From Where Things Are to What They Are For: Benchmarking Spatial–Functional Intelligence in Multimodal LLMs

Le Zhang[1] Jihan Yang[2] Soundarya Krishnan[3] Jimit Majmudar[3]
Xiou Ge[3] Prasoon Puri[3] Prathamesh Saraf[3] Shruti Bhargava[3] Dhivya Piraviperumal[3]
Yinan Ling[3] Cindy Pan[3] Hong Yu[3] Aishwarya Agrawal[1] Bo-Hsiang Tseng[3]

[1]Mila - Québec AI Institute, UdeM [2]NYU [3]Apple

Code Project Page HuggingFace

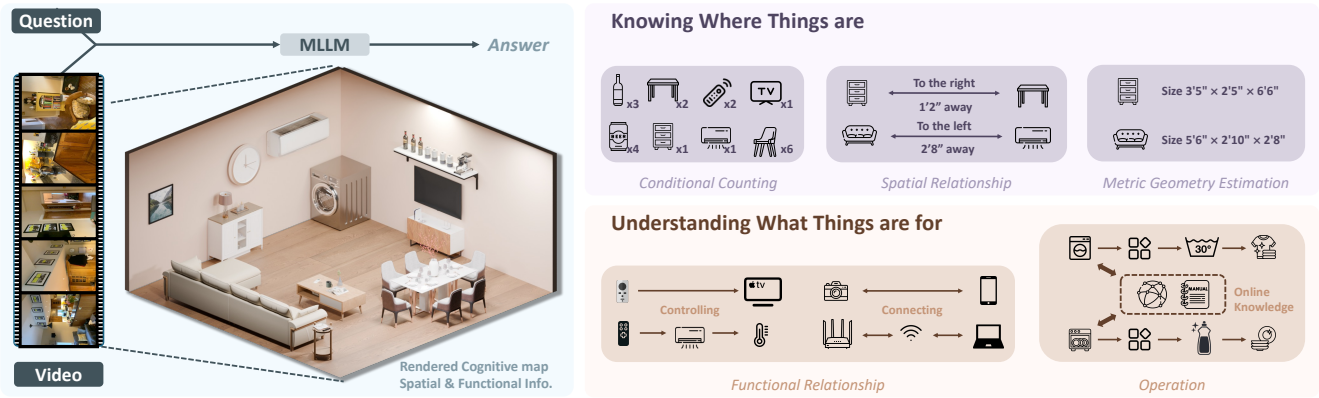


Figure 1. **From Spatial Cognition to Intelligent Agents. Left:** Task pipeline. A video is provided as input, and a multimodal model must reason across frames over both spatial and temporal context to answer a question. **Right:** Two complementary reasoning abilities evaluated in our benchmark. **Top (*Where Things Are*)**: spatial reasoning that requires understanding the scene layout and geometric relationships among objects (e.g., counting, directions, distances, and size). **Bottom (*What They Are For*)**: functional reasoning that requires understanding object affordances and functional relationships within the environment, enabling goal-oriented action and planning.

Abstract

Human-level agentic intelligence extends beyond low-level geometric perception, evolving from recognizing where things are to understanding what they are for. While existing benchmarks effectively evaluate the geometric perception capabilities of multimodal large language models (MLLMs), they fall short of probing the higher-order cognitive abilities required for grounded intelligence. To address this gap, we introduce the Spatial–Functional Intelligence Benchmark (SFI-Bench), a video-based benchmark with over 1,500 expert-annotated questions derived from diverse egocentric indoor video scans. SFI-Bench systematically evaluates two complementary dimensions of advanced reasoning: (1) Structured Spatial Reasoning, which requires understanding complex layouts and forming coherent spatial representations, and (2) Functional Reasoning, which involves inferring object affordances and their context-dependent utility. The benchmark includes tasks such as conditional counting, multi-hop relational reasoning, functional pairing, and knowledge-grounded troubleshooting, directly challenging models to integrate perception, memory, and inference. Our experiments reveal that current MLLMs consistently struggle to combine spatial memory with functional reasoning and external knowledge, highlighting a critical bottleneck in achieving grounded intelligence. SFI-Bench therefore provides a diagnostic tool for measuring progress toward more cognitively capable and truly grounded multimodal agents.

1. Introduction

Humans navigate and act in the world by forming internal *cognitive maps*—structured representations of where objects are and how they can be used. Such representations support tasks ranging from spatial memory to purposeful interaction with tools. For artificial agents, matching this capability requires moving beyond visual recognition to construct two complementary internal models: a **spatial map** that captures object layouts and relational structure, and a **functional map** that encodes affordances, purpose, and contextual usage.

Recent advances in Multimodal Large Language Models (MLLMs) [3, 14, 22, 24, 44, 60] have brought us closer to this goal, powering modern vision–language–action (VLA) systems [5, 9, 10, 26]. Yet, systematically evaluating whether these models truly acquire such spatial and functional intelligence remains challenging [1, 5]. Existing benchmarks such as VSI-Bench [63] primarily probe the *first* step of this developmental hierarchy—testing geometric perception and factual recall—while leaving the higher cognitive stages of structured map construction, affordance inference, and knowledge-grounded reasoning largely unexamined.

To close this gap, we introduce the **Spatial–Functional Intelligence Benchmark (SFI-Bench)**, which holistically evaluates cognitive intelligence across progressive stages (see Fig. 1). While prior works contain tasks labeled as counting or spatial relations, these are typically formulated as *perceptual* recognition problems. SFI-Bench instead reformulates them as *cognition-level* challenges: conditional counting demands logical and compositional inference over attributes and relations (*e.g., finding the maximum number of same-brand bottles on a cabinet*) and path reasoning requires integrating spatial cues across multiple views over time to infer a coherent global layout. These tasks incentivize models to build coherent internal representations of space rather than merely reacting to local cues.

Beyond spatial cognition, SFI-Bench incorporates functional and knowledge-grounded reasoning, probing whether models understand what objects in the scene are for, how they are operated, and how failures can be diagnosed. Tasks such as functional pairing, operational planning, and causal troubleshooting assess whether a model can bridge perception to action, mirroring the functional reasoning that underlies human goal-directed behavior. This shifts the evaluation from testing spatial memory to evaluating the broader pre-action cognitive abilities required for agentic behaviour.

Evaluating state-of-the-art MLLMs on SFI-Bench reveals a consistent pattern: while current models excel at local perception, they remain brittle in maintaining global spatial memory, grounding affordances, and composing multi-step functional plans. Our analyses uncover several key findings. First, longer reasoning chains do not lead to better decisions; reasoning quality saturates once a moderate budget is reached, beyond which overthinking introduces semantic drift. Second, cognitive map construction depends strongly on visual evidence rather than textual descriptions, and—unlike humans—models exhibit surprising insensitivity to temporal continuity. Third, systematic failure modes arise across tasks, including visual ambiguity, object recognition errors, spatial layout inconsistencies, and affordance overgeneralization.

In addition, SFI-Bench reveals the crucial role of *external knowledge acquisition*. For operational and troubleshooting tasks, GPT-5 can exhibit performance gaps exceeding 20 points depending solely on whether web search is enabled, highlighting that many functional questions fundamentally require grounding in up-to-date or device-specific knowledge. This underscores an often-overlooked challenge for multimodal reasoning systems [21, 41]: the need to seamlessly integrate visual perception with dynamic, external knowledge sources rather than relying on closed-world parametric memory alone.

Together, these findings point to a critical frontier for multimodal AI: moving beyond perception-oriented models toward systems capable of integrated spatial–functional cognition—constructing, maintaining, and exploiting coherent cognitive maps while flexibly retrieving and applying external knowledge to support purposeful action in real-world environments.

2. SFI-Bench

2.1. Dataset Overview

We introduce **SFI-Bench** to evaluate how multimodal foundation models acquire cognitive abilities for intelligent agents. SFI-Bench is a *video-based multiple-choice question answering* benchmark with 1561 *human-annotated* questions from 200 real-world egocentric indoor videos, sourced from ARKitScenes [4] and ScanNet++ [67], covering diverse spatial layouts and functional contexts in residential, professional, and industrial environments. SFI-Bench spans six core tasks grouped into two fundamental cognitive capabilities central to agentic intelligence:

2.1.1. Cognitive Spatial Reasoning.

These tasks assesse whether a model can move beyond frame-level perception to construct *structured cognitive spatial maps*. Rather than recognizing objects in isolation, the model must compositionally integrate attributes, absolute and relative positions, and multi-view spatial cues distributed across the video. This requires stitching together fragmented observations to form a coherent and temporally

Global and Conditional Counting. Reformulates counting as a *compositional & logical reasoning* task. Beyond simple enumeration, models must apply attribute constraints and perform set-based operations—such as intersec-

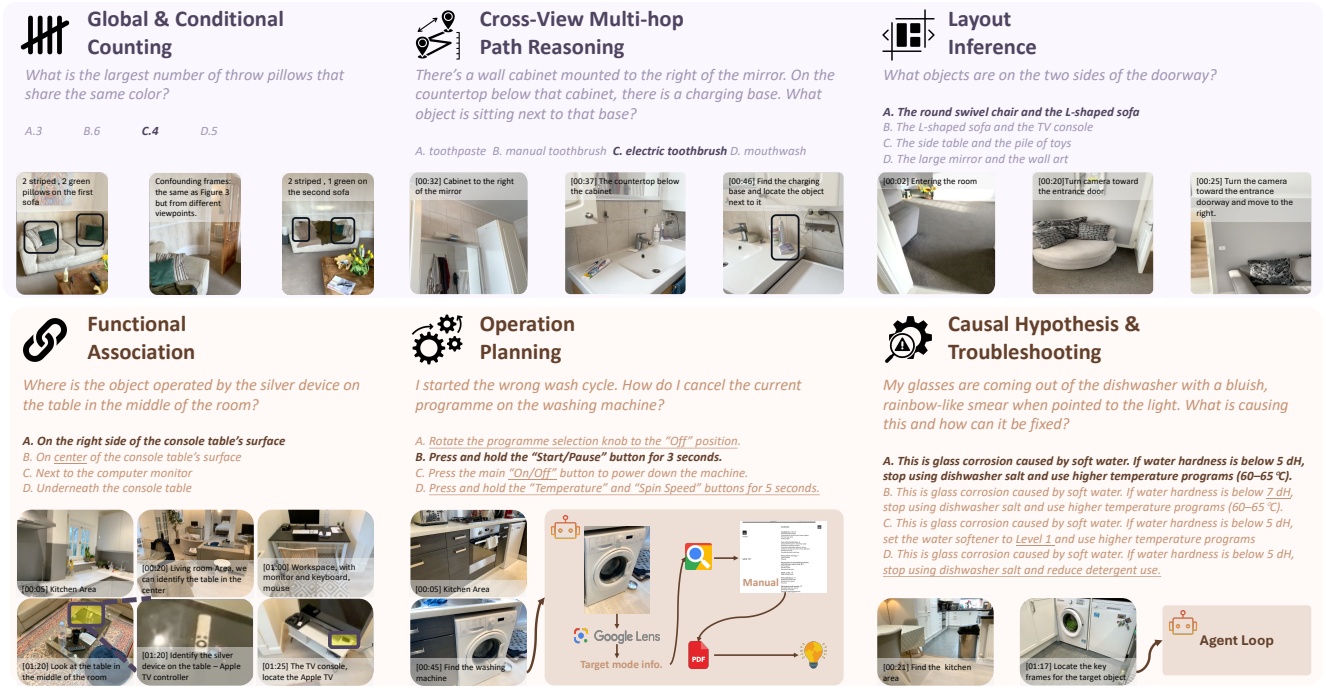


Figure 2. **Task examples in SFI-Bench.** Required grounding cues and reasoning evidence are highlighted using bounding boxes and accompanying text. For functional reasoning tasks, potentially confounding options are underlined. Answers are simplified for readability. For the final two functional reasoning tasks, successful completion requires online search to access relevant operational manuals.

tion, union, and complement—along with group-level aggregation (e.g., identifying the largest subset of same-brand bottles on a cabinet). This shifts counting from perceptual detection to structured logical inference.

Cross-View Multi-hop Path Reasoning. Evaluates the ability to integrate spatial evidence across time and viewpoints to infer relationships not visible in any single frame. Success requires constructing a coherent multi-hop spatial memory and recovering implicit connections between objects and locations beyond immediate perception.

Layout Inference. Evaluates whether the model can integrate distributed cues into a coherent global scene layout and reason about *occlusion relationships*. Because referenced objects often never appear together, the model must infer their relative arrangement and visibility ordering across frames. This reflects real-world navigation, where understanding occlusions is essential for building a consistent spatial map.

2.1.2. Functional Reasoning.

These tasks evaluate whether a model can move from spatial understanding to *functional cognition*—inferring object affordances, interactions, and context-dependent use. These tasks require integrating visual evidence with *external knowledge sources* (e.g., device manuals, online instructions), testing whether models can retrieve, interpret, and apply functional knowledge in real-world scenarios.

Functional Association. Tests whether the model can infer *affordance relationships* between objects. Objects often never co-occur in the same frame; thus, the model must link them through cues such as brand, design, or spatial context (e.g., associating a remote with the correct television), reflecting early functional map construction.

Operation Planning. Probes whether a model can determine *how an object should be used*. Solving these questions requires searching for device-specific information (e.g., manuals), interpreting retrieved knowledge, and assembling multi-step action plans grounded in the videos.

Causal Hypothesis and Troubleshooting. Assesses a model's ability to diagnose problems by combining scene understanding with external knowledge. The model must hypothesize plausible failure modes, consult relevant documentation via *web search*, and integrate the two sources to generate a grounded and actionable solution.

2.2. Benchmark Curation Process

SFI-Bench is constructed through a three-stage pipeline designed to produce high-quality, temporally grounded questions. (details in Appendix C, pipeline shown in Fig. 3).

Automatic Question Generation. We repurpose egocentric scans from ARKitScenes and ScanNet++ and use Gemini-2.5-Pro to extract fine-grained metadata for each video. Multiple passes of metadata extraction are merged

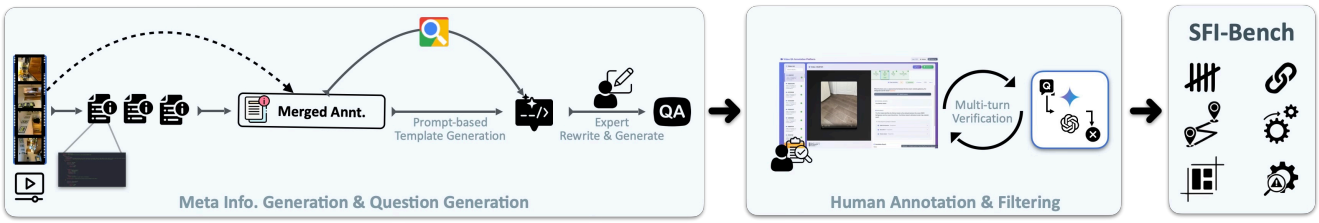


Figure 3. **Benchmark curation pipeline.** Metadata is extracted and consolidated across multiple MLLM passes, and combined with task-specific templates and few-shot examples to generate candidate questions. Annotators verify all questions and provide answers for the first four tasks, while answers for the two knowledge-grounded tasks are derived from expert-retrieved manuals. Finally, all questions undergo multi-turn human–AI collaborative post-hoc filtering to ensure quality and consistency. Samples that models fail to answer are manually rechecked, and questions that can be solved without visual grounding are removed.

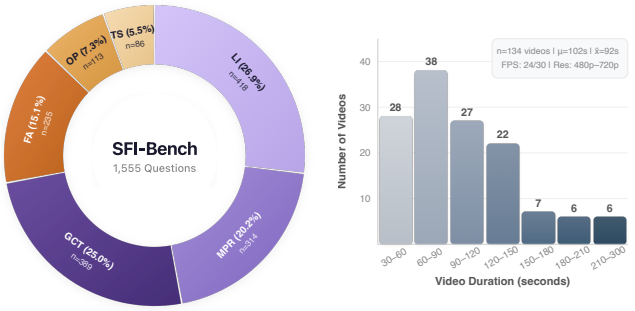


Figure 4. **Dataset Statistics.** Task and video length distribution. Task names are abbreviated for brevity.

and cross-validated against the raw video to obtain a reliable structured description of objects, attributes, spatial relations, and functional roles. Task-specific templates then generate candidate questions; for knowledge-grounded tasks, relevant manuals are retrieved online by annotators manually and integrated into the generation prompts. Annotators *refine all generated questions*, correct mismatches, and add additional items that better capture each task's cognitive requirements. The statistics are illustrated at Fig. 4

Human Verification and Answer Annotation. For the first four tasks, annotators watch each video, validate every question, and provide ground-truth answers based on visual evidence. For the two knowledge-grounded tasks, answers are automatically generated from retrieved device manuals.

Post-hoc Quality Filtering. All questions undergo automated and manual validation. Each item is first evaluated using Gemini-2.5 Pro and GPT-5, and any incorrectly answered case is then reviewed through a *multi-turn human–AI verification process* to diagnose potential issues and refine the question or options when necessary. Questions that can be solved without videos are removed to ensure visual dependence.

3. Benchmarking on SFI-Bench

3.1. Evaluation Setups

Baseline Models. We comprehensively evaluate a wide range of MLLMs *capable of processing video inputs*, spanning both open-source and proprietary systems. Among proprietary systems, we benchmark Gemini-2.5 [14], GPT-5.4/5 [44], and o4-mini under default configurations. Open-source models include Qwen3-VL [60], InternVL-3.5 [76], GLM-4.5 [23], LLaVA-OneVision [27], and LLaVA-Video [73]. All evaluations are conducted in a zero-shot setting using same prompt templates to ensure fairness and reproducibility.

Evaluation Metric. All questions in SFI-Bench are multiple-choice (MCQ), each with four candidate options (25% random chance). Performance is measured by answer accuracy. For the first four tasks—*Conditional Counting*, *Path Reasoning*, *Layout Inference*, and *Functional Association*—models are directly prompted to select the correct option. For the remaining tasks—*Functional Planning* and *Causal Hypothesis & Troubleshooting*—models equipped with search tools are allowed to retrieve external knowledge (e.g., user manuals) via web search before answering. Models without tool-use or web-access capabilities are evaluated in the same offline setting as the first four tasks.

3.2. Main Results

Table 1 presents the overall results on SFI-Bench.

Proprietary Models. Among proprietary systems, reasoning-enabled variants consistently deliver substantial gains. GPT-5.4-high achieves the strongest overall performance, outperforming GPT-5.4, while Gemini-2.5-Pro similarly surpasses Gemini-2.5-Flash across all tasks. Across all models, counting emerges as a key bottleneck, revealing persistent limitations in compositional and logical reasoning. While leading proprietary models exhibit strong capabilities in spatial cognitive map construction, their performance on functional reasoning tasks remains comparatively weaker. This gap becomes more pronounced on the

Methods	Rank	Avg.	GCT.	MPR.	LI.	FA.	OP.	TS.
Proprietary Models (API)								
†GPT-5.4-high ‡	1	72.1	58.4	82.8	81.1	76.2	65.5	68.8
†GPT-5 ‡	2	69.4	58.4	83.0	81.5	75.3	60.2	58.1
†GPT-5.4 ‡	3	67.3	54.5	79.6	83.0	66.4	63.4	57.0
†Gemini-2.5 Pro ‡	4	67.1	54.4	80.7	83.8	65.5	60.2	58.1
†o4-mini ‡	5	66.8	51.0	73.7	82.4	68.5	65.0	60.4
†Qwen3-VL-plus	6	58.1	51.3	64.3	73.6	61.3	50.4	47.7
†Gemini-2.5 Flash ‡	7	55.3	41.5	66.8	73.3	52.1	50.4	47.7
Open-source Instruct Models								
†Qwen3-VL-235B-A22B-Instruct	3	60.7	52.3	66.6	78.8	55.5	53.0	58.1
†Qwen3-VL-32B-Instruct	4	59.0	50.0	64.3	76.7	55.5	53.1	54.6
†Qwen3-VL-30B-A3B-Instruct	9	52.7	42.1	57.6	75.5	46.2	49.6	45.3
†Qwen3-VL-8B-Instruct	8	53.3	41.5	56.3	73.1	45.4	54.9	48.8
InternVL3.5-30B-A3B	5	55.9	48.5	59.5	74.0	50.0	51.3	52.3
InternVL3.5-14B	6	55.3	44.6	63.6	72.1	44.5	52.2	54.7
InternVL3.5-8B	7	53.9	44.4	57.9	69.0	46.6	53.1	52.3
LLaVA-OneVision-7B	11	50.4	40.5	57.3	60.2	44.5	49.9	50.3
LLaVA-OneVision-72B	2	61.3	52.8	64.2	68.6	60.1	58.4	61.0
LLaVA-Video-7B	10	50.9	55.4	61.1	67.9	49.2	38.9	32.6
LLaVA-Video-72B	1	64.9	57.9	70.3	75.2	56.7	58.4	50.9
Open-source Reasoning Models								
†Qwen3-VL-235B-A22B-Thinking	1	57.9	53.8	62.4	74.0	60.9	51.3	45.3
†Qwen3-VL-32B-Thinking	2	55.9	49.5	64.0	75.7	59.7	42.5	44.2
†Qwen3-VL-30B-A3B-Thinking	3	52.1	41.5	59.9	75.0	46.6	39.8	50.0
†Qwen3-VL-8B-Thinking	4	51.4	42.6	58.3	70.7	48.3	40.7	47.7
†GLM-4.5V-Thinking	5	45.1	28.7	53.5	65.5	41.2	42.5	39.5

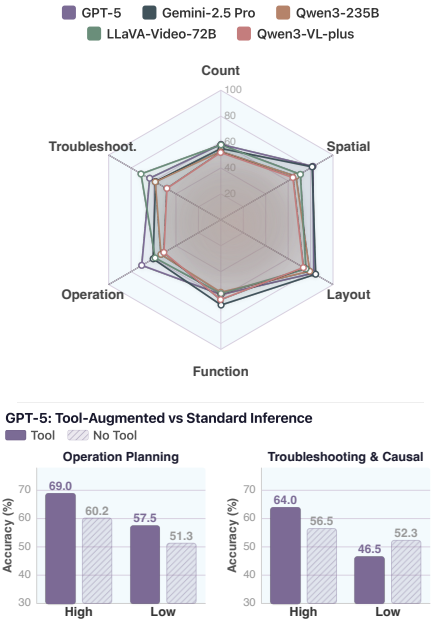


Table 1. **Evaluation on SFI-Bench. Left:** Avg. is macro-average accuracy. Dark gray indicates the best result among all models and light gray indicates the best result among open-source models. Cyan highlights spatial reasoning tasks, while Yellow highlights functional reasoning tasks. † indicates calling model with official API. ‡ indicates model with internet search tool for last two tasks. **Top Right:** Radar plot of the best-performing models. **Bottom Right:** GPT-5 performance on the last two tasks under different reasoning modes and with/without web search tool.

two knowledge-grounded tasks (Tab. 1 bottom right): GPT-5 equipped with web search tool significantly outperforms its offline counterpart under high reasoning budgets.

However, tool use introduces additional noise and can degrade performance when reasoning capacity is limited. Even within GPT-5, the low-reasoning variant performs worse with web search than without it on troubleshooting tasks. A similar trend is observed for Gemini-2.5-Flash and Qwen3-VL-Plus, where reasoning-enabled variants underperform their instruction-tuned counterparts. These findings suggest that strong reasoning ability is a prerequisite for effective tool use.

Open-source Models. Among open-source systems, video-based models such as LLaVA-Video-72B achieve strong spatial reasoning performance, even surpassing Gemini-2.5-Flash on several tasks. Nevertheless, the overall open-source ecosystem remains substantially behind API-based models. Global conditional counting persists as the primary bottleneck, while layout inference is comparatively easier. Models without internet access must rely solely on parametric knowledge, yielding accuracies near 50% on functional reasoning tasks.

Notably, open-source reasoning models show minimal improvement over instruct counterparts.

> Spatial and layout understanding can be improved with valid reasoning, but current open-source multimodal models fail to transfer reasoning capacity from visual math problems to spatial-function tasks.

4. Limitation of Current MLLMs

We investigate how modern MLLMs reason about *space* and *functionality*—specifically, how they construct and utilize cognitive spatial maps and functionality maps when solving complex multimodal reasoning tasks. By examining both reasoning traces and systematic failure patterns, we aim to uncover the mechanisms and limitations underlying spatial understanding, functional inference, and their interaction in current vision–language systems.

4.1. Probing via Reasoning Traces

To probe how state-of-the-art MLLMs internalize spatial and functional structure, we analyze the reasoning traces of Gemini-2.5 Pro. Modern MLLMs often employ reinforcement-learning-enhanced reasoning to produce explicit textual chains of thought [14, 20, 24], which provide a valuable lens into their decision processes [40, 58].

Case Study. Fig. 5 presents representative examples of spatial and functional reasoning. In successful cases, the model decomposes questions into manageable subtasks (e.g., lo-

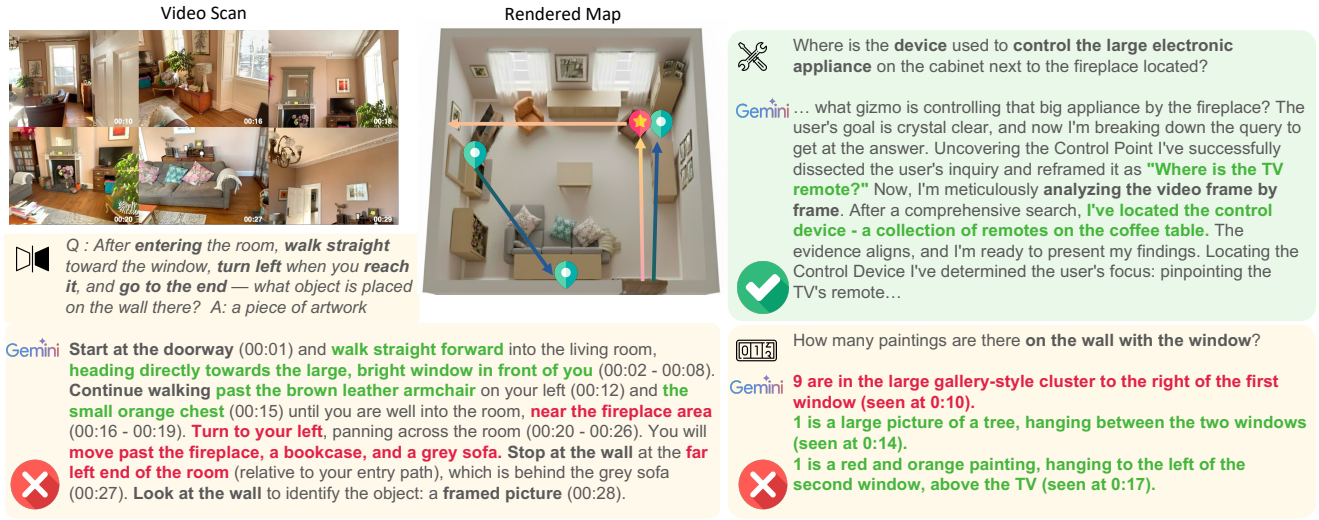


Figure 5. **Human-annotated analysis of MLLM spatial–functional reasoning.** The central panel shows a conceptual cognitive map reconstructed from the video and annotated based on the model's reasoning. Red arrows denote the ground-truth path, while blue arrows indicate incorrect navigation. Key steps are labeled, with correct and incorrect reasoning highlighted, revealing both spatial reconstruction and failure modes in model's reasoning process.

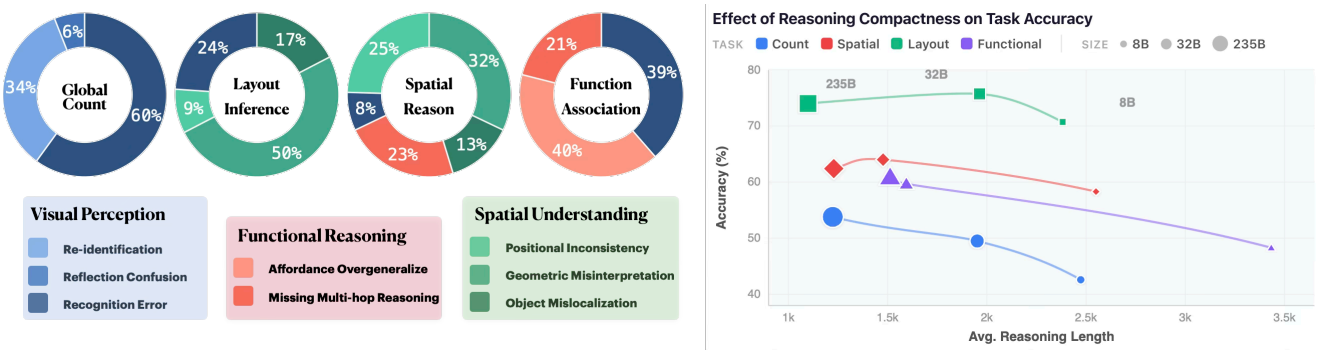


Figure 6. **Left:** Human analysis cattegorizing task specific failures. **Right:** Relationship between reasoning compactness and task accuracy. Colors denote task types, while point size encodes Qwen3-VL model scale (8B, 32B, 235B). Regression lines indicate that larger models tend to produce shorter reasoning chains, which consistently correlate with higher accuracy.

cating relevant regions, aligning cross-view cues, inferring relationships), integrates multi-frame evidence, and constructs a coherent internal scene representation. These behaviors suggest early-stage cognitive spatial mapping, reflecting an emergent understanding of object permanence and spatial continuity.

However, performance degrades when performing multi-hop reasoning. In cross-view reasoning (left), the model exhibits fragmented spatial continuity, "jumping" between distant objects (e.g., from a chest to a fireplace) without maintaining consistent self-location. In conditional counting (bottom right), the model misinterprets contextual modifiers (e.g., "on the wall with the window"), revealing a gap between linguistic conditioning and geometric grounding.

General functional reasoning tasks amplify these challenges. While the model can recognize object affordances (e.g., linking a coffee machine with mugs), it struggles with multi-step reasoning, such as identifying prerequisites ("fill with water") or understanding control dependencies ("the remote operates the TV"). In causal troubleshooting, it fails to integrate visual cues with external knowledge, often generating superficial or hallucinated explanations instead of grounded diagnoses.

Failure Mode Analysis. To identify the sources of failure in state-of-the-art MLLMs' reasoning on *spatial–functional intelligence*, we analyze 120 erroneous samples produced by Gemini-2.5-Pro across the first four tasks (30 samples per task). We identify three main categories of failure

modes (see Appendix for details):

1. **Visual Perception:** Failures in object recognition and visual interpretation, including *missing objects*, *misclassification*, *attribute mislabeling*, *re-identification failure*, and *reflection confusion*.
2. **Spatial Understanding:** Errors in maintaining spatial consistency and inferring spatial relation, such as *positional inconsistency*, *geometric misinterpretation*, and *object mislocalization*.
3. **Functional Reasoning:** Failures related to the model's ability to understand functional relationships and perform grounded, compositional reasoning. This includes: *Affordance overgeneralization*, where the model assumes functional relationships based on commonsense (e.g., assuming any remote controls a TV) without verifying the specific context; and *Missing multi-hop reasoning*, where the model fails to complete complex multi-step inferences over functional chain of objects.

Fig. 6 (left) shows that error patterns across tasks. It highlights that visual perception issues are common across all tasks, pointing to the ongoing difficulty of video understanding for existing MLLMs. Spatial understanding is critical for spatial and layout reasoning tasks, while functional reasoning errors primarily affect spatial reasoning and functional association, where the model often overlooks fine-grained functional relationships and struggles with object recognition. This trend reveals a clear progression:

> As tasks become more cognition-intensive, errors shift from perceptual ambiguities to deficiencies in reasoning and spatial–functional grounding.

4.2. Limited Reasoning Gains from RLVR

Tab. 1 reveals that open-source reasoning variants trained with reinforcement learning with verifiable rewards (RLVR) offer only marginal improvements over their instruction-tuned counterparts. To investigate this, we analyze the reasoning lengths of the Qwen3-VL family across model scales and tasks (Fig. 6 right; Fig. 7). Across all settings, longer reasoning does not correlate with higher accuracy. Notably, samples that were answered correctly in the non-reasoning mode but failed in the reasoning mode exhibit substantially longer reasoning chains—1.41×, 1.12×, and 1.22× longer than the global averages for the 235B, 32B, and 8B models, respectively (Fig. 7 left). The above results suggests that excessive reasoning often leads to over-explanation and semantic drift rather than deeper inference.

As shown in Fig. 6 (right), larger models (235B) generate shorter yet more effective reasoning, achieving higher accuracy through compact, well-grounded planning. Smaller models (8B), by contrast, rely on verbose but shallow reasoning to compensate for weaker internal representations. Overall, reasoning length proves to be an unreliable indicator of reasoning quality—beyond a moderate range (≈1.2–1.5k tokens), linguistic noise dominate.

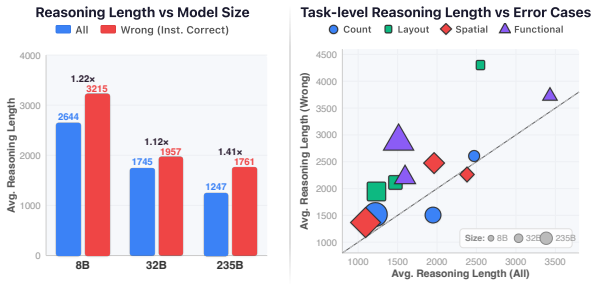


Figure 7. **Left:** Average reasoning lengths across Qwen3-VL model scales (8B, 32B, 235B). *Wrong* denotes cases where the reasoning model failed while the instruction model succeeded. **Right:** Task-level reasoning length comparison across models. The Y-axis shows the average token length for *Wrong* cases, and the X-axis for all cases. Points above the diagonal and that lie farther away indicate tasks where models tend to produce unnecessarily long and erroneous reasoning chains.

SR	Overall	Count	Layout	Spatial	Func.
0	75.5	60.0	88.0	82.0	72.0
25%	75.5	62.0	88.0	80.0	72.0
50%	71.5	50.0	84.0	80.0	72.0
75%	68.5	50.0	78.0	76.0	70.0
100%	75.0	58.0	86.0	78.0	78.0

Table 2. **Effect of frame order shuffling rate (SR) on task accuracy (%)** Higher SR indicates stronger temporal disruption.

At the task level (see Fig. 7 and Fig. 6), *functional association* shows the most pronounced overextension effect, where longer reasoning chains lead to increased failure rates. This indicates that the model tends to overthink, resulting in overly long reasoning chains that introduce noise. *Spatial reasoning* remains relatively stable across lengths, while *layout inference* shows occasional inflation caused by repetitive scene descriptions. *Counting* is the most concise and stable task, reflecting strong visual grounding.

> Effective reasoning is not about producing longer explanations but about maintaining alignment with perceptual evidence. Concise, grounded, and semantically coherent reasoning yields the most reliable improvements across tasks and model scales.

5. Determinants of Performance on SFI-Bench

5.1. Temporal Consistency Dependence

Humans construct cognitive maps by integrating perceptual inputs over continuous time. Neuroscience studies show that temporal continuity is essential for spatial learning, as hippocampal and entorhinal systems encode spatial

relations through sequential experience [16, 43]. To test whether large multimodal models share this property, we progressively shuffled the input frame order for GPT-5 at different shuffle ratios (SR). As shown in Table 2, performance remained largely stable, with only moderate degradation in counting tasks. This experiment was conducted on a subset of 200 samples (50 per task type), and the results indicate that MLLMs rely primarily on aggregated visual evidence rather than temporally coherent dynamics, forming a static spatial abstraction rather than a time-dependent cognitive map.

5.2. Visual Grounding vs. Textual Descriptions

Input	Count	Layout	Spatial	Func.
Visual	58.4	83.0	81.5	75.3
Caption-only	57.2	51.6	55.4	67.6

Table 3. GPT-5 performance comparison between visual and textual inputs. The model's performance is shown for both visual (with full context) and textual (with local descriptions) conditions.

To assess whether cognitive map construction relies on global visual signals or can be inferred from structured textual descriptions, we evaluate GPT-5 on 200 samples under two conditions: (1) full visual input, and (2) structured textual input derived from captions, similar to socratic models approach [70]. The captions are generated by the same model by first processing the video and producing detailed descriptions of object attributes, relations, and layout.

As shown in Tab. 3, performance remains strong with visual input but drops consistently when only structured textual descriptions are provided. The degradation is most pronounced in tasks requiring spatial and layout understanding.

These results suggest that even detailed, model-generated textual descriptions are insufficient for accurate cognitive map construction, highlighting the critical role of direct visual grounding.

6. Related Work

6.1. Multimodal Large Language Models

Multimodal large language models (MLLMs) have achieved significant success [2, 12, 25, 28, 30, 35, 51, 53, 57] by effectively integrating rich visual semantics from vision encoders [46, 56, 71] with the sophisticated reasoning abilities of LLMs [7, 55, 60]. As a natural extension of this momentum, this architecture has been adapted for the temporal domain, leading to the rapid development of video-based MLLMs [3, 28, 31, 32, 39, 50, 72, 74, 76, 77] sampling separate frames in video and concatenating their per frame features. Concurrently, evaluation is also fundamental to drive model evolution. Massive VQA benchmarks have emerged [18, 37, 38, 53, 54, 61, 69], to test knowledge recall and semantic understanding of MLLMs. However, most existing benchmarks primarily focus on content or activity level understanding, ignoring a fundamental primitive: spatial layout within the video. Recent works are beginning to draw the community's attention to this gap, emphasizing the need to examine the spatial reasoning and underlying world understanding capabilities of MLLMs in both images and videos [62, 66, 68].

6.2. Visual Spatial Intelligence

Spatial intelligence refers to the ability to perceive, represent, and reason about spatial relationships, a foundational concept in cognitive psychology [19, 42, 49]. Humans excel at building mental maps of their surroundings, performing egocentric–allocentric transformations, and leveraging spatial memory for navigation and problem-solving. For MLLMs to perceive and interact with the physical world, a robust understanding of spatial relationships from visual inputs is crucial. Recognizing this requirements, recent research has advanced in two parallel directions: designing physically and spatially grounded benchmarks [6, 33, 37, 47, 52, 59, 62, 66, 68], and developing new methods to enhance the spatial reasoning capabilities of MLLMs [8, 11, 13, 15, 17, 29, 34, 36, 45, 48, 64, 65, 75].

While prior efforts focus on spatial reasoning in MLLMs, we extend this scope, moving beyond prior efforts that focused solely on spatial layout from video, and incorporating an evaluation of object functionality and potential interactions.

7. Conclusion and Future Work

We introduced SFI-Bench, a benchmark that advances multimodal evaluation from basic perception to spatial reasoning and functional understanding. Our results show that while current MLLMs excel at basic perception, they struggle with maintaining spatial memory, integrating functional knowledge, and linking perception to external knowledge. Open-source models face difficulties transferring reasoning from visual math problems to spatial-functional tasks, highlighting the need for further exploration. Furthermore, web search is essential, as models leveraging external knowledge significantly outperform offline variants.

Future work should focus on improving memory mechanisms for retaining spatial representations, developing more robust compositional reasoning frameworks, and expanding the external knowledge dimension to enable retrieval-augmented reasoning for intelligent spatial agents.

References

[1] Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*, 2025. 2

[2] 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. 8

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 8

[4] 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. 2

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2

[6] Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025. 8

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 8

[8] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *ICRA*, 2025. 8

[9] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, et al. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025. 2

[10] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024. 2

[11] 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

[12] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 8

[13] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 8

[14] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 4, 5

[15] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *ACL*, 2024. 8

[16] Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513, 2017. 8

[17] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 8

[18] Chaoyou Fu, Yuhan Dai, Yongdong 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. In *CVPR*, 2025. 8

[19] Howard Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011. 8

[20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5

[21] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025. 2

[22] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 2

[23] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 4

[24] 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, 5

[25] 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. 8

[26] 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. *arXiv preprint arXiv:2406.09246*, 2024. 2

[27] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, 2024. 4

[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 8

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

[30] 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. 8

[31] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 8

[32] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 2024. 8

[33] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *ICCV*, 2025. 8

[34] Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondences boost spatial-temporal reasoning in multimodal language model. In *CVPR*, 2025. 8

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 8

[36] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *NeurIPS*, 2025. 8

[37] 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

[38] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 8

[39] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 8

[40] Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. Deepseek-r1 thoughtology: Let's think about llm reasoning. *arXiv preprint arXiv:2504.07128*, 2025. 5

[41] Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*, 2025. 2

[42] Nora S Newcombe. Spatial cognition. *Memory and Cognitive Processes*, 3:113–163, 2004. 8

[43] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978. 8

[44] OpenAI. GPT-5 System Card. Online, 2025. Accessed: October 8, 2025. 2, 4

[45] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 8

[46] 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

[47] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *ICLR*, 2025. 8

[48] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Spatial Aptitude Training for Multimodal Language Models. In *COLM*, 2025. 8

[49] Roger N Shepard and Lynn A Cooper. *Mental images and their transformations.* The MIT Press, 1986. 8

[50] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024. 8

[51] 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. 8

[52] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. 8

[53] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *NeurIPS*, 2024. 8

[54] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 8

[55] 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. 8

[56] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 8

[57] 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

[58] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025. 5

[59] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025. 8

[60] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 4, 8

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

[62] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. In *CVPR*, 2024. 8

[63] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 2

[64] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025. 8

[65] Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025. 8

[66] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025. 8

[67] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2

[68] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025. 8

[69] 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. 8

[70] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 8

[71] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 8

[72] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023. 8

[73] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. *URL https://arxiv. org/abs/2410.02713*, 17. 4

[74] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *TMLR*, 2025. 8

[75] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. In *ICCV*, 2025. 8

[76] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4, 8

[77] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *CVPR*, 2025. 8

From Where Things Are to What They Are For: Benchmarking Spatial–Functional Intelligence in Multimodal LLMs

Supplementary Material

.1. Influence of Reasoning Budget

In the last section, we analyzed the open-source model's performance based on its scale without explicitly controlling the reasoning parameter. However, to investigate whether RL-based optimization improves reasoning on SFI-Bench, we now **actively control the *reasoning budget***—the maximum token allowance for the model's reasoning chain—on proprietary models. By varying the budget (see Fig. 8), we can directly assess how changes in reasoning depth affect performance across all SFI-Bench tasks.

Both GPT-5 and Gemini-2.5-Pro show accuracy gains with increased reasoning budgets, indicating that extended reasoning allows for better integration of spatial and functional cues. However, the performance curve flattens beyond approximately 2k tokens, as further exploration reveals that Gemini-2.5-Pro does not utilize reasoning beyond 2k tokens, reaching its limit in reasoning depth at this point.

By examining the reasoning content, we observe that longer reasoning budgets tend to reveal interesting patterns, such as frequent self-checking and the development of complex, ordered plans for task-solving.

Overall, the results suggest that reasoning efficiency, rather than token capacity alone, drives performance. The model's reasoning chain has an upper limit, with a moderate reasoning budget (around 2k tokens) offering an optimal balance between expressivity and stability. Beyond this point, further expansion yields diminishing returns, as good models do not continue reasoning indefinitely but instead reach a point of effective problem-solving.

A. Task Examples

Additional task examples are provided in Fig. 9 and Fig. 10.

B. Detailed Failure Mode Categorization

Below is the detailed failure mode categorization for the analysis introduced in Sec. 4.1:

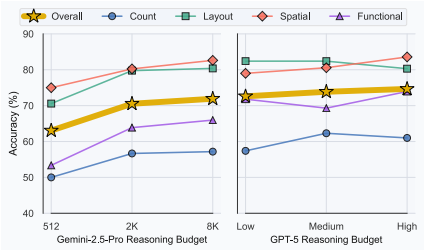
1. **Visual Perception:** Failures related to object recognition and visual data interpretation. This includes: *Missing objects*, where the model overlooks visible entities; *Object misclassification*, where objects are assigned the wrong labels; *Attribute mislabeling*, where attributes such as color, size, or brand are incorrectly identified; and *Re-identification failure*, where the model mistakenly counts the same object multiple times when viewed from different perspectives. Additionally, *Reflection confusion* occurs when the model misinterprets mirror reflections as real objects.
2. **Spatial Understanding:** Errors related to the model's ability to maintain consistent and accurate spatial representations. This includes: *Positional inconsistency*, where objects shift positions or lose continuity across frames (e.g., teleportation effects); *Geometric misinterpretation*, where the model fails to infer correct geometric relationships between objects (e.g., alignment or linearity); and *Object mislocalization*, where the model places objects in incorrect locations, such as confusing left/right or near/far positioning.
3. **Functional Reasoning:** Failures related to the model's ability to understand functional relationships and perform grounded, compositional reasoning. This includes: *Affordance overgeneralization*, where the model assumes functional relationships based on commonsense (e.g., assuming any remote controls a TV) without verifying the specific context; and *Missing multi-hop reasoning*, where the model fails to complete complex multi-step inferences over functional chain of objects.

C. Data Curation Process Details

This section provides additional details on the dataset construction process, including the metadata format, prompting setup, and human annotation pipeline. Unless otherwise specified, Gemini-2.5-Pro is used as the default MLLM assistant. For clarity, the prompts shown here are lightly streamlined.

Metadata Generation. We begin by generating structured metadata for each video using an MLLM. The metadata includes global scene descriptions, object instances with timestamps, category labels, fine-grained attributes (e.g., color, material, brand), spatial relations, and functional tags. An example (simplified for readability) is shown in Fig. 11. The system prompt used for metadata generation is provided in Fig. 12. Outputs from multiple passes are merged and subsequently verified against the original video.

Task Templates. Given high-quality metadata, task-specific questions are produced using carefully designed templates. The templates for each task are shown in: Fig. 13 (global conditional counting), Fig. 14 (cross-view multi-hop path reasoning), Fig. 15 (layout inference), and Fig. 16



Q: What is the appliance **on the far left** of the counter **surface on the right side** as you enter? A: Juicer

*Okay, let's break this down… "As you enter" is crucial – it sets the frame of reference. This means the perspective is from the doorway, looking into the kitchen. "On the right side" narrows it down to the counters and cabinets to the right as I enter. And "counter surface" means I'm looking for something *on* the countertop, not built-in or hanging… It starts at the doorway. I need to fast-forward and look for the right-hand side of the kitchen, specifically the counter… The answer is "Coffee maker"* ***--- Gemini 2.5 Pro Low thinking budge 273 token***

*right, let's break this down… I need to understand the terms… As you enter" is key; it's the camera's initial perspective. "Right side" is the countertop on the right wall from that point, and "far left" pinpoints the appliance at the very end of that counter run… Now, which counter is the question referring to? … Now, which is it? The question phrasing is a bit ambiguous... It could refer to *either* counter run on the right side. I need to be precise. Next, I look at the multiple-choice options … Yes, bread bin.* ***--- Gemini 2.5 Pro Medium thinking budge 482 token***

*… I need to nail this multiple-choice question… It starts with a wide shot, then moves in. There's a kitchen on the right. The camera pans along the counter. Aha**!** There's a black and silver something-or-other at the beginning. It looks like a coffee machine or a hot water dispenser… I **rewatch** parts of the video ... So, I **re-read** the question… I go back to the video to re-examine the two key locations… Then I try to think about the camera perspective… I focus on the camera moves… I go back to `01:14` for a final check of the "right side" counter. Okay, Bread bin, Juicer, Coffee Maker, Toaster. I then check my reasoning.1. Right side as you enter, means the whole kitchen on your right as you enter.2. The counter.3. Far left means the "point" on the counter that is furthest to the left.4. From the entrance, that is the counter by the window, the far wall.5. The objects on that part of the counter, from right to left are: the Toaster, Coffee Maker, Juicer, and then the Bread Bin.6. The question is "What is the appliance on the far left?"7. That would be the juicer (if we are excluding the bread bin)… The answer is "Juicer"* ***--- Gemini 2.5 Pro High thinking budge 1121 token***

Figure 8. **(Left)** Performance of Gemini-2.5 Pro and GPT-5 under different reasoning budgets. **(Right)** illustrate how increased reasoning budget alters reasoning patterns, higher budgets reveal frequent reasoning behaviors such as self-rechecking and complex reasoning chain.

(functional association). Each template is tailored to ensure consistency, task fidelity, and sufficient reasoning complexity.

Human Verification. The dataset is curated by 11 experienced annotators, all trained machine learning engineers. Each question is manually reviewed by at least one expert to ensure correctness and alignment with the associated video. Annotations are performed using an HTML-based platform that displays the question-generation reasoning chain, relevant objects with timestamps, and the corresponding video interface, as illustrated in Fig. 17.

D. Evaluation Setup

All videos are preprocessed to a fixed resolution of **720p** and **24 fps** before being uploaded to any model. This standardization reduces data size and ensures consistent visual quality across systems, preventing performance differences arising from input variability.

For proprietary models, we evaluate the Qwen3-VL family through the official **Aliyun API**, Gemini models via the **Google Cloud API**, and GPT models using the **OpenAI API**. All API-based experiments were conducted during the first two weeks of **September 2025** to ensure fair comparison under stable model deployments.

For open-source models, we adopt the **VLMEvalKit** framework to provide a unified inference pipeline and standardized evaluation protocol. This setup enables consistent multi-image and video processing across LLaVA-based, InternVL-based, and GLM-based models.

To promote reproducibility, we will release our **evaluation scripts**, **preprocessing pipeline**, and the full **SFI-Bench dataset** upon publication.

Global & Conditional Counting

How many pieces of furniture at or next to the desk contain metal?

Answer: 5

Functional Association

Where is the object that provides the content for the black rectangular device with a red and black stand?
Answer: On right side of the topped desk.

Cross-View Multi-hop Path Reasoning

Starting from the gaming chair, what electronic device is on the tall metal device that is positioned next to the desk?

Answer: headphones

Functional Association

Where is the object with a white frame and burgundy compartment offers a place for both resting and putting things away?

Answer: Opposite the window.

Layout Inference

Which obstacle is blocking the straight-line walking route between the Red Entrance Door and the Hallway Trash Can?

Answer: The Shoe Rack.

Layout Inference

What object should be removed to enable a direct straight path from the Laundry Basket to the Radiator?

Answer: The Standing Fan.

Global & Conditional Counting

How many power outlets are there on the wall opposite the entrance?

Answer: 12

Operational Planning

How do I load paper into the main paper tray for the printer on the desk?

Answer: Pull the main tray (Tray 2) completely out of the printer, open the paper guides, place the paper stack in the tray, adjust the guides to fit, and then reinsert the tray into the printer.

Causal Hypothesis & Trouble Shooting

The printer on the desk is pulling multiple sheets of paper at once from the tray. What's causing this and how do I fix it?

Answer: This is caused by the paper stack or tray condition. To fix it, remove the paper from the tray, flex it, rotate it 180 degrees, and flip it over. Also, ensure the tray is not overfilled and that the paper guides are adjusted correctly.

Global & Conditional Counting

How many heavy-duty staplers are the long-reach type?
Answer: 2

Global & Conditional Counting

How many office supplies with the same function are there on the desk?

Answer: 4

Figure 9. Representative task examples from SFI-Bench. Note that not all videos are equally suitable for every task; for instance, the bottom example depicts a small office with overly simple spatial structure, making it difficult to generate challenging layout or spatial reasoning questions.

Global & Conditional Counting

How many of the wall-mounted reading lights are turned on?

Answer: 2

Cross-View Multi-hop Path Reasoning

What appliance is on the far right of the countertop directly opposite the bed?
Answer: headphones

Answer: Coffee Machine

Layout Inference

Is there a direct path from the purple armchair to the bed?

Answer: Yes

Operational Planning

How do I make a cappuccino using the coffee machine on the counter?

Answer: Connect the suction hose to the spout and place the other end in the milk container. Place a cup under the spout. Turn the rotary control until 'CAPPUCCINO' is displayed, then press the rotary control once to start the process.

Causal Hypothesis & Trouble Shooting

The coffee machine's orange service light above the buttons is on continuously and isn't flashing. What does this mean and what should I do?

Answer: A solid orange light means the machine is clogged with scale and needs to be descaled.

Operational Planning

How do I replace the small, round CMOS battery on the motherboard of the dell desktop computer on the bottom of the right most shelf?

Answer: After opening the case, locate the coin-cell battery on the system board. Press the release latch away from the battery to allow it to pop up, then lift it out. To install the new one, press it into the slot until the latch secures it.

Causal Hypothesis & Trouble Shooting

The power button on the dell desktop computer on the bottom of the right most shelf is blinking amber. What does this mean?

Answer: A blinking amber power light indicates that a problem has occurred with the system board.

Global & Conditional Counting

What is the largest number of desktop computers that are all the same model?

Answer: 12

Operational Planning

How do I open the side cover to access the internal components of the dell desktop computer on the bottom of the right most shelf?

Answer: Lay the computer on its side, then lift the cover-release latch located on the top. Lift the cover upward to a 45-degree angle and remove it from the chassis.

Causal Hypothesis & Trouble Shooting

The diagnostic lights on the front of the dell desktop computer on the bottom of the right most shelf are showing lights 3 and 4 are on. What is the problem and what should I do?

Answer: This light pattern indicates a memory power failure has occurred. To fix it, you should remove all the memory modules, then reinstall one module and restart the computer. If it starts, continue installing modules one by one to identify the faulty one.

Global & Conditional Counting

How many Fujitsu Esprimo desktop computers are present?

Answer: 7

Figure 10. Representative task examples from SFI-Bench. Note that not all videos are equally suitable for every task; for instance, the bottom example depicts a small office with overly simple spatial structure, making it difficult to generate challenging layout or spatial reasoning questions.

```
{
  "video_file": "41069048.mp4",
  "scene_overview": {
    "room": "Bathroom",
    "style": "Modern, clean, functional, hotel-like",
    "palette": "White, grey-blue, chrome"
  },
  "objects": [
    {
      "id": "toilet_001",
      "class": "Toilet",
      "prominent_ts": "00:10",
      "vis_segments": [ {"start": "00:00", "end": "00:02"}, {"start": "00:08", "end": "00:13"} ],
      "attributes": { "type": "Two-piece", "mat": "ceramic", "color": "white" },
      "state": { "lid": "closed", "condition": "clean" },
      "location": "Positioned between trash can and bathtub, against wall",
      "functionality": { "primary": "Waste disposal", "secondary": [] },
      "relations": {
        "type": "functional_group",
        "related": ["toilet_brush_001", "trash_can_001", "toilet_paper_holder_001"]
      }
    },
    {
      "id": "bathtub_001",
      "class": "Bathtub",
      "prominent_ts": "00:19",
      "vis_segments": [ {"start": "00:11", "end": "00:20"} ],
      "attributes": { "type": "Shower-tub combo", "mat": "acrylic", "color": "white" },
      "state": { "fill_level": "empty", "condition": "clean" },
      "location": "Adjacent to sink and toilet, against wall",
      "functionality": { "primary": "Bathing/Showering", "secondary": [] },
      "relations": {
        "type": "integrated_system",
        "core_components": ["shower_system_001", "shower_screen_001", "grab_bar_001"]
      }
    }
    ...
  ],
"spatialLayout": {
      "mainPathway": "The camera moves in a circular path around the small bathroom, starting near the toilet, panning
           up the wall, across the ceiling, down to the shower/tub, across to the sink, and back towards the door.",
      "relativePositions": "The toilet and heated towel rail are on one side of the room. The bathtub and sink are on
          the opposite side. A large mirror is mounted above the sink.",
      "anomaliesOrAbsences": "A DVD case is on the bathroom floor, which is an unusual location for such an item. The
          toilet paper holder is empty."
    },

"functionalEcosystem": {
         "hygiene_and_grooming": {
            "core_objects": [
               "sink_001","toilet_001","bathtub_001","shower_system_001"
            ],
            "supporting_objects": [
            "faucet_001","soap_bar_001","mirror_001","towel_radiator_001","towel_001","towel_002","towel_003","
                toilet_brush_001","trash_can_001"
            ],
            "description": "A complete system for personal hygiene, including washing, bathing, grooming, and waste
                disposal, with all necessary fixtures and accessories present."
         }
      }
}
```


Figure 11. Meta Information Examples.

System Prompt: Meta Information Generation

TASK: Analyze this video and generate a comprehensive, structured JSON representation of the scene and its contents. The goal is to create a rich, machine-readable format suitable for detailed Q&A and object-level analysis. **OUTPUT FORMAT:** `Pure JSON object only` **KEY PRINCIPLES:**

- **Unique Instance Tracking:** Every discrete object (e.g., each individual cup, book, or chair) should be a unique entry in the `objectInventory`. Assign a persistent `instance_id` to each object (e.g., `book_001`, `mug_001`). This ID is crucial for referring to a specific object unambiguously, even if it looks identical to others or reappears after being hidden. For this reason, avoid using a summary "quantity" field.
- **Structured Properties:** Favor structured key-value pairs within an `attributes` object over a single, long description string. This allows for more precise and easily queryable information about each object's visual characteristics.
- **Temporal and State Awareness:** Accurately document when objects are visible and how their state might change. Use `visibility_segments` to track an object's presence over time, which is essential for understanding dynamic scenes with occlusions. For each object, identify the `most_prominent_timestamp` — the single moment when the object appears clearest and largest in the frame.
- **Functionality and Relationship Analysis:** For each object, analyze and describe its primary functionality and its relationships with other objects. Use reasoning to infer functional connections, especially when objects share brands or have complementary purposes.
- **Accuracy and Completeness:** Be as exhaustive and accurate as possible. Capture both large furniture/appliances and smaller, everyday items like books, bottles, cups, toys, remote controls, or electronic gadgets. If any text is clearly legible (e.g., a brand name, a book title), capture it. If a piece of information cannot be determined, use a `null` value for that key.

Here is the required JSON structure. The examples illustrate how to apply the principles above:

```
{
  "sceneOverview": {
    "roomType": "Living Room with Open Kitchen",
    "styleAndAtmosphere": "Modern minimalist, bright with good natural light",
    "mainColorPalette": "Primarily white and natural wood tones, with accents of blue"
  },
  "objectInventory": [
    {
      "instance_id": "phone_charger_001",
      "object_class": "Phone Charger",
      "visibility_segments": [ { "start": "00:08", "end": "01:25" } ],
      "attributes": {
        "type": "USB-C cable with wall adapter",
        "color": "white",
        "cable_length": "1 meter",
        "brandAndModel": "Apple 20W USB-C Power Adapter",
        "recognizedText": "Apple"
      },
      "state": {
        "connection_status": "plugged into wall outlet",
        "cable_condition": "coiled neatly"
      },
      "relational_location": "On the nightstand next to the bed, near the wall outlet.",
      "functionality_relation": {
        "target_objects": ["iphone_001", "ipad_001"],
        "reasoning": "Apple charger is specifically designed for Apple devices, and there's an iPhone visible on
            the nightstand with the same charging port",
        "relationship_type": "power_supply",
        "compatibility": "brand_specific"
      }
    }
  ],
  "spatialLayout": {
    "mainPathway": "The path from the room's entrance to the sofa area is clear, but a floor lamp partially
        obstructs the path between the sofa and the balcony.",
    "relativePositions": "The bookshelf is to the left of the sofa. The window is on the south wall of the room
        .",
    "anomaliesOrAbsences": "A dumbbell is visible on the kitchen counter, which is unusual for a kitchen."
  },
  "functionalEcosystem": {
    "entertainment_zone": {
      "core_objects": ["tv_001", "sofa_001", "remote_control_001"],
      "description": "Integrated entertainment setup where TV, seating, and control devices work together"
    },
    "connectivity_infrastructure": {
      "core_objects": ["router_001"],
      "dependent_objects": ["smart_tv_001", "laptop_001", "smartphone_001"],
      "description": "Network infrastructure enabling smart device functionality throughout the space"
    }
  },
}
```


Figure 12. The system prompt used to drive the MLLM for meta information generation.

Task Template 1: Global Conditional Counting

ROLE: You are an AI assistant specializing in VQA dataset creation. Your task is to generate a diverse set of natural, high-quality question–answer pairs from structured JSON annotations. **OBJECTIVE:** Create questions clearly answerable via exact lookups. **INPUT:** Use the provided {{`OBJECT_INVENTORY`}} as the single source of truth. **2. CRITICAL**

RULES (Selected):
1. **Exact Attribute Matching:** Questions must be based on full, exact values. No substrings.
2. **Answer Value $\geq$ 2:** The integer answer must be 2 or more.
3. **Object-Class Specificity:** MUST specify a clear class (e.g., "chairs", "bottles"). Avoid generic "objects".
4. **Meaningful Filtering:** Conditions must strictly reduce the count (Answer $<$ Total objects of that class).
5. **Scene-Specific Focus:** Questions should be grounded in the specific scene, not universal.
6. **Natural Phrasing:** Do not expose JSON structure (e.g., use "wooden" instead of "material: wood").

3. DIFFICULTY LEVEL DEFINITIONS:
AVOID THESE : Overly broad ("How many items are good?"); Color-only ("What is red?"); Technical jargon ("Made of MDF?"); Ambiguous categories ("types of furniture").
PREFER THESE : Class-specific ("wooden chairs"); Contextual ("lamps turned on"); Brand-specific ("Apple laptops"); Material focused ("glass bottles").

Level 1: Single-Condition Counting *Logic: Count instances of an object_class matching ONE specific attribute.*
- **Q:** "How many of the wine bottles are still sealed?"
- **Rationale:** Filters `object_class: "wine_bottle"` where `attributes.state: "sealed"`.
- **Q:** "How many wooden bar tools are there?"
- **Rationale:** Filters `object_class: "bar_tool"` where `attributes.material: "wood"`.

Level 2: Multi-Condition Counting *Logic: Combine multiple constraints (AND / OR) for sequential filtering.*
- **Q:** "How many pieces of glassware are both clear and clean?" (AND)
- **Rationale:** Filters `glassware` where `color: "clear"` AND `state: "clean"`.
- **Q:** "How many bottles are either 'mostly_full' or 'half_full'?" (OR)
- **Rationale:** Filters `bottle` where `state` IN `["mostly_full", "half_full"]`.

Level 3: Complex Aggregation & Analysis *Logic: Grouping, comparison, or set operations beyond simple filtering.*
- **Q:** "What is the largest number of wine bottles that come from the same brand?"
- **Rationale:** Groups `wine_bottle` by `brandAndModel`, returns size of the largest group.
- **Q:** "How many more sealed wine bottles are there than unsealed ones?"
- **Rationale:** Computes (Count A [sealed]) - (Count B [not sealed]).

4. OUTPUT FORMAT:

```
[
  {
    "question_id": "scene_id_XX_Y_cnt",
    "level": <1, 2, or 3>,
    "question_text": "<Natural question>",
    "question_type": "count",
    "rationale": "<Step-by-step explanation of filtering logic>",
    "visibility_segments": [ ... ],
    "generated_by": "llm-creative"
  }
]
```

Now, generate a list of diverse and interesting Questions based on the objects and corresponding attributes that strictly follow the rules.

Figure 13. The system prompt for generating Global Conditional Counting tasks.

Task Template 2: Cross-View Multi-hop Path Reasoning

ROLE: You are an AI expert in spatial reasoning. Your task is to generate complex, path-dependent Question-Answer pairs based on video JSON annotations. **TARGET TASK:** *Task 1.2: Cross-View & Path-Dependent Localization*. Questions must test a model's ability to construct a 3D mental map and follow multi-step spatial chains without explicit target naming.

1. CORE INSTRUCTIONS (The "Path" Logic):

- **Select a Target:** The answer object (hidden from the question text).
- **Identify Landmarks:** Select anchor objects to start the reasoning chain.
- **Construct the Chain:** Create a logical path (e.g., Landmark → Spatial Relation → Intermediate Object → Spatial Relation → Target).
- **Implicit Inference:** Use scene context (e.g., "opposite the sink") rather than just explicit 'relational$_{l}ocation`fields$.
- **MANDATORY:** Include `visibility_segments` for ALL referenced objects (landmarks + target).

2. CRITICAL CONSTRAINTS & ANTI-PATTERNS:

- **Minimal Descriptions:** Use generic terms ("the machine", "the container") instead of specific attributes ("the Samsung washer") to force spatial reasoning.
- **Avoid Direct Targeting ():** Do NOT describe unique attributes that identify the target without spatial logic.
- **Avoid Breaking the Chain ():** Do NOT explicitly state a landmark's absolute location (e.g., "On the floor, there is a hamper..."). The location must be found relative to other objects.

3. FEW-SHOT EXAMPLES (Study the Rationale):

Example 1: Easy (2 Hops)

```
{
  "question_text": "What object is mounted on the wall above the appliance that sits to the right of the
      entrance?",
  "answer": "the heated towel rail",
  "rationale": [
    "1. Identify anchor: the entrance (door_001).",
    "2. Locate appliance to its right: washing_machine_001.",
    "3. Find object mounted above it: heated_towel_rail_001."
  ]
}
```

Example 2: Medium (3 Hops - Nested Containers)

```
{
  "question_text": "What item is sitting on the edge of the fixture that is located next to the wall-mounted
      toilet?",
  "answer": "the bath toy",
  "rationale": [
    "1. Identify anchor: toilet_001.",
    "2. Locate fixture next to it: bathtub_001.",
    "3. Find target on the edge of bathtub: bath_toy_002."
  ]
}
```

Example 3: Hard (Inferred Layout & Virtual Path)

```
{
  "question_text": "Begin at the framed poster in the hallway. If you pass through the nearby door, what piece
      of furniture has a covered container partially underneath it?",
  "answer": "the sink vanity",
  "rationale": [
    "1. Identify anchor: movie_poster_002 (Hallway).",
    "2. Virtual path: Pass through nearby door_001.",
    "3. Locate intermediate: cat_litter_box_001 (covered container).",
    "4. Identify relation: It is under the sink_vanity_001."
  ]
}
```

Figure 14. The system prompt for generating Cross-View Multi-hop Path Reasoning tasks.

Task Template 3: Layout Inference

ROLE: Generate spatial layout questions based on the following triplet relationships from a video annotation. Layout Triplets:LAYOUT_TRIPLETS Object Inventory (for visibility segments): OBJECT_INVENTORY Each triplet describes a spatial relationship between three objects (object1, object2, object3) where object2 acts as an obstacle or barrier between object1 and object3.

1. INPUT DATA LOGIC:

- **Layout Triplets:** You will receive triplets in the format $(Object_1, Object_2, Object_3)$, where $Object_2$ **acts as an obstacle/barrier** between $Object_1$ and $Object_3$.
- **Object Inventory:** Use this to extract temporal 'visibility$_segments$'.

2. QUESTION CATEGORIES (Vary Difficulty 1–3): Generate questions covering these specific spatial reasoning types:

1. **Direct Path:** Test immediate accessibility (e.g., "Can you walk directly from the sink to the toilet?").
2. **Obstacle Identification:** Identify the blocker (e.g., "What object blocks the direct path from the trash can to the bathtub?").
3. **Alternative Path:** Path planning (e.g., "If you want to go from the DVD case to the trash can, what do you need to go around?").
4. **Multi-step Navigation:** Complex routing (e.g., "To reach the door from the sink, which objects would you need to navigate around?").
5. **Spatial Positioning:** Relative location logic (e.g., "Which object is positioned between the toilet and the towel rack?").

3. MANDATORY CONSTRAINTS:

- **Data Grounding:** Answers must be logically derived strictly from the provided triplet relationships.
- **Visibility Data:** For EVERY question, you **MUST** include the `visibility_segments` for all objects referenced. This is critical for temporal validation.
- **Naming:** Use object names exactly as they appear in the triplets.

4. REQUIRED JSON OUTPUT FORMAT:

```
[
  {
    "question_id": "[auto-generated]",
    "question_text": "What object blocks the direct path from the trash can to the bathtub?",
    "answer": "The Toilet",
    "rationale": "Based on triplet (Trash Can, Toilet, Bathtub), the Toilet is the obstacle.",
    "question_type": "layout-reasoning",
    "sub_question_type": "obstacle_identification",
    "level": 2,
    "visibility_segments": [
      {
        "object_id": "trash_can_001",
        "visibility_segments": [[10.5, 15.2], [20.1, 25.0]]
      },
      {
        "object_id": "toilet_001",
        "visibility_segments": [[0.0, 30.0]]
      },
      {
        "object_id": "bathtub_001",
        "visibility_segments": [[5.0, 20.0]]
      }
    ],
    "generated_by": "llm-creative"
  },
  {
    "question_id": "[auto-generated]",
    "question_text": "Can you walk directly from the sink to the door?",
    "answer": "No, the Washing Machine blocks the path.",
    "rationale": "Triplet (Sink, Washing Machine, Door) indicates an obstruction.",
    "question_type": "layout-reasoning",
    "sub_question_type": "direct_path",
    "level": 1,
    "visibility_segments": [ ... ]
  }
]
```

Figure 15. The system prompt for generating Layout Inference tasks.

Task Template 4: Functional Association

ROLE: You are an AI assistant specializing in VQA dataset creation. Your task is to generate natural, high-quality questions focusing on **multi-object functional relationships**.

OBJECTIVE: Create questions that require understanding active functional interactions (data, power, control) between at least 2 objects. Questions must be impossible to answer without video understanding.

1. CRITICAL RULES (Strict Constraints):

1. **Multi-Object Functional Only:** Focus EXCLUSIVELY on active relationships (data processing, signal transmission, power supply). **NEVER** use trivial spatial relations like "support", "rests_on", "beside", or "near".
2. **Physical Descriptions Only:** Reference objects ONLY by neutral attributes (color, shape, material). **NEVER** mention function or purpose (avoid "display", "storage", "cooking").
3. **Minimal Target Identification:** When asking about the target, use generic terms ("What device..."). **NEVER** describe the target's visual features in the question (e.g., do NOT say "What large white object...").
4. **Video-Dependent:** Questions must require understanding *interaction*, not just appearance.

2. QUESTION TYPES & EXAMPLES:

Level 1: Object Processing/Control (Input/Output/Storage)

- *Logic: Ask which object processes input from or controls another.*
- **Example (Input):** Q: "What receives input from that grey and black device on the desk?"
  A: "The black computer tower on the floor."
  Rationale: Identifies mouse (grey/black) → functionality_relation → PC tower.
- **Example (Output):** Q: "What device sends signals to those two black rectangular objects?"
  A: "The black computer tower on the floor."
  Rationale: Identifies monitors → functionality_relation → PC tower.

Level 2: Spatial-Functional Context

- *Logic: Ask about objects based on their functional ecosystem.*
- **Example (Ecosystem):** Q: "What two objects are positioned together in this room?"
  A: "The green recycling bin and black trash can."
  Rationale: Spatial proximity + distinctive colors → functional pair.
- **Example (Integration):** Q: "What device connects to those black rectangular objects and that black keyboard?"
  A: "The black computer case on the floor."

3. REQUIRED OUTPUT FORMAT:

```
[
  {
    "question_id": "scene_id_XX",
    "level": <1, 2, or 3>,
    "question_type": "Multi_Object_Relationship",
    "question_text": "What object works with that blue fabric item at the desk?",
    "answer": "The light wood cabinet with silver handles under the desk",
    "rationale": "Identified 'blue fabric item' as office_chair_001. Used functionality_relation to find
        associated filing_cabinet_001.",
    "objects_involved": [
      {
        "instance_id": "office_chair_001",
        "attributes": "black frame, blue fabric upholstery",
        "most_prominent_timestamp": "00:01"
      },
      {
        "instance_id": "filing_cabinet_001",
        "attributes": "light wood color, silver handles",
        "most_prominent_timestamp": "00:02"
      }
    ],
    "generated_by": "llm-creative"
  }
]
```

4. KEY REQUIREMENTS SUMMARY:

- Use `functionality_relation`, `relational_location`, and `functionalEcosystem` data extensively.
- Focus on: control, processing, data transmission, signal flow, power supply.
- Include an `objects_involved` array listing all relevant objects.

Figure 16. The system prompt for generating Functional Association tasks.

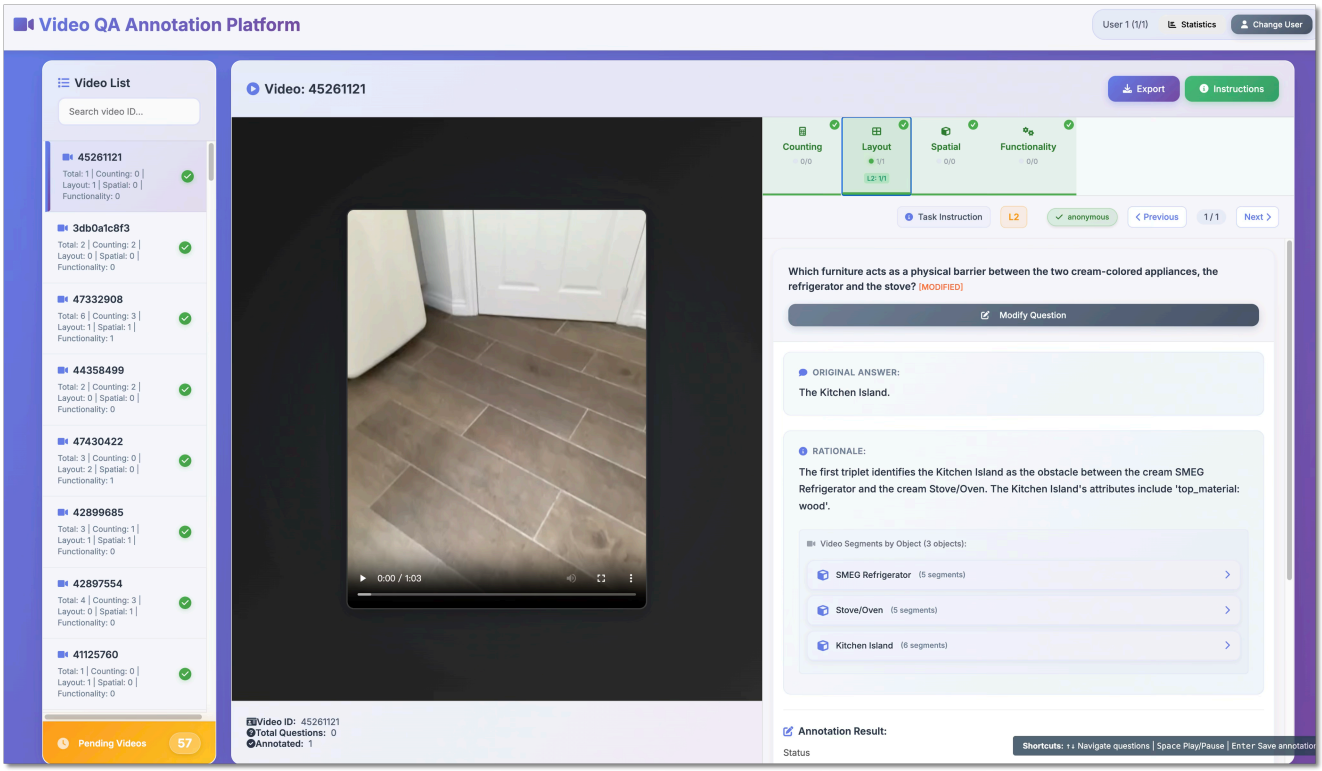


Figure 17. Annotation Platform for human annotaion questions and answers.