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++ [97], 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 × 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 [106].

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 [38, 89], 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 [87], we prompt MLLMs to generate multiple answers for a given question under Zero-shot-CoT [38] 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 [95], 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. 16. Fig. 17, and Fig. 18 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. 9 and Fig. 10. 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. 10. 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.

B.5. Cognitive Map on More MLLMs

We evaluate two more MLLMs, LLaVA-Video-7B and LLaVA-Video-72B. Tab. 5 validates our Sec. 6.1 finding of significantly stronger local than global accuracy. Regarding Sec. 6.2, as shown in Tab. 6, LLaVA-Video-72B achieves an 8% performance gain. In contrast, LLaVA-Video-7B performance decreases, likely due to its limited model capacity, which impairs cog. map prediction (Tab. 5 shows its suboptimal acc. on cog. map compared to Gemini-1.5 Pro and LLaVA-Video-72B).

Distance	[1.0, 2.1]	(2.1, 3.3]	(3.3, 4.4]	(4.4, 5.5]	(5.5, 6.6]	(6.6, 7.8]	(7.8, 8.9]	(8.9, 10.0]
Gemini-1.5 Pro	0.64	0.48	0.35	0.35	0.28	0.12	0.06	0.00
LLaVA-Video-72B	0.59	0.45	0.42	0.30	0.15	0.23	0.16	0.00
LLaVA-Video-7B	0.50	0.43	0.34	0.29	0.19	0.18	0.14	0.00

Table 5. **Locality of cognitive maps.**

Models	LLaVA-Video-72B	LLaVA-Video-7B
w/o. Cog. Map	36.0	40.0
w/. Cog. Map	42.0	32.0

Table 6. **Rel. dist. task with cognitive maps.**

C. Evaluation Details

C.1. General Evaluation Setup

Our evaluation processes are primarily conducted using the `LMMs-Eval` project [100]. 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. 10.

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.

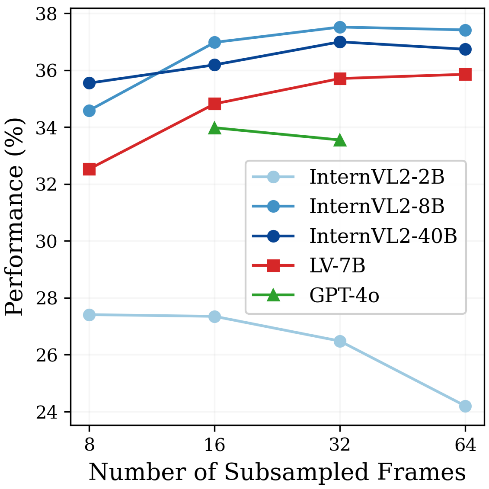


Figure 11. **Analysis of different # sampled frames.**

In addition, we provide the human evaluation on another `VSI-Bench` subset with 560 samples optimized to minimize the average performance gap between this subset and full set for all MLLMs. As shown in Tab. 11, this subset has an average performance discrepancy compared to full set (see Tab. 1) just 0.5% and a maximum of 2.9%.

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 [100], 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. 9. We believe that frame sampling strategies are a model design choice separate from the benchmark design. Established benchmarks (*e.g.*, VideoMME [24] and EgoSchema [55]) also employ default sampling, reinforcing this perspective. In addition, as shown in the Fig. 11, the # of sampled frames only marginally affects performance—it is not the primary bottleneck.

C.4. More Evaluation Results

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

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. 12, the consistent improvements in "Enabled−Disabled" and general degradation in "Disabled−Chance" demonstrates that video is essential

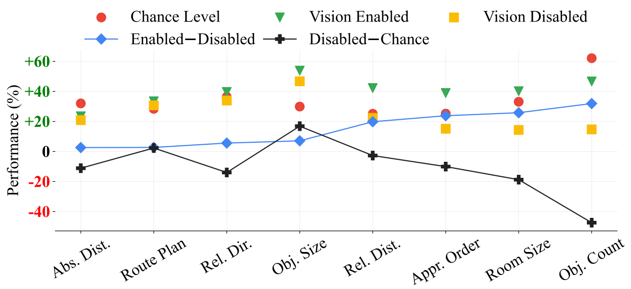


Figure 12. **Performance comparisons between Vision Enabled (w/ video), Vision Disabled (w/o video) and Chance Level (Freq.).** Enabled−Disabled indicates the gap between Vision Enabled and Vision Disabled, and Disabled−Chance betokens the gap between Vision Disabled and Chance Level (Freq.). Tasks are sorted by Enable−Disable for better understanding.

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.

In addition, as shown in Tab. 13, 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.

Socratic LLMs with Frame Captions. Following OpenEQA [54] and HourVideo [12], we implement a Socratic variant of GPT-4o using LLaVA-Video-72B as the captioner and GPT-4o as the answering LLM. As shown in Tab. 7, Socratic lags behind the standard GPT-4o by 4.7%.

GPT-4o	Standard	Socratic	Blind
Avg.	34.0	29.3	14.5

Table 7. **Socratic LLMs with Frame Captions.**

`VSI-Bench` (tiny) Results. As shown in Tab. 12, 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.

Vision Enabled − Vision Disabled. Tab. 14 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.

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 8. Ablations on the video input sequence and repetition.

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 [49, 78], 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. 8 (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. Specifically, input is formatted as: `[Video] [Context] [Video]` with identical video, where the system prompt explicitly informs the model of the redundancy of input video. As shown in Tab. 8 (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.

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. 13 and Fig. 14, we provide more examples from `VSI-Bench` to illustrate the structure and format of tasks, questions, and answers.

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

Table 9. **Number of frames used in evaluation.**

E.2. Error Analysis Examples

In Fig. 15, 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. 16, 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. 17, each example is paired with five independent responses. The final answer is then determined by a majority vote. For Tree-of-Thought, Fig. 18 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. 19, 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

Object Size

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

Answer: 282

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

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

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 Counting

How many bookshelf(s) are in this room?

Answer: 2

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

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

Object Size

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

Answer: 158

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

Appearance Order

No Question

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 Counting

How many chair(s) are in this room?

Answer: 3

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).

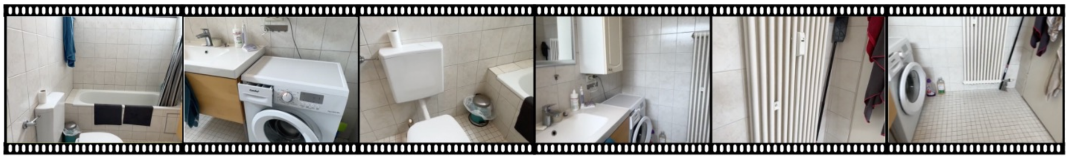
A. front-right
B. back-left
C. back-right
D. front-left

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

Figure 13. **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).

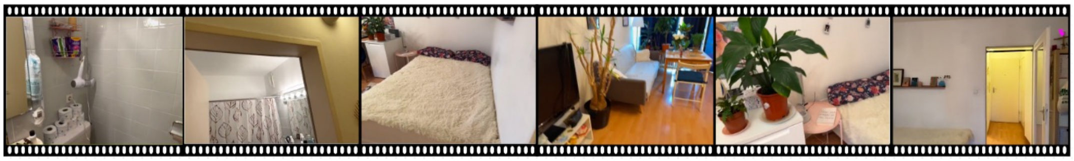
A. back-left B. front-right
C. front-left D. back-right

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).

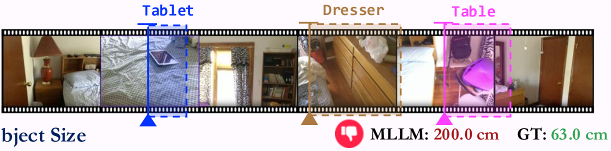
A. front-left B. back-left
C. back-right D. front-right

Route Plan

No question

Figure 14. **`VSI-Bench` Examples (Part 2).**

Visual Perception Error



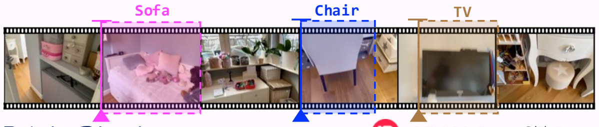
Object Size

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

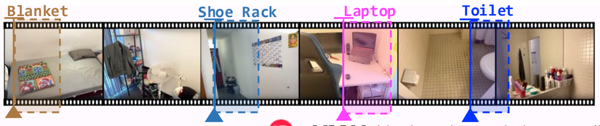
MLLM: 2.5 Meters GT: 1.6 Meters

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

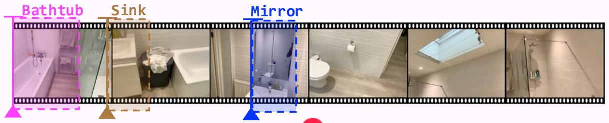
MLLM: Front-Left GT: Front-Right

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.

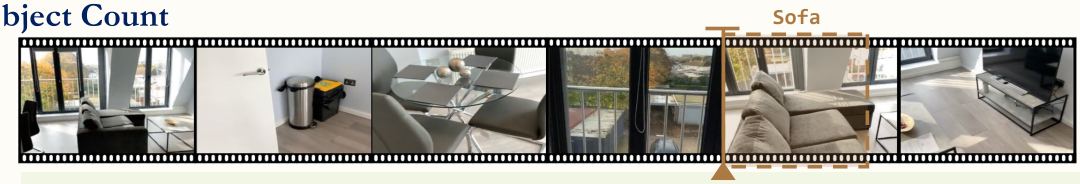
MLLM ...1. Starting Point: The robot begins at the door, facing the display cabinet. 2. Objective: The goal is to reach the 3-seater sofa. 3. Path: The robot moves forward, closer to the display cabinet. **To reach the 3-seater sofa, which is located to the right of the display cabinet from the robot's 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 15. **Additional Error Analysis Examples.**

Object Count

Question



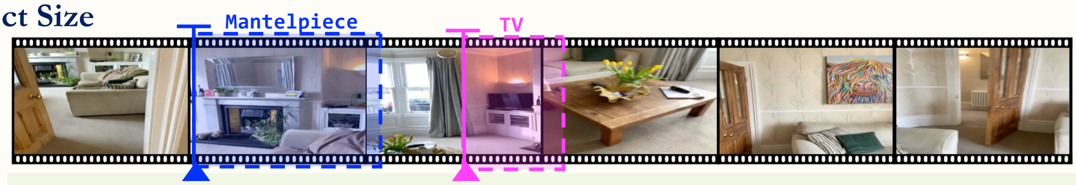
These are frames of a video. How many sofa(s) are in this room? Let's think step by step.

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



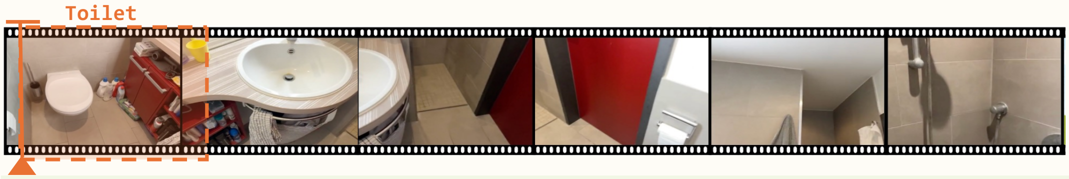
These are frames of a video. What is the length of the longest dimension (length, width, or height) of the fireplace, measured in centimeters? Let's think step by step.

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

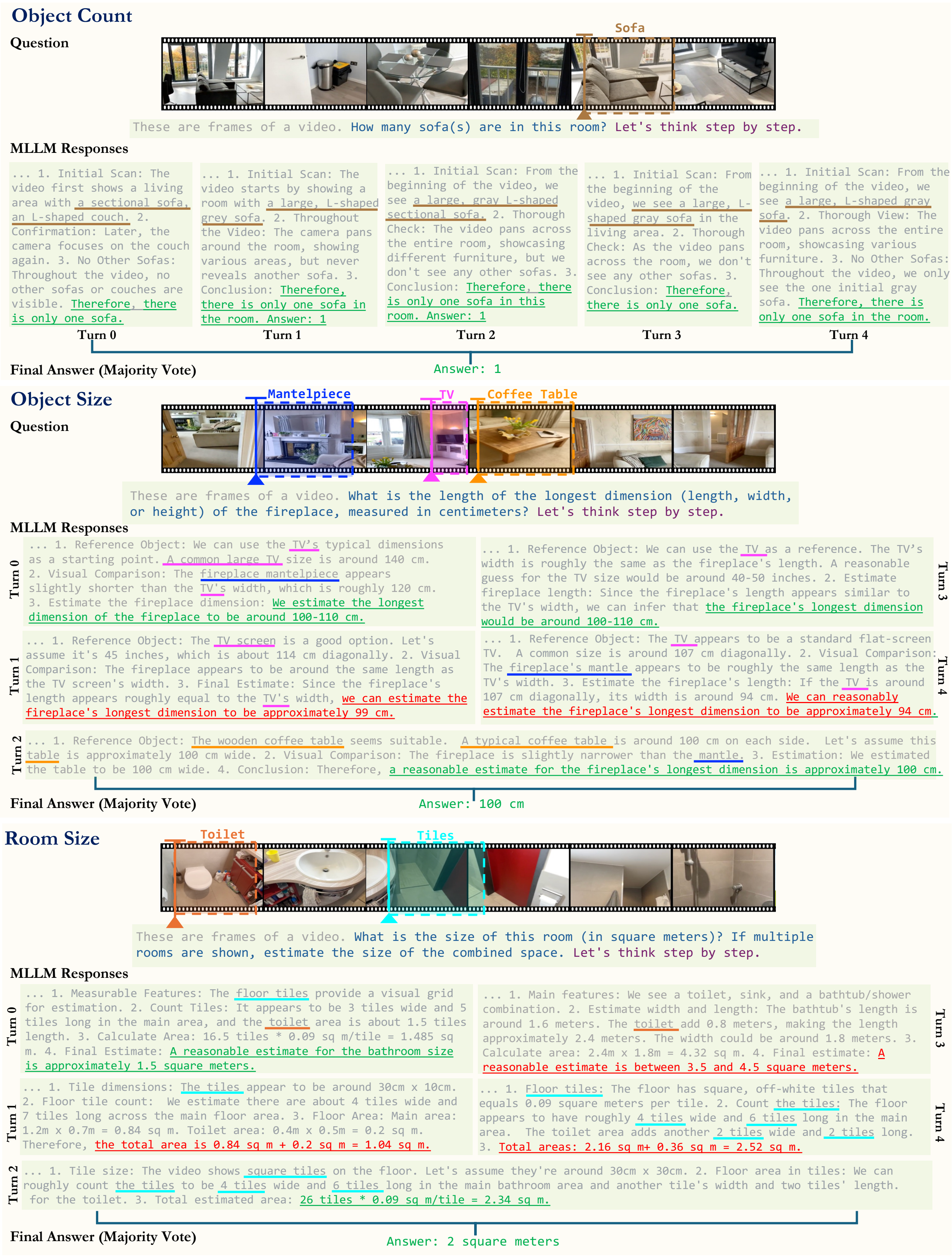


These are frames of a video. What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space. Let's think step by step.

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 16. **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: ... 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.

Turn 1: ... 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

Turn 2: ... 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

Turn 3: ... 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.

Turn 4: ... 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.

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.

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.

Turn 3: ... 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 4: ... 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.

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.

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.

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.

Turn 3: ... 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.

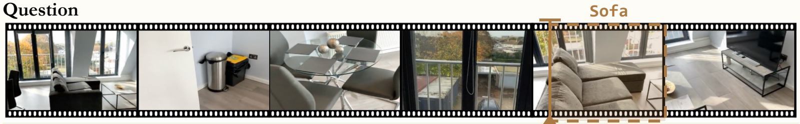
Turn 4: ... 1. Floor tiles: The floor has square, off-white tiles that equals 0.09 square meters per tile. 2. Count the tiles: The floor appears to have roughly 4 tiles wide and 6 tiles long in the main area. The toilet area adds another 2 tiles wide and 2 tiles long. 3. Total areas: 2.16 sq m+ 0.36 sq m = 2.52 sq m.

Final Answer (Majority Vote) Answer: 2 square meters

Figure 17. **Self-Consistency w/ CoT Examples.**

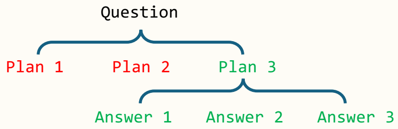
Object Count

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.

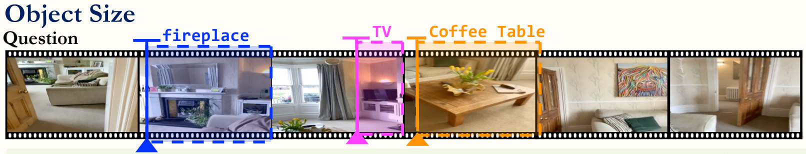
Final Answer

One Sofa

Answer 1-3 (3/3 Votes)

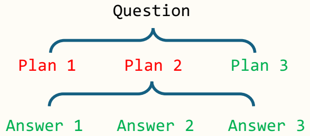
Object Size

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.

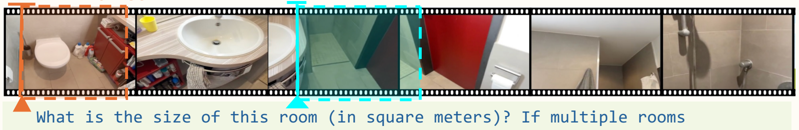
Final Answer

115 centimeters

Answer 1-3 (3/3 Votes)

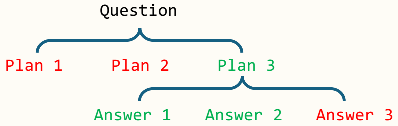
Room Size

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 18. **Tree-of-Thought Examples.**

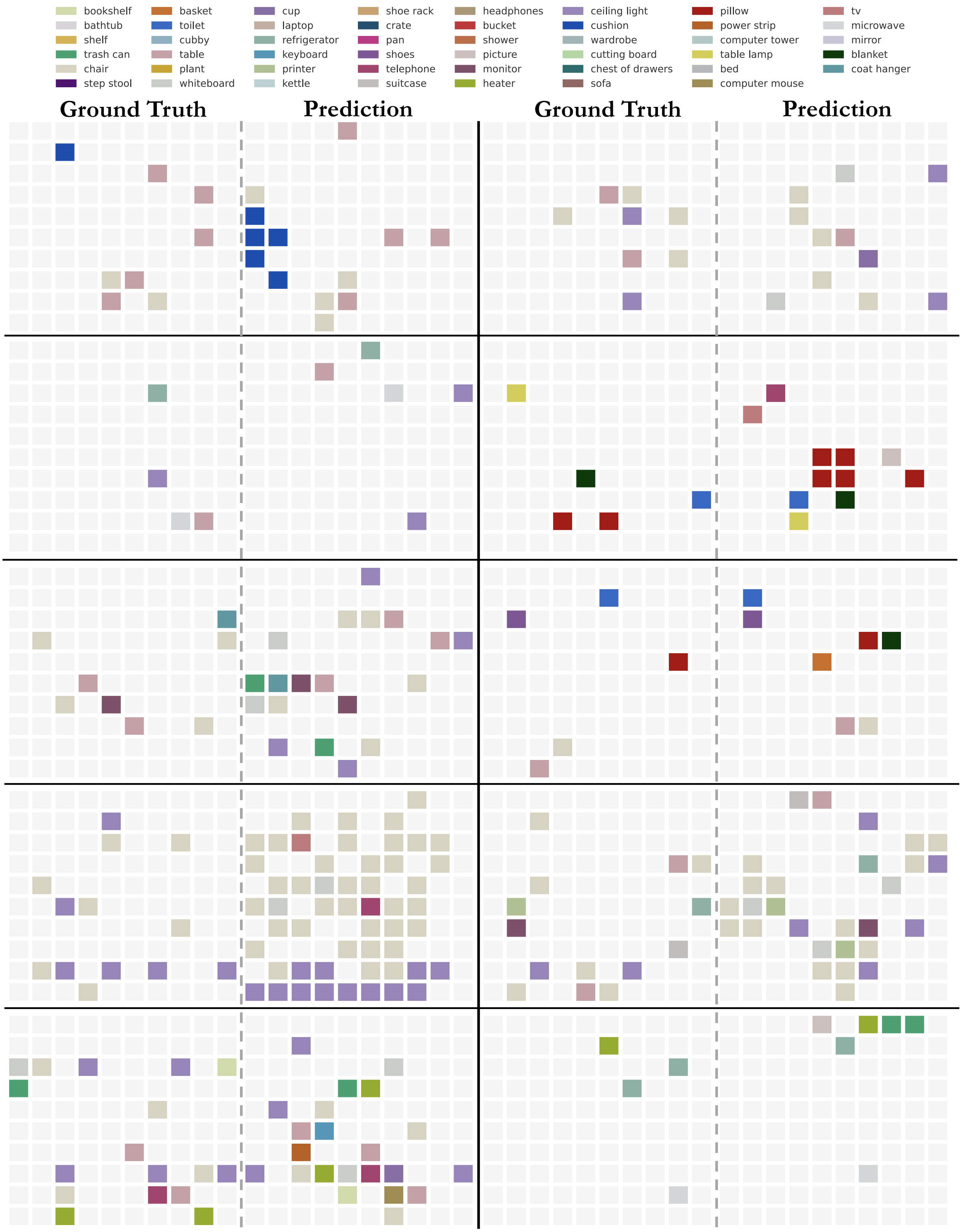


Figure 19. **Additional predicted cognitive map examples.**

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

		Obj. Count	*Abs. Dist.*	*Obj. Size*	*Room Size*	*Rel. Dist.*	*Rel. Dir.*	*Route Plan*	*Appr. Order*
Methods	Avg.	Numerical Answer				Multiple-Choice Answer			
Proprietary Models (API)									
Gemini-1.5 Flash	41.6	49.1	30.3	52.7	53.7	37.1	40.8	31.4	37.1
Gemini-1.5 Pro	44.9	55.1	30.3	63.1	43.3	50.0	45.9	35.7	35.7
Open-source Models									
InternVL2-2B	27.0	22.4	24.9	21.1	34.1	32.9	43.5	30.0	7.1
InternVL2-8B	34.1	22.6	28.3	47.6	39.6	35.7	30.4	30.0	38.6
InternVL2-40B	35.5	34.4	26.9	45.6	31.3	41.4	31.7	32.9	40.0
LongVILA-8B	21.0	28.7	8.6	16.3	0.0	28.6	30.5	31.4	24.3
VILA-1.5-8B	28.4	17.3	21.6	49.9	18.6	31.4	34.4	30.0	24.3
VILA-1.5-40B	30.8	21.4	24.4	48.3	21.9	40.0	25.0	30.0	35.7
LongVA-7B	29.0	38.1	16.9	38.1	21.7	32.9	42.8	25.7	15.7
LLaVA-Video-7B	34.9	47.9	13.4	46.7	23.9	42.9	41.9	32.9	30.0
LLaVA-Video-72B	40.5	48.3	22.6	56.7	34.6	41.4	36.5	35.7	48.6
LLaVA-OneVision-0.5B	27.6	45.1	27.9	14.7	27.9	28.6	37.0	34.3	5.7
LLaVA-OneVision-7B	32.1	46.9	19.9	46.9	12.1	41.4	35.1	30.0	24.3
LLaVA-OneVision-72B	39.6	42.7	23.7	56.7	36.9	41.4	39.5	31.4	44.3

Table 11. **Evaluation results on `VSI-Bench` 560 samples subset.**

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-Video-7B	35.7	49.0	12.8	48.6	21.4	40.0	43.5	34.0	36.0
LLaVA-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 12. **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-Video-7B	25.2	14.8	14.6	32.5	26.1	26.8	45.0	33.0	8.5
LLaVA-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 13. **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	8.7	20.3	0.3	10.8	29.2	5.2	2.9	2.5	-1.6
InternVL2-8B	9.9	-0.6	2.2	10.6	43.5	10.9	-5.8	-4.1	22.8
InternVL2-40B	12.6	35.9	-2.9	9.0	26.8	17.3	-5.0	9.9	20.0
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-Video-7B	10.5	33.8	-0.6	15.2	-1.9	16.7	-2.7	1.0	22.1
LLaVA-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 14. **Results of Vision Enabled − Vision Disabled.**