] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *NeurIPS*, 2024. 8

[58] Lynn Nadel. *The Hippocampus and Context Revisited*. Oxford University Press, 2008. 7

[59] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023. 2

[60] Nora S. Newcombe. *Spatial Cognition*. MIT Press, 2024. https://oecs.mit.edu/pub/or750iar. 2, 7

[61] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023. 8

[62] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 2, 8

[63] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 8

[64] Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *ACL*, 2024. 6

[65] Alec Radford. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. 2, 8

[66] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 8

[67] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 8

[68] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024. 8

[69] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *KDD*, 2016. 6

[70] Julia Rozanova, Deborah Ferreira, Krishna Dubba, Weiwei Cheng, Dell Zhang, and Andre Freitas. Grounding natural language instructions: Can large language models capture spatial information? *arXiv preprint arXiv:2109.08634*, 2021. 8

[71] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA, 1986. 4

[72] Shenna Shepard and Douglas Metzler. Mental rotation: effects of dimensionality of objects and type of task. *Journal of experimental psychology: Human perception and performance*, 14(1):3, 1988. 2

[73] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023. 7

[74] 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. 8

[75] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 8

[76] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent,

Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2, 4, 6, 8, 16

[77] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *CoRL*, 2024. 2

[78] E. C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. 2, 7

[79] 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. 2, 8

[80] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 8

[81] David Ed Waller and Lynn Ed Nadel. *Handbook of spatial cognition.* American Psychological Association, 2013. 2

[82] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *TMLR*, 2023. 7

[83] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[84] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 7, 15

[85] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *TMLR*, 2022. 2

[86] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022. 7, 15

[87] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Visualization-of-thought elicits spatial reasoning in large language models. *NeurIPS*, 2024. 8

[88] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 4, 8

[89] Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models. *TMLR*, 2024. 8

[90] 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. 8

[91] Jihan Yang, Runyu Ding, Ellis Brown, Xiaojuan Qi, and Saining Xie. V-irl: Grounding virtual intelligence in real life. In *ECCV*, 2024. 8

[92] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2024. 7, 15

[93] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mm-ego: Towards building egocentric multimodal llms. *arXiv preprint arXiv:2410.07177*, 2024. 8

[94] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 3, 13

[95] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ICML*, 2024. 8

[96] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 4, 6, 8

[97] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 16

[98] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 4

[99] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 4, 8

[100] Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. Unveiling linguistic regions in large language models. In *ACL*, 2024. 2

[101] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 13

[102] 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. 8

A. Appendix Outline

In these supplementary materials, we provide:

- Technical details about VSI-Bench construction and our linguistic and visual analysis (Appendix B);
- Evaluation setup and full evaluation results for VSI-Bench sub-experiments (Appendix C);
- Analysis on input sequencing and repetition (Appendix D);
- Additional visualization results (Appendix E).

B. Technical Details for VSI-Bench Construction and Analysis

In this section, we provide more technical details on the construction of VSI-Bench and analyzing MLLM thinking via self-explanations, Chain-of-Thought-based methods, and cognitive maps.

B.1. VSI-Bench Construction Pipeline

Here, we discuss the concrete setup for each stage in the benchmark construction pipeline.

Dataset Collection and Unification. We curate our evaluation dataset by collecting 150 samples from ARKitScenes [5], 50 samples from ScanNet++ [94], and 88 samples from ScanNet [19]. For video processing, we convert ScanNet's individual frames into continuous videos at 24 FPS, while subsampling ScanNet++ and ARKitScenes videos to 30 FPS. All videos are standardized to a resolution of 640 $\times$ 480 pixels. Given that ARKitScenes contains videos with varying orientations, we normalize their rotation to maintain a consistent upward orientation across all samples.

Due to varying annotation structures across the three datasets, we unify them into a standardized meta-information format for each scene with the following attributes: *dataset*, *video path*, *room size*, *room center*, *object counts*, and *object bounding boxes*. The room size is calculated by the Alpha shape algorithm* with the scene's point cloud. The room center is calculated as the geometric center of the minimal bounding box of the scene's point cloud. Object counts record the number of instances for each category. As for the object bounding boxes, we unify different annotation formats to the format of OrientedBoundingBox in Open3D [101].

For the categories included in the meta-information, we carefully curate a subset of categories from the three source datasets. Since our benchmark aims to evaluate the visual-spatial intelligence of MLLMs, we exclude both rare categories and those with extremely small object sizes to reduce perceptual challenges. Additionally, we implement category remapping to ensure vocabulary consistency and intuitive understanding across the benchmark. This category remapping is also iteratively refined during human review.

QA-Pair Generation. Each QA-pair contains the following attributes: *question ID*, *source dataset*, *task type*, *video path*, *question*, *multiple-choice options w/ letter answer*, and *verbal or numerical ground truth*. Of the eight tasks in VSI-Bench, the QA-pairs for seven tasks are derived from the unified meta-information and the Route Plan QA-pairs from human-annotated routes.

We evaluate the multiple-choice answer (MCA) tasks via accuracy and the numerical-answer (NA) tasks via mean relative accuracy ($\mathcal{MRA}$), but our VQA dataset also includes generated multiple-choice options and letter answers for the NA tasks. The generated multiple-choice options are sampled between a lower and upper bound factor of the ground truth numerical answer and are re-sampled if any two options are within a given threshold of each other. We sub-sample the number of questions for each scene for each task to prevent over-representation of any scene or task and to create a more balanced dataset. For MCA tasks, the letter answers are distributed as uniformly as possible.

For the *object counting* task, objects with counts of one are not included. For the *relative distance* task, only unique-instance objects are used for the primary category; multiple-instance objects are allowed for the object choices. If there are multiple instances of an object category, the minimum absolute distance to the primary object is used. If any of the four option distances are within a threshold (30 cm for rooms with size greater than 40 sq m, 15 cm otherwise) of each other, the question is considered ambiguous. For the *relative direction* task, to make sure the direction is clear, questions are considered ambiguous if they violate lower and upper bounds on the distance between any two objects or a threshold for proximity to angle boundaries. For the *appearance order* task, first appearance is considered to be the timestamp where the number of object pixels cross a set threshold, and timestamps too close together are considered ambiguous. For the *object size* task, the ground truth is taken as the longest dimension of the unique object's bounding box. For the *room size* task, room size is calculated by the alpha shape algorithm, as specified earlier. For the *absolute distance* task, we first uniformly sample points within the bounding boxes of the two objects. The distance is the minimum Euclidean distance among pairwise points. For the *route planning* task, humans construct routes given a template and instructions to choose any two unique objects as the start and end position, respectively, such that the route between them can be described in approximately two to five movements. Routes are comprised of two actions: "Go forward until [unique object]" and "Turn [left / right / back]". After collection, filtering and standardization are done. In the question, the "turn" directions are replaced with "[please fill in]".

*https://en.wikipedia.org/wiki/Alpha_shape

Task	Question Template
Object Counting	*How many {category}(s) are in this room?*
Relative Distance	*Measuring from the closest point of each object, which of these objects ({choice a}, {choice b}, {choice c}, {choice d}) is the closest to the {category}?*
Relative Direction	To create a comprehensive test of relative direction, three difficulty levels were created: • **Easy:** *If I am standing by the {positioning object} and facing the {orienting object}, is the {querying object} to the left or the right of the {orienting object}?* • **Medium:** *If I am standing by the {positioning object} and facing the {orienting object}, is the {querying object} to my left, right, or back? An object is to my back if I would have to turn at least 135 degrees in order to face it.* • **Hard:** *If I am standing by the {positioning object} and facing the {orienting object}, is the {querying object} to my front-left, front-right, back-left, or back-right? Directions refer to the quadrants of a Cartesian plane (assuming I am at the origin and facing the positive y-axis).*
Appearance Order	*What will be the first-time appearance order of the following categories in the video: {choice a}, {choice b}, {choice c}, {choice d}?*
Object Size	*What is the length of the longest dimension (length, width, or height) of the {category}, measured in centimeters?*
Absolute Distance	*Measuring from the closest point of each object, what is the direct distance between the {object 1} and the {object 2} (in meters)?*
Room Size	*What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.*
Route Plan	*You are a robot beginning at {the bed facing the tv}. You want to navigate to {the toilet}. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): {1. Go forward until the TV 2. [please fill in] 3. Go forward until the shower 4. [please fill in] 5. Go forward until the toilet.} You have reached the final destination.*

Table 4. **Question Templates for tasks in `VSI-Bench`.** We replace the highlighted part in the question template from scene to scene to construct our benchmark. Note that a complete example question is provided for Route Plan.

The question templates for the generation of each task are listed in Tab. 4.

Human-in-the-loop Quality Review. The quality review process occurs throughout two stages of our pipeline. During dataset collection, we manually filter the validation set by removing scenes with a high ratio of incomplete 3D mesh reconstruction that could misalign 3D annotations with visible video content. After generating scene meta-information, we manually verify its correctness, with a specific focus on ensuring the correctness of *object counts*.

In the QA pairs generation stage, we customize a web interface for human quality review. Human evaluators are asked to answer the benchmark questions without prior knowledge of the correct answers. They flag QA pairs where they believe the answers are incorrect. When evaluators identify ambiguous or erroneous questions, we trace the source of the errors and take corrective actions, such as removing problematic data samples or adjusting the meta-information, question templates, or modifying QA generation rules to prevent similar issues in the future. We iterate this procedure multiple times to ensure the quality.

B.2. Probing MLLM via Self-Explanations

Here, we provide more concrete implementations for the self-explanations and error analysis.

Self-Explanations. To conduct error analysis on a model's reasoning chains behind its predictions, we explicitly extract the reasoning chains that support the model's question-answering process. Specifically, after the model predicts an answer to a given question, it is further prompted with *"Please explain your answer step by step."* to generate the internal rationale leading to its prediction. It is important to note that this process is fundamentally different from *Chain-of-Thought* reasoning, where the model is asked to generate reasoning chains first and then predict the answer.

Error Analysis. For error analysis, we manually review within `VSI-Bench` (tiny) all error cases for tasks in multiple-choice answers and the bottom half of the worst-performing cases for tasks in numerical answers, which totals 163 samples. For each error case, human examiners are required to classify its primary error into one of four primary categories: *visual perception error*, *linguistic intel-*

ligence error, *relational reasoning error*, and *egocentric-allocentric transformation error*. If an incorrect prediction is attributed to multiple reasons, it is proportionally assigned as $\frac{1}{n}$ to each applicable category, where n is the number of error categories.

B.3. Implementation Details of CoT Methods

As detailed in our paper, we evaluate several advanced linguistic prompting methods on our benchmark, including *Chain-of-Thought*, *Self-Consistency*, and *Tree-of-Thoughts*. In this section, we elaborate on the implementation details of these three methods.

- *Chain-of-Thought* prompting. Following Zero-shot-CoT [37, 86], we append the phrase "Let's think step by step." to each question to elicit step-by-step reasoning from the large language model. The temperature, top-p, and top-k parameters are set to 0, 1, and 1, respectively. After the model generates its prediction, we initiate an additional turn of dialogue to prompt the model to extract its answer explicitly (*e.g.*, the letter corresponding to the correct option for multiple-choice questions or a numerical value for numerical questions). This approach mitigates errors arising from fuzzy matching.
- *Self-Consistency w/ CoT*. In line with Self-Consistency [84], we prompt MLLMs to generate multiple answers for a given question under Zero-shot-CoT [37] prompting. To encourage diversity among runs, we set the temperature to 0.7, top-p to 1, and top-k to 40. Initially, the model is prompted to provide an answer with step-by-step reasoning (using Zero-shot-CoT). As with Zero-shot-CoT, an additional dialogue turn is added to explicitly extract the prediction from the model's response. For each question, we perform 5 independent runs and take the majority prediction as the final answer.
- *Tree-of-Thoughts*. Inspired by the "Creative Writing" practice in [92], we divide the problem-solving process into two steps: plan generation and answer prediction. The temperature, top-p, and top-k parameters remain consistent with the Self-Consistency setup. For the plan generation step, we ask the model to generate 3 distinct plans to answer the given question. We then start a new dialogue and prompt the model to select the most promising plan based on the video, the question and the generated plans. This voting process is repeated 3 times, with the majority-selected plan chosen for the next step. In the answer prediction step, based on the video and the selected plan, the model is asked to predict the answer. Similar to the previous step, 3 independent predictions are generated, and the model votes 3 times to determine the most confident answer. A majority vote determines the final prediction.

Fig. 15. Fig. 16, and Fig. 17 illustrate these three prompting techniques and model outputs under the different strategies.

B.4. Cognitive Map

Generation. To generate the cognitive map for each video, we specify the target categories of interest and prompt the MLLM to predict the central position for each of these categories. The following prompt is used:

Cognitive Map Prompt

[Task]
This video captures an indoor scene. Your objective is to identify specific objects within the video, understand the spatial arrangement of the scene, and estimate the center point of each object, assuming the entire scene is represented by a 10x10 grid.

[Rule]
1. We provide the categories to care about in this scene: {categories_of_interest}. Focus ONLY on these categories.
2. Estimate the center location of each instance within the provided categories, assuming the entire scene is represented by a 10x10 grid.
3. If a category contains multiple instances, include all of them.
4. Each object's estimated location should accurately reflect its real position in the scene, preserving the relative spatial relationships among all objects.

[Output]
Present the estimated center locations for each object as a list within a dictionary. STRICTLY follow this JSON format: {"category name": [(x_1, y_1), ...], ...}

For the categories of interest, we include all potential categories as shown in Fig. 10 and Fig. 11. Such setup facilitates our focus on assessing the spatial awareness of the MLLM rather than its perceptual capabilities. In contrast, for benchmark tasks such as evaluating relative distance (as shown in Tab. 3), we restrict the provided categories to those explicitly mentioned in each question. This ensures that no additional information apart from the question is included.

Distance Locality Calculation. To quantitatively evaluate the cognitive maps, we measure inter-category distances as illustrated in Fig. 11. Specifically, for each category, we compute its Euclidean distance to all other categories. When a category contains multiple objects, we define the inter-category distance as the shortest distance between any two objects from the respective categories. We perform these distance calculations on both MLLM-predicted and ground truth cognitive maps and consider an MLLM's predicted distance between two categories to be correct if it

differs from the ground truth distance by no more than one grid unit. We apply this evaluation process across all cognitive maps and group the distances into eight bins to calculate the average accuracy on different bins.

C. Evaluation Details

C.1. General Evaluation Setup

Our evaluation processes are primarily conducted using the `LMMs-Eval` project [97]. To ensure reproducibility, unless otherwise specified, we adopt a greedy decoding strategy for all models (*i.e.*, the temperature is set to 0, and both top-p and top-k are set to 1). The input for the models is formatted as follows: `[Video Frames][Pre-prompt][Question][Post-prompt]`, where `Question` includes the question and any available options. The specific `Pre-prompt` and `Post-prompt` for different models and question types are detailed in Tab. 7.

C.2. Human Evaluation Setup

During the evaluation of human-level performance on `VSI-Bench` (tiny), human evaluators are allowed unlimited time to answer questions to the best of their ability. They receive both the questions and corresponding videos simultaneously and can review the videos multiple times to gather comprehensive information. We do not restrict the number of times evaluators can review videos for two key reasons. First, MLLMs auto-regressively generate answers, enabling them to analyze videos repeatedly during the response generation process. Second, MLLMs are designed to achieve and exceed typical human-level performance for practical real-world applications.

C.3. Number of Frames Setup

Typically, MLLMs subsample a fixed number of frames for evaluation. For all open-source models and the GPT-4 API, following [97], we manually sample video frames from the entire video at evenly spaced time intervals. For the Gemini API, we follow its instructions, uploading and feeding the entire video to the model. The number of frames used for each model are provided in Tab. 6.

C.4. More Evaluation Results

Here, we provide more evaluation results on our benchmark, including the full evaluation results of `VSI-Bench` (tiny), blind evaluation results, and vision-enabled − vision-disabled results.

`VSI-Bench` (tiny) Results. As shown in Tab. 8, we provide the evaluation results of all models on `VSI-Bench` (tiny). The rankings and average accuracy of MLLMs on `VSI-Bench` (tiny) remain consistent to the results reported in Tab. 1. This consistency suggests that the human evaluation and analysis results conducted on `VSI-Bench` (tiny) are reliable.

Order	Avg.
Video first	**48.8**
Question first	46.3

(a) Input Sequence

# Times	Avg.
1	48.8
2	**50.9**

(b) Video Repetition Times

Table 5. Ablations on the video input sequence and repetition.

Blind Evaluation. As shown in Tab. 9, we present the evaluation results for all MLLMs on `VSI-Bench`. Generally, larger variants within the same model family often demonstrate better performance in blind evaluations, as seen in comparisons such as Gemini-1.5 Flash *vs*. Gemini-1.5 Pro and VILA-1.5-8B *vs*. VILA-1.5-40B. The blind evaluation also highlights LLM biases across tasks. For instance, LongVILA-8B achieves 47.5% accuracy on the object count task, benefiting from a bias that frequently leads it to predict 2 as the answer.

Vision Enabled − Vision Disabled. Tab. 10 presents the improvement of MLLMs from using visual signals to answer `VSI-Bench`. Almost all MLLMs obtain improvements from visual signals, with notable improvements in tasks such as object count, room size, relative distance and appearance order.

D. Input Sequencing and Repetition Analysis

Human performance in visual problem-solving improves when they know the question before viewing the visual content, as it helps direct their attention to relevant visual cues. However, current MLLMs typically rely on a visual-first paradigm [47, 76], leading us to examine how the presentation order of video-question pairs impacts model performance. To investigate, we conduct experiments using Gemini-1.5 Pro on `VSI-Bench` (tiny).

MLLM's performance degrades with question-first paradigm. As shown in Tab. 5 (a), switching to a video-first approach results in a 2.5% decrease in overall performance for Gemini compared to the question-first approach.

MLLM benefits from multiple video views. In addition, humans often improve their VQA performance by reviewing visual content multiple times, inspiring us to implement a similar setup for MLLMs. As shown in Tab. 5 (b), Gemini achieves a notable 2.1% performance gain with two repeated videos as input. This is surprising, as autoregressive MLLMs theoretically have the capability to revisit the video multiple times during answer generation, even if the video is only presented once. This finding suggests that, despite its remarkable capabilities, a powerful MLLM like Gemini still has suboptimal reasoning processes for Video QA.

Methods	# of Frames
Proprietary Models (API)	
GPT-4o	16
Gemini-1.5 Flash	-
Gemini-1.5 Pro	-
Open-source Models	
InternVL2-2B	8
InternVL2-8B	8
InternVL2-40B	8
LongVILA-8B	32
VILA-1.5-8B	32
VILA-1.5-40B	32
LongVA-7B	32
LLaVA-NeXT-Video-7B	32
LLaVA-NeXT-Video-72B	32
LLaVA-OneVision-0.5B	32
LLaVA-OneVision-7B	32
LLaVA-OneVision-72B	32

Table 6. **Number of frames used in evaluation.**

E. Visualization Results

In this section, we present more qualitative results, including more examples of `VSI-Bench`, further error analysis case studies, examples of Chain-of-Thought promptings, and additional cognitive maps.

E.1. `VSI-Bench` Examples

In Fig. 12 and Fig. 13, we provide more examples from `VSI-Bench` to illustrate the structure and format of tasks, questions, and answers.

E.2. Error Analysis Examples

In Fig. 14, we present more case studies for our human-conducted error analysis on `VSI-Bench`. In the error analysis, we identify the categorized error types and highlight the relevant parts of the explanation.

E.3. Linguistic Prompting Examples

We provide examples for the three CoT prompting methods discussed in Sec. 5.2 to illustrate their concrete reasoning procedure in detail. We include examples of three selected tasks: object count, object size, and room size. For Zero-Shot Chain of Thought, as shown in Fig. 15, we highlight each step of the MLLM's reasoning process to offer insights into how it arrives at its final decision. For Self-Consistency w/ CoT, as illustrated in Fig. 16, each example is paired with five independent responses. The final answer is then determined by a majority vote. For Tree-of-Thought, Fig. 17 details how each depth of the decision tree is reached. At the first depth, the MLLM generates three potential plans and conducts a choice analysis to select the optimal plan. At the second and final depth, the selected plan is used to generate three potential answers, with the final output determined through a majority vote.

E.4. Cognitive Map Examples

In Fig. 18, we include 10 additional cognitive maps and pair each prediction with its corresponding ground truth map to provide insight into the alignment between predicted and ground truth layouts.

Absolute Distance

Measuring from the closest point of each object, what is the distance between the kettle and the suitcase (in meters)?

Answer: 1.8

Room Size

What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Answer: 54.1

Object Size

What is the length of the longest dimension (length, width, or height) of the sofa, measured in centimeters?

Answer: 282

Object Counting

How many bookshelf(s) are in this room?

Answer: 2

Relative Distance

Measuring from the closest point of each object, which of these objects (microwave, trash can, pillow, plant) is the closest to the shoe rack?

A. microwave
B. trash can
C. pillow
D. plant

Relative Direction

If I am standing by the sofa and facing the suitcase, is the microwave to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).

A. front-right
B. back-left
C. back-right
D. front-left

Appearance Order

What will be the first-time appearance order of the following categories in the video: microwave, sofa, trash can, pillow?

A. sofa, pillow, trash can, microwave
B. trash can, sofa, pillow, microwave
C. microwave, sofa, trash can, pillow
D. sofa, trash can, microwave, pillow

Route Plan

You are a robot beginning at the door facing the table. You want to navigate to the power strip. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. Go forward until the table 2. [please fill in] 3. Go forward until the power strip. You have reached the final destination.

A. Turn Left
B. Turn Right
C. Turn Back

Absolute Distance

Measuring from the closest point of each object, what is the distance between the tv and the stove (in meters)?

Answer: 4.7

Room Size

What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Answer: 38.7

Object Size

What is the length of the longest dimension (length, width, or height) of the stove, measured in centimeters?

Answer: 158

Object Counting

How many chair(s) are in this room?

Answer: 3

Relative Distance

Measuring from the closest point of each object, which of these objects (chair, stool, stove, sofa) is the closest to the tv?

A. chair
B. stool
C. stove
D. sofa

Measuring from the closest point of each object, which of these objects (chair, table, tv, sofa) is the closest to the stool?

A. chair
B. table
C. tv
D. sofa

Relative Direction

If I am standing by the stove and facing the tv, is the stool to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).

A. front-right
B. back-left
C. back-right
D. front-left

Route Plan

You are a robot beginning at the tv facing the tv. You want to navigate to the sofa. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the blue desk 3. [please fill in] 4. Go forward until the sofa. You have reached the final destination.

A. Turn Left, Turn Left
B. Turn Back, Turn Right
C. Turn Right, Turn Left
D. Turn Right, Turn Right

Appearance Order

No Question

Figure 12. **VSI-Bench Examples (Part 1).**

Absolute Distance

Measuring from the closest point of each object, what is the distance between the door and the cup (in meters)?

Answer: 1.6

Object Size

What is the length of the longest dimension (length, width, or height) of the heater, measured in centimeters?

Answer: 152

Relative Distance

Measuring from the closest point of each object, which of these objects (heater, cup, ceiling light, toilet) is the closest to the door?

A. heater B. cup
C. ceiling light D. toilet

Appearance Order

What will be the first-time appearance order of the following categories in the video: ceiling light, cup, heater, door?

A. cup, door, heater, ceiling light
B. ceiling light, door, cup, heater
C. heater, cup, door, ceiling light
D. ceiling light, cup, heater, door

Room Size

What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Answer: 5.8

Object Counting

No question

Relative Direction

If I am standing by the ceiling light and facing the door, is the cup to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).

A. back-left B. front-right
C. front-left D. back-right

If I am standing by the heater and facing the cup, is the toilet to my left, right, or back? An object is to my back if I would have to turn at least 135 degrees in order to face it.

A. left
B. back
C. right

Route Plan

No question

Absolute Distance

Measuring from the closest point of each object, what is the distance between the bed and the chair (in meters)?

Answer: 2.0

Object Size

What is the length of the longest dimension (length, width, or height) of the toilet, measured in centimeters?

Answer: 105

Relative Distance

Measuring from the closest point of each object, which of these objects (basket, pillow, door, heater) is the closest to the ceiling light?

A. basket B. pillow
C. door D. heater

Appearance Order

What will be the first-time appearance order of the following categories in the video: basket, blanket, toilet, ceiling light?

A. basket, blanket, toilet, ceiling light
B. blanket, toilet, basket, ceiling light
C. toilet, ceiling light, basket, blanket
D. toilet, basket, blanket, ceiling light

Room Size

What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Answer: 26.5

Object Counting

How many pillow(s) are in this room?

Answer: 4

Relative Direction

If I am standing by the sofa and facing the chair, is the tv to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).

A. front-left *B. front-right*
C. back-right D. back-left

If I am standing by the toilet and facing the basket, is the tv to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).

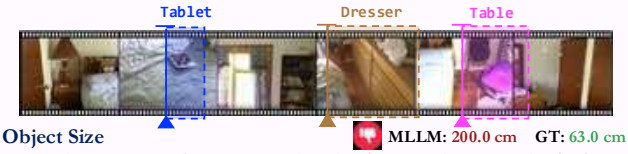
A. front-left B. back-left
C. back-right D. front-right

Route Plan

No question

Figure 13. **VSI-Bench Examples (Part 2).**

Visual Perception Error

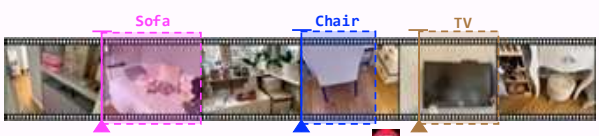


Object Size MLLM: 200.0 cm GT: 63.0 cm

Q: What is the length of the longest dimension (length, width, or height) of the table, measured in centimeters?

MLLM: 1. Identify the table: **The table is the wooden dresser/chest of drawers shown in the video.** 2. A reference object: The video shows a tablet. Tablets typically have standardized sizes. An iPad, for example, has a diagonal of approximately 24.6 cm. 3.Estimate the table's length relative to the tablet: The tablet appears to fit along the dresser's length roughly eight times. Multiplying the tablet's estimated diagonal by 8 gives us an approximate dresser length of 196.8 cm. 5. Account for perspective errors: Rounding up to 200 cm provides a reasonable estimate for the longest dimension of the table.

Error Analysis MLLM misidentified the wooden dresser as the table when answering.



Relative Direction MLLM: Left GT: Right

Q: If I am standing by the sofa and facing the tv, is the chair to the left or the right of the tv?

MLLM: Here's how to figure out the chair's position relative to the TV: 1. Locate the sofa and TV: The video starts in a hallway, then enters a bedroom. **The sofa and TV are visible at [00:00:16].** 2. Orient yourself: Imagine you are standing by the sofa, facing the TV. The camera pans across the room, passing the bed and then the desk. 3. Find the chair: The chair is visible next to the desk at [00:00:16] and again at [00:00:35]. 4. Determine left or right: From your imagined position facing the TV, the chair is on your left.

Error Analysis MLLM failed to identify the sofa in question.

Relational Reasoning Error

Absolute Distance MLLM: 3.5 Meters GT: 1.4 Meters

Q: Measuring from the closest point of each object, what is the distance between the table and the sofa (in meters)?

MLLM: ... 1. Establish a Reference: The video pans across the room, showing both the sofa and the table. The wooden floor planks provide a consistent unit of measurement. 2. Count the Planks: **Observe how many floor planks separate the closest points of the sofa and the table. It appears to be approximately 7 planks. 3. Estimate Plank Width: Standard wooden floor planks are often around 0.5 meters wide.** 4. Calculate Distance: Multiply the number of planks (7) by the estimated width of each plank (0.5 meters). 7 * 0.5 = 3.5 meters. Therefore, the estimated distance between the table and the sofa is 3.5 meters.

Error Analysis MLLM misestimated the number and width of the wood planks.

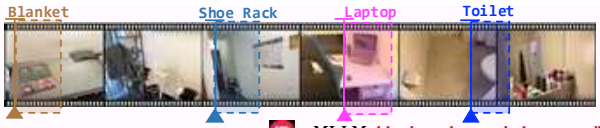
Absolute Distance MLLM: 2.5 Meters GT: 1.6 Meters

Q: Measuring from the closest point of each object, what is the distance between the stool and the fireplace (in meters)?

MLLM: ... 1. Identify the objects: Locate the stool and the fireplace. The stool is a small, square, blue upholstered piece of furniture with wooden legs. The fireplace is a brick structure with a metal fire screen. 2. Find the closest points: Determine the points on the stool and the fireplace that are closest to each other. This appears to be the front edge of the stool and the front edge of the fireplace hearth. 3. Estimate the distance: Estimate the distance between these two closest points. **Based on the size of the other furniture and the room's dimensions, the distance appears to be roughly 2.5 meters.**

Error Analysis MLLM misjudged the distance between the stool and the fireplace.

Linguistic Intelligence Error

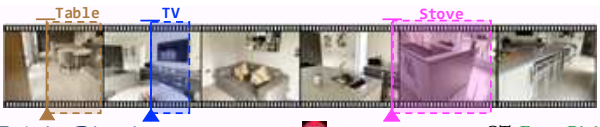


Appearance Order MLLM: blanket, shoe rack, laptop, toilet GT: blanket, toilet, laptop, shoe rack

Q: What will be the first-time appearance order of the following categories in the video: blanket, toilet, laptop, shoe rack?

MLLM: ...1. Blanket (0:00): The video begins with a shot of a bed, clearly showing a folded blanket on top. 2. Toilet (0:28): At [00:00:28], the toilet becomes visible within a small bathroom off the hallway. This is the second category to appear. 3. **Laptop (0:36):** Continuing the pan, the camera reveals a desk area. A laptop is clearly visible on the desk. **This is the third category to appear.** 4. **Shoe Rack (0:30):** Finally, still panning along the hallway, a shoe rack with several pairs of shoes is seen next to the entryway door. **This is the last of the four categories to appear. Therefore, the order is blanket, toilet, laptop, shoe rack, which corresponds to option C.**

Error Analysis MLLM thinks 0:36 < 0:30 and ranked appearance order incorrectly.



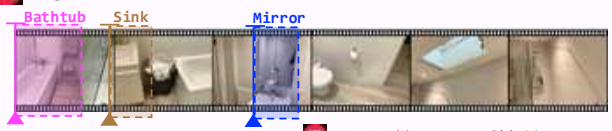
Relative Direction MLLM: Front-Left GT: Front-Right

Q: If I am standing by the stove and facing the tv, is the table to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).

MLLM: 1. Orient yourself: Imagine you are standing by the stove, facing the TV. The TV is directly in front of you. 2. Visualize the layout: Recall the video's layout. The dining table is situated between the stove/kitchen island area and the seating area with the sofa. **3. Determine the direction: Since the table is between you (at the stove) and the sofa, and the sofa is to your right, the table must be on your left.** Because it's in the space between you and the TV (which is in front of you), **the table is to your front-left.**

Error Analysis Identified object positions then did not linguistically reason correctly.

Egocentric-Allocentric Transformation Error



Route Plan MLLM: Turn Back GT: Turn Right

Q: You are a robot beginning at the bathroom sink facing the mirror. You want to navigate to the bathtub. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the bathtub.

MLLM: ...1. Starting Point: The prompt states you begin at the sink, facing the mirror. In the video, this position is shown around the 0:13 mark. 2. Destination: The goal is to reach the bathtub. The bathtub is located behind the initial position at the sink. 3. Determining the Turn: **Since the bathtub is behind the starting position, the robot needs to turn back from its position facing the mirror to head towards the bathtub.** 4. Moving to the Destination: After turning back, the robot needs to go forward to reach the bathtub. **Therefore, the correct first action is to turn back.**

Error Analysis MLLM followed video's perspective instead of the given direction.

Route Plan MLLM: Turn Right GT: Turn Left

Q: You are a robot beginning at the door and facing the display cabinet. You want to navigate to the 3-seater sofa. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. Go forward until the display cabinet 2. [please fill in] 3. Go forward until the 3-seater sofa.

MLLM: ...1. Starting Point: The robot begins at the door, facing the display cabinet. 2. Objective: The goal is to reach the 3-seater sofa. 3. Path: The robot moves forward, closer to the display cabinet. **To reach the 3-seater sofa, which is located to the right of the display cabinet from the robot's perspective [00:00:10], the robot must turn right.** After turning right, moving forward takes the robot directly to the sofa. **Therefore, only turning right allows the robot to reach the destination.**

Error Analysis MLLM followed video's perspective instead of the given direction.

Figure 14. **Additional Error Analysis Examples.**

Object Count

Question

These are frames of a video. How many sofa(s) are in this room? Let's think step by step.

MLLM

Here's how we can determine the number of sofas:
1. Initial Scan: From the beginning, we see a large, L-shaped gray sofa in the living area.
2. Thorough Check: The video pans the entire room, showcasing different angles and areas. We don't see any other sofas besides the initial one.
3. Conclusion: Therefore, there is only **one** sofa in the room.

Object Size

Question

These are frames of a video. What is the length of the longest dimension (length, width, or height) of the fireplace, measured in centimeters? Let's think step by step.

MLLM

Here's how we can estimate the fireplace's longest dimension:
1. Identify a reference object: The TV appears to be a standard flat-screen model. We can use its typical size as a reference. A common large TV size is around 55 inches diagonally, which is approximately 140 cm.
2. Compare the fireplace to the reference: In the video, the fireplace's mantelpiece (the shelf above the firebox) appears to be slightly shorter than the TV's width. The TV stand is about the same width as the fireplace.
3. Estimate the fireplace's length: Given that the TV is roughly 140 cm diagonally, and the fireplace mantel appears slightly shorter than the TV's width, we can estimate the fireplace's length to be around 110-120 cm. This is a rough estimate, as the exact dimensions of the TV are unknown, and perspective can distort the relative sizes of objects in the video.
Therefore, a reasonable estimate for the fireplace's longest dimension is between 110 and 120 cm.

Room Size

Question

These are frames of a video. What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space. Let's think step by step.

MLLM

Here's a step-by-step estimation of the bathroom size:
1. Establish a reference: The toilet appears to be a standard size. A typical toilet's width is around 0.4 meters.
2. Measure the room using the reference: Visually comparing the toilet width to the room's dimensions, the room seems roughly 2 toilets wide and 3 toilets long.
3. Calculate the dimensions: 2 toilets wide * 0.4 meters/toilet = 0.8 meters wide. 3 toilets long * 0.4 meters/toilet = 1.2 meters long.
4. Calculate the area: 0.8 meters * 1.2 meters = 0.96 square meters.
Therefore, the bathroom is approximately 1 square meter. This is a rough estimate, and the actual size could be slightly larger or smaller.

Figure 15. **Zero-Shot CoT Examples.**

Object Count

Question

These are frames of a video. How many sofa(s) are in this room? Let's think step by step.

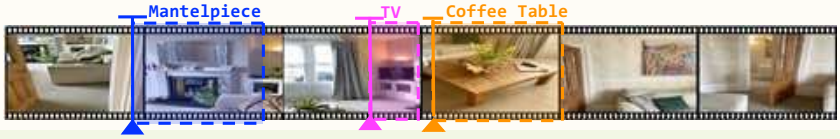
MLLM Responses

Turn 0	Turn 1	Turn 2	Turn 3	Turn 4
... 1. Initial Scan: The video first shows a living area with a sectional sofa, an L-shaped couch. 2. Confirmation: Later, the camera focuses on the couch again. 3. No Other Sofas: Throughout the video, no other sofas or couches are visible. Therefore, there is only one sofa.	... 1. Initial Scan: The video starts by showing a room with a large, L-shaped grey sofa. 2. Throughout the Video: The camera pans around the room, showing various areas, but never reveals another sofa. 3. Conclusion: Therefore, there is only one sofa in the room. Answer: 1	... 1. Initial Scan: From the beginning of the video, we see a large, gray L-shaped sectional sofa. 2. Thorough Check: The video pans across the entire room, showcasing different furniture, but we don't see any other sofas. 3. Conclusion: Therefore, there is only one sofa in this room. Answer: 1	... 1. Initial Scan: From the beginning of the video, we see a large, L-shaped gray sofa in the living area. 2. Thorough Check: As the video pans across the room, we don't see any other sofas. 3. Conclusion: Therefore, there is only one sofa.	... 1. Initial Scan: From the beginning of the video, we see a large, L-shaped gray sofa. 2. Thorough View: The video pans across the entire room, showcasing various furniture. 3. No Other Sofas: Throughout the video, we only see the one initial gray sofa. Therefore, there is only one sofa in the room.

Final Answer (Majority Vote) Answer: 1

Object Size

Question



These are frames of a video. What is the length of the longest dimension (length, width, or height) of the fireplace, measured in centimeters? Let's think step by step.

MLLM Responses

Turn 0	... 1. Reference Object: We can use the TV's typical dimensions as a starting point. A common large TV size is around 140 cm. 2. Visual Comparison: The fireplace mantelpiece appears slightly shorter than the TV's width, which is roughly 120 cm. 3. Estimate the fireplace dimension: We estimate the longest dimension of the fireplace to be around 100-110 cm.	... 1. Reference Object: We can use the TV as a reference. The TV's width is roughly the same as the fireplace's length. A reasonable guess for the TV size would be around 40-50 inches. 2. Estimate fireplace length: Since the fireplace's length appears similar to the TV's width, we can infer that the fireplace's longest dimension would be around 100-110 cm.
Turn 1	... 1. Reference Object: The TV screen is a good option. Let's assume it's 45 inches, which is about 114 cm diagonally. 2. Visual Comparison: The fireplace appears to be around the same length as the TV screen's width. 3. Final Estimate: Since the fireplace's length appears roughly equal to the TV's width, we can estimate the fireplace's longest dimension to be approximately 99 cm.	... 1. Reference Object: The TV appears to be a standard flat-screen TV. A common size is around 107 cm diagonally. 2. Visual Comparison: The fireplace's mantle appears to be roughly the same length as the TV's width. 3. Estimate the fireplace's length: If the TV is around 107 cm diagonally, its width is around 94 cm. We can reasonably estimate the fireplace's longest dimension to be approximately 94 cm.
Turn 2	... 1. Reference Object: The wooden coffee table seems suitable. A typical coffee table is around 100 cm on each side. Let's assume this table is approximately 100 cm wide. 2. Visual Comparison: The fireplace is slightly narrower than the mantle. 3. Estimation: We estimated the table to be 100 cm wide. 4. Conclusion: Therefore, a reasonable estimate for the fireplace's longest dimension is approximately 100 cm.	

Final Answer (Majority Vote) Answer: 100 cm

Room Size

These are frames of a video. What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space. Let's think step by step.

MLLM Responses

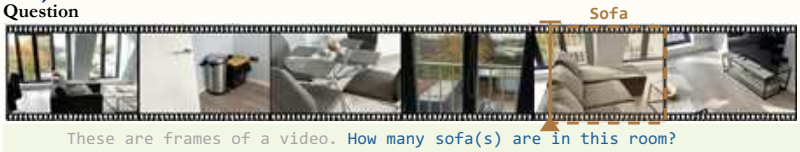
Turn 0	... 1. Measurable Features: The floor tiles provide a visual grid for estimation. 2. Count Tiles: It appears to be 3 tiles wide and 5 tiles long in the main area, and the toilet area is about 1.5 tiles length. 3. Calculate Area: 16.5 tiles * 0.09 sq m/tile = 1.485 sq m. 4. Final Estimate: A reasonable estimate for the bathroom size is approximately 1.5 square meters.	... 1. Main features: We see a toilet, sink, and a bathtub/shower combination. 2. Estimate width and length: The bathtub's length is around 1.6 meters. The toilet add 0.8 meters, making the length approximately 2.4 meters. The width could be around 1.8 meters. 3. Calculate area: 2.4m x 1.8m = 4.32 sq m. 4. Final estimate: A reasonable estimate is between 3.5 and 4.5 square meters.
Turn 1	... 1. Tile dimensions: The tiles appear to be around 30cm x 10cm. 2. Floor tile count: We estimate there are about 4 tiles wide and 7 tiles long across the main floor area. 3. Floor Area: Main area: 1.2m x 0.7m = 0.84 sq m. Toilet area: 0.4m x 0.5m = 0.2 sq m. Therefore, the total area is 0.84 sq m + 0.2 sq m = 1.04 sq m.	... 1. Floor tiles: The floor has square, off-white tiles that equals 0.09 square meters per tile. 2. Count the tiles: The floor appears to have roughly 4 tiles wide and 6 tiles long in the main area. The toilet area adds another 2 tiles wide and 2 tiles long. 3. Total areas: 2.16 sq m+ 0.36 sq m = 2.52 sq m.
Turn 2	... 1. Tile size: The video shows square tiles on the floor. Let's assume they're around 30cm x 30cm. 2. Floor area in tiles: We can roughly count the tiles to be 4 tiles wide and 6 tiles long in the main bathroom area and another tile's width and two tiles' length. for the toilet. 3. Total estimated area: 26 tiles * 0.09 sq m/tile = 2.34 sq m.	

Final Answer (Majority Vote) Answer: 2 square meters

Figure 16. **Self-Consistency w/ CoT Examples.**

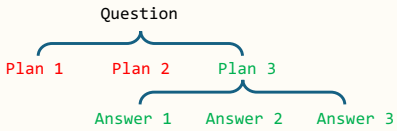
Object Count

Question



These are frames of a video. How many sofa(s) are in this room?

Tree-of-Thought



Plan Generations

1. Review the video.
2. Count the number of sofas in the room.

Plan 1 (0/3 Votes)

1. Watch the video and focus on the living area.
2. Count the number of sofas present.

Plan 2 (3/3 Votes)

1. Examine the room as the camera enters and pans.
2. Count the number of sofas that are visible.

Plan 3 (0/3 Votes)

Choice Analysis

Choice 1 is fine, but lacks some details, such as the room's identifying characteristics, which would make the plan more likely to deliver the correct answer. Choice 2 is more specific about focusing specifically on the living area of the room, but it still lacks details about what actions need to be taken to find the sofas and how to count them. Choice 3 includes instructions to follow the camera as it enters and pans around the room, which helps identify the proper room. The visibility note and explicit viewing plan make this the most robust choice. **The best choice is 3.**

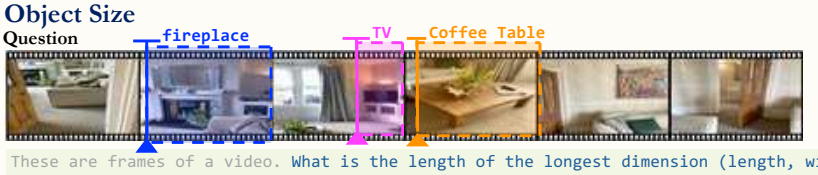
Final Answer

One Sofa

Answer 1-3 (3/3 Votes)

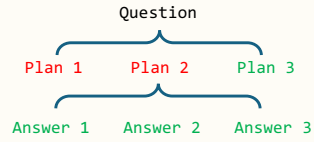
Object Size

Question



These are frames of a video. What is the length of the longest dimension (length, width, or height) of the fireplace, measured in centimeters?

Tree-of-Thought



Plan Generations

1. Find a frame where the fireplace is visible. 2. Estimate the fireplace's length relative to a known object. The flatscreen television is roughly the same length as the fireplace. 3. Research typical flatscreen TV dimensions. 4. Assume the TV is a typical size and has a standard 16:9 aspect ratio. This would suggest a width of 124cm and a height of 69cm. 5. Since the TV and fireplace appear similar in length, approximate the fireplace's longest dimension to be around 124 cm.

Plan 1 (0/3 Votes)

1. Identify a reference object. A wooden coffee table is visible, and its dimensions are approximately length = 115 cm and width = 115 cm, as indicated in the original listing. 2. Estimate the length of the fireplace relative to the coffee table. The fireplace's longest dimension appears roughly the same as that of the coffee table. 3. Estimate the length of the fireplace: The longest dimension is approximately the same as the table, which is 115 cm.

Plan 2 (3/3 Votes)

1. Determine the length of the fireplace using the available video frames. 2. Since the video doesn't provide a reference object with known dimensions near the fireplace, it's not possible to directly measure its length. 3. Due to the lack of information in the video frames, attempt to estimate the length in centimeters using visual clues. 4. Compare the size of the fireplace to elements of known or estimated size.

Plan 3 (0/3 Votes)

Choice Analysis

Choice 1 has a few flaws. First, while the approximation of the fireplace's length relative to the TV is reasonable, the TV's actual dimensions could be different, leading to inaccuracies. Choice 2 is more promising because it uses a specific reference object with known measurements, the coffee table. Comparing the fireplace's length to the coffee table allows for a more accurate estimation. Choice 3 does not offer a concrete plan for measurement. It recognizes the need for a reference object but doesn't provide a solution. **The best choice is 2.**

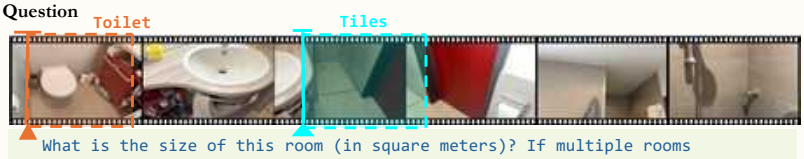
Final Answer

115 centimeters

Answer 1-3 (3/3 Votes)

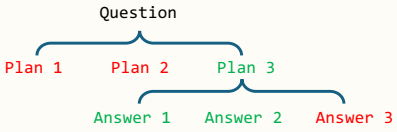
Room Size

Question



What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Tree-of-Thought



Plan Generations

1. Identify the dimensions using the toilet and tiles as reference. 2. The toilet appears to be a standard size (0.4m wide). 3. The tiles seem to be square and roughly 0.3m wide. 4. Multiply the number of tiles by the tile width to estimate length and width. 5. Multiply the length and width to get an approximate area in square meters.

Plan 1 (1/3 Votes)

1. Identify the shape of the room. 2. Observe the dimensions of the room's features (toilet, sink, shower stall) to estimate the lengths of walls, recognizing that some objects may visually distort the room's size. 3. Estimate length and width in meters. 4. Calculate area in square meters: length x width.

Plan 2 (0/3 Votes)

1. Identify main features of the room. 2. Use these features to visually estimate the dimensions of the bathroom. 3. Assume the toilet is roughly 0.4m wide and the tiles are approximately 0.3m x 0.3m. 4. Count the tiles to approximate the dimensions. 5. Multiply the estimated length and width to get an approximate area. 6. Adjust the area calculation slightly downwards to account for the narrow shower.

Plan 3 (2/3 Votes)

Choice Analysis

Choice 1 is a good starting point, but it relies heavily on the assumption that the tiles are square. This assumption might be inaccurate. Choice 2 is overly simplistic. While it encourages estimating the lengths of walls using room features, it doesn't provide a concrete method for doing so. Choice 3 offers a more detailed and seemingly accurate approach. It uses familiar bathroom fixtures (toilet) and the floor tiles for estimation. The plan also accounts for the narrower shower area and suggests an appropriate adjustment to the calculation. **The best choice is 3.**

Final Answer

2.5 Square Meters

Answer 1-2 (2/3 Votes)

2.4 Square Meters

Answer 3 (1/3 Votes)

Figure 17. **Tree-of-Thought Examples.**

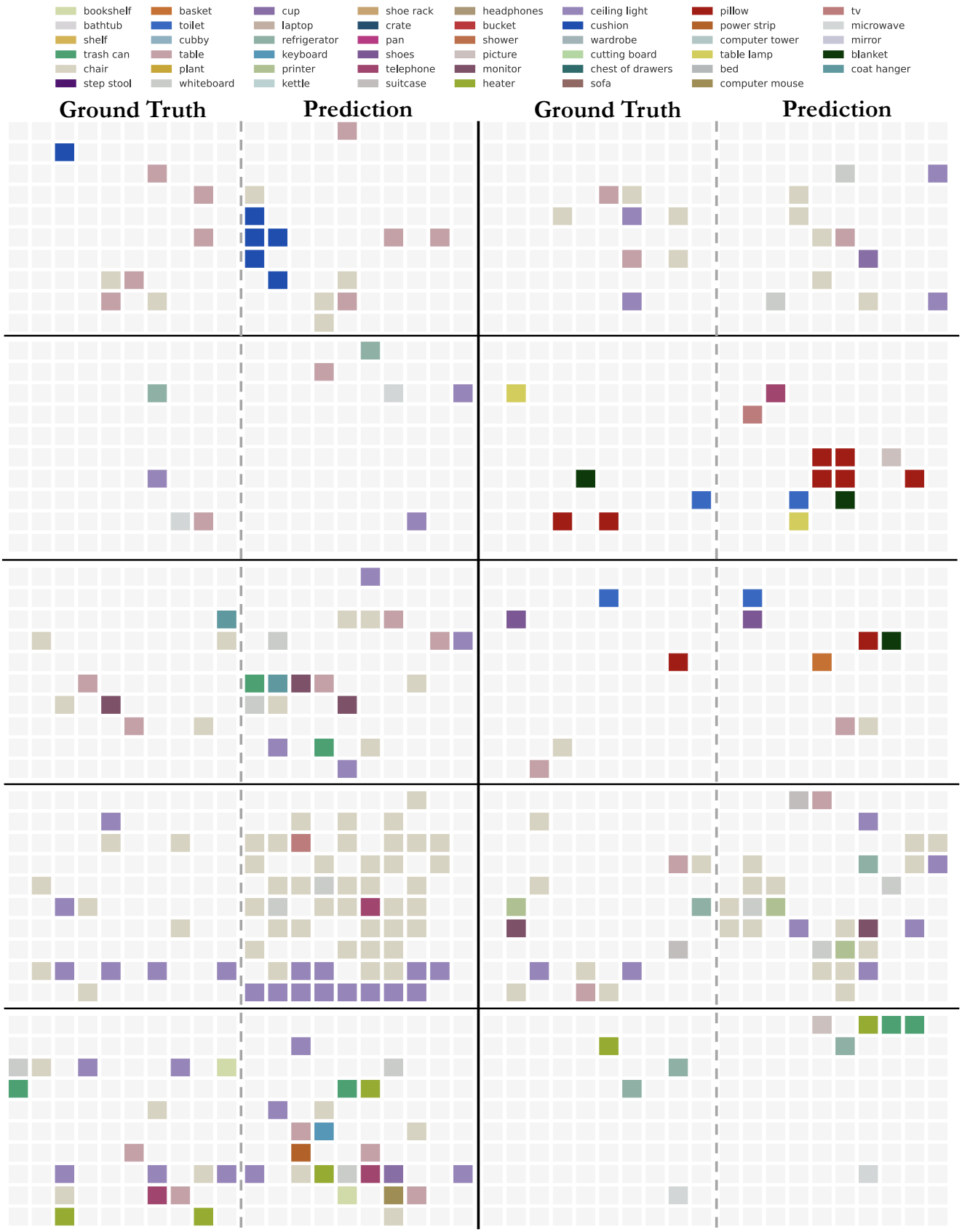


Figure 18. **Additional predicted cognitive map examples.**

	Models	QA. Type	Prompt
`Pre-Prompt`	-	-	*These are frames of a video.*
`Post-Prompt`	Open-source Models	NA	*Please answer the question using a single word or phrase.*
		MCA	*Answer with the option's letter from the given choices directly.*
	Proprietary Models	NA	*Do not respond with anything other than a single number!*
		MCA	*Answer with the option's letter from the given choices directly.*

Table 7. **Prompts used in evaluation.** NA and MAC indicates questions with *Numerical Answer* and *Multiple Choice Answer* respectively.

Methods	Avg.	*Obj. Count*	*Abs. Dist.*	*Obj. Size*	*Room Size*	*Rel. Dist.*	*Rel. Dir.*	*Route Plan*	*Appr. Order*
		Numerical Answer				Multiple-Choice Answer			
Proprietary Models (API)									
GPT-4o	35.6	36.2	4.6	47.2	40.4	40.0	46.2	32.0	38.0
Gemini-1.5 Flash	45.7	50.8	33.6	56.5	45.2	48.0	39.8	32.7	59.2
Gemini-1.5 Pro	48.8	49.6	28.8	58.6	49.4	46.0	48.1	42.0	68.0
Gemini-2.0 Flash	45.4	52.4	30.6	66.7	31.8	56.0	46.3	24.5	55.1
Open-source Models									
InternVL2-2B	25.5	30.6	20.4	26.0	29.6	28.0	39.2	28.0	2.0
InternVL2-8B	32.9	26.4	25.4	43.8	41.6	30.0	32.2	20.0	44.0
InternVL2-40B	37.6	40.8	23.8	48.0	26.0	46.0	30.1	42.0	44.0
LongVILA-8B	19.1	23.4	10.8	11.4	0.0	20.0	33.1	28.0	26.0
VILA-1.5-8B	31.4	12.2	23.4	51.4	18.6	36.0	41.5	42.0	26.0
VILA-1.5-40B	32.3	14.6	21.0	48.0	20.6	42.0	22.0	40.0	50.0
LongVA-7B	31.8	41.2	17.4	39.6	25.4	30.0	52.8	34.0	14.0
LLaVA-NeXT-Video-7B	35.7	49.0	12.8	48.6	21.4	40.0	43.5	34.0	36.0
LLaVA-NeXT-Video-72B	39.3	41.4	26.6	55.6	31.6	36.0	25.6	42.0	56.0
LLaVA-OneVision-0.5B	27.7	44.0	23.0	18.8	28.4	30.0	33.4	36.0	8.0
LLaVA-OneVision-7B	33.8	48.2	22.0	44.4	14.0	44.0	31.9	34.0	32.0
LLaVA-OneVision-72B	41.6	38.0	31.6	54.4	35.2	44.0	39.7	32.0	58.0

Table 8. **Complete `VSI-Bench` (tiny) evaluation results.**

Methods	Avg.	*Obj. Count*	*Abs. Dist.*	*Obj. Size*	*Room Size*	*Rel. Dist.*	*Rel. Dir.*	*Route Plan*	*Appr. Order*
		Numerical Answer				Multiple-Choice Answer			
Proprietary Models (API)									
GPT-4o	14.5	0.1	5.2	36.7	0.0	10.8	23.2	26.9	13.1
Gemini-1.5 Flash	19.9	25.0	30.3	52.5	0.0	0.0	21.2	29.9	0.2
Gemini-1.5 Pro	32.3	30.6	11.5	51.5	33.1	33.8	44.6	33.5	20.2
Open-source Models									
InternVL2-2B	17.8	5.4	23.7	9.2	0.0	26.9	41.2	27.9	7.9
InternVL2-8B	27.6	31.9	26.8	38.3	0.7	27.1	39.2	33.0	23.6
InternVL2-40B	24.4	5.4	29.1	39.2	0.7	30.3	37.7	27.9	24.7
LongVILA-8B	20.2	47.4	12.6	8.7	0.6	24.3	27.0	27.4	13.9
VILA-1.5-8B	21.5	7.4	7.6	45.7	0.0	25.4	39.1	29.4	17.6
VILA-1.5-40B	25.5	5.3	27.6	46.5	0.7	30.2	37.1	31.5	25.0
LongVA-7B	21.9	5.1	18.1	27.4	26.1	23.4	39.8	26.9	8.7
LLaVA-NeXT-Video-7B	25.2	14.8	14.6	32.5	26.1	26.8	45.0	33.0	8.5
LLaVA-NeXT-Video-72B	29.1	19.0	25.4	46.3	26.1	29.0	38.8	33.0	15.5
LLaVA-OneVision-0.5B	28.6	38.4	30.1	32.0	24.3	22.0	41.8	34.5	5.4
LLaVA-OneVision-7B	25.3	13.8	8.5	45.5	26.1	28.6	41.2	27.9	11.1
LLaVA-OneVision-72B	28.9	8.2	23.8	54.1	26.1	30.4	38.1	33.0	17.1

Table 9. **Complete blind evaluation results.**

Methods	Avg.	*Obj. Count*	*Abs. Dist.*	*Obj. Size*	*Room Size*	*Rel. Dist.*	*Rel. Dir.*	*Route Plan*	*Appr. Order*
		Numerical Answer				Multiple-Choice Answer			
Proprietary Models (API)									
GPT-4o	19.5	46.1	0.1	7.1	38.2	26.2	18.0	4.6	15.4
Gemini-1.5 Flash	22.2	24.9	0.5	1.0	54.4	37.7	19.9	1.5	37.7
Gemini-1.5 Pro	13.0	25.5	19.5	12.6	10.6	17.5	1.7	2.5	14.4
Open-source Models									
InternVL2-2B	9.6	16.4	1.2	12.8	35.0	6.9	3.0	2.5	-0.8
InternVL2-8B	7.0	-8.8	1.9	9.9	39.1	9.7	-8.5	-3.0	16.0
InternVL2-40B	11.6	29.6	-2.2	7.3	31.1	11.8	-5.5	6.1	14.9
LongVILA-8B	1.4	-18.2	-3.5	7.9	-0.6	5.3	3.7	5.1	11.5
VILA-1.5-8B	7.3	10.0	14.2	4.6	18.8	6.7	-4.4	1.5	7.2
VILA-1.5-40B	5.7	17.1	-2.8	2.2	22.0	10.4	-11.4	0.0	7.9
LongVA-7B	7.2	32.9	-1.5	11.5	-3.9	9.7	3.5	-1.5	7.1
LLaVA-NeXT-Video-7B	10.5	33.8	-0.6	15.2	-1.9	16.7	-2.7	1.0	22.1
LLaVA-NeXT-Video-72B	11.7	29.9	-2.6	11.1	9.2	13.3	-2.0	2.0	33.0
LLaVA-OneVision-0.5B	-0.5	7.8	-1.7	-16.6	4.0	6.9	-5.0	0.0	0.3
LLaVA-OneVision-7B	7.0	33.9	11.7	1.9	-13.9	13.9	-6.0	1.5	13.3
LLaVA-OneVision-72B	11.4	35.4	0.1	3.5	11.4	12.1	1.8	-0.5	27.4

Table 10. **Results of Vision Enabled − Vision Disabled.**