SEEING ISN'T KNOWING: DO VLMS KNOW WHEN NOT TO ANSWER SPATIAL QUESTIONS (AND WHY)?

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning benchmarks typically evaluate whether vision-language models can derive the correct answer from a visual observation. Yet in real 3D environments, the observation itself may be unreliable: occlusion can remove task-relevant evidence, while perspective can make visible geometry misleading. Reliable spatial reasoning therefore requires more than answering a question correctly. A model must also assess whether its current observation provides sufficient and trustworthy evidence for that answer. We introduce SPATIALUNCERTAIN, a controlled evaluation framework for studying *viewpoint-dependent observational uncertainty*. By holding the underlying 3D scene fixed while systematically varying the observation, we study two complementary failure modes: missing evidence caused by occlusion and misleading evidence caused by perspective. We further evaluate whether models can recognize when the current view is unreliable and identify a more informative observation. Across eight open- and closed-source vision-language models from the Qwen2.5-VL, InternVL3, GPT, and Gemini families, we find that model behavior does not track the reliability of visual evidence. Models do not reliably become more cautious as evidence disappears, and under perspective conflict, their judgments increasingly follow projected appearance rather than the unchanged physical 3D relation. Internal analysis suggests a corresponding representational asymmetry: projected 2D relations are readily available, whereas the underlying physical 3D relation is barely decodable. Moreover, models that can identify an informative viewpoint when explicitly asked often fail to recognize when such an additional view is needed. These failures are not fully resolved by prompting or fine-tuning, and providing a better viewpoint is substantially more effective than adding depth information to the same misleading observation. Our results identify assessing the reliability of visual observations as a distinct and missing component of current spatial reasoning evaluation.

1 INTRODUCTION

Recent advances in Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Singh et al., 2025; Deepmind, 2025a) have enabled intelligent agents to perceive and interact with their environments, bringing us closer to practical embodied systems (Zhang et al., 2024a). A fundamental capability underlying these systems is spatial reasoning, which has been extensively studied through a growing number of benchmarks (Yang et al., 2025a;b; Pothiraj et al., 2025; Wang et al., 2024a; Liu et al., 2025; Jia et al., 2025; Stogiannidis et al., 2025; Wang et al., 2025). These benchmarks measure models' ability to answer spatial questions, such as object relations, distance, and object size, from visual observations including images or videos. However, they typically assume that the provided visual observation contains sufficient and reliable evidence for answering the question (see Fig. 1(a)).

Visual observations are inherently 2D projections of a 3D world, and the spatial evidence they provide can become incomplete or misleading in different ways. In the occlusion setting (see Fig. 1(b)), a target object may become partially or fully hidden, making its spatial relations increasingly difficult or even impossible to determine from the current view. In the perspective setting (see Fig. 1(c)), objects can remain fully visible while viewpoint changes distort their apparent geometry, potentially creating conflicts between projected 2D appearance and the underlying physical 3D relation. Such unreliable observations are particularly challenging for embodied agents (Zhang et al., 2024a; Duan et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

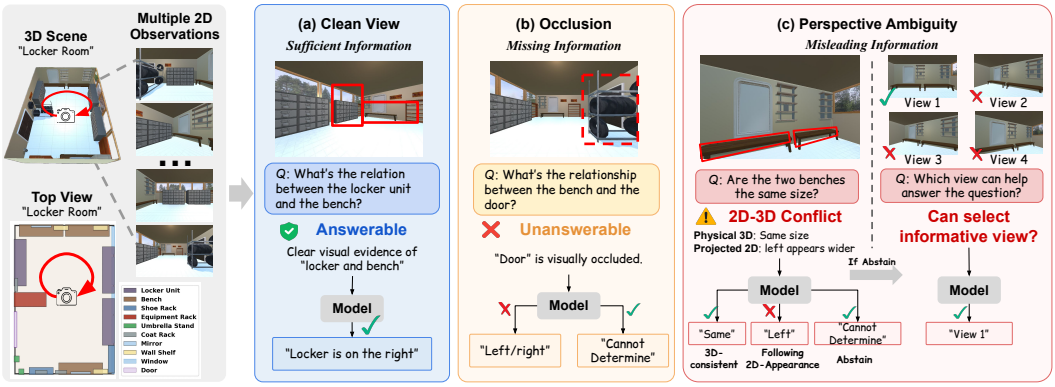


Figure 1: Visual observations are inherently 2D projections of a 3D world and can provide sufficient, missing, or misleading evidence for spatial reasoning. (a) Under clean views, spatial questions are answerable from direct visual evidence. (b) Under occlusion, task-relevant evidence may become invisible and require abstention with *Cannot determine*. (c) Under perspective ambiguity, projected 2D appearance can conflict with the underlying physical 3D relation, requiring models to recognize conflicting evidence and identify a more informative view for reliable reasoning.

2022; Zhang & Kordjamshidi, 2023; Yu et al., 2026a), where acting on missing or misleading visual evidence can lead to incorrect action decisions or unsafe behaviors. Ideally, a model should recognize when the current evidence is insufficient or unreliable, abstain when necessary, and seek additional observations when the current view does not support a reliable judgment. A growing body of work studies uncertainty and abstention across language and multimodal models, focusing on whether a model can recognize when it does not know the answer (Manakul et al., 2023; Stengel-Eskin et al., 2024; Wen et al., 2025; Madhusudhan et al., 2026). A complementary challenge in visual spatial reasoning is whether the model can assess the reliability of the observation itself, namely, whether the available visual evidence is sufficient and trustworthy for making a spatial judgment.

To address this gap, we introduce SPATIALUNCERTAIN, a controlled evaluation framework for studying *viewpoint-dependent observational uncertainty* in spatial reasoning, where the reliability of visual evidence changes with the observation viewpoint while the underlying 3D scene remains fixed. We construct paired observations with systematically varying degrees of evidence reliability, rather than a single binary distinction between answerable and unanswerable cases. For occlusion, we progressively vary and quantify target visibility, producing clean, partially occluded, and fully occluded conditions that control how much task-relevant evidence remains observable. For perspective, we vary the camera viewpoint to induce different degrees of disagreement between projected 2D appearance and the corresponding physical 3D relation. We evaluate whether model behavior is appropriate to the evidence available in the current view. Under occlusion, this means answering when sufficient evidence remains and abstaining when it does not; under perspective conflict, we characterize whether predictions remain consistent with the physical 3D relation, follow the projected 2D appearance, or abstain. For perspective cases, we further evaluate whether models can identify a more informative viewpoint through two complementary tasks: ViewSel, which measures viewpoint selection in isolation, and AbstainViewSel, which jointly tests whether a model can first recognize an unreliable observation and then select an alternative view that better supports the spatial judgment.

Using this controlled setup, we evaluate eight vision-language models spanning open-source (Qwen2.5-VL-7B, Qwen2.5-VL-32B, InternVL3-8B) and closed-source (GPT-4o, GPT-5-mini, GPT-5.4, Gemini-2.5-Flash, Gemini-3.0-Flash) families. Our results show that model behavior often fails to track the reliability of the available visual evidence (Sec. 4.2). Under occlusion, abstention increases only weakly and inconsistently as target visibility decreases. Under perspective conflict, models increasingly follow the projected 2D appearance as disagreement with the underlying physical 3D relation becomes stronger. Although stronger models can often identify an informative viewpoint when explicitly asked, their performance drops substantially when they must first determine whether the current observation is reliable enough to answer or whether another view is needed. We further diagnose these failures (Sec. 4.3). The physical 3D relation is barely decodable from model representations, while the projected 2D relation is readily available. Model attention is also biased

toward the object that appears larger in projection. The same behavior persists under open-ended evaluation, and a real-image consistency check shows the same pattern: viewpoint changes can induce contradictory physical-size judgments despite unchanged object sizes. Finally, we examine what helps models handle unreliable observations (Sec. 4.4). Structured prompting improves performance on unanswerable cases but introduces a trade-off with answerable accuracy. Fine-tuning on diverse ambiguity conditions substantially reduces 2D-following, though largely by shifting predictions toward abstention rather than recovering the physical 3D relation. At inference time, providing an alternative viewpoint is more effective than augmenting the same misleading observation with depth information. Together, these results suggest that reliable spatial reasoning requires not only reasoning from visual observations, but also assessing when those observations should be trusted and when better evidence should be sought.

2 Related Work

Spatial reasoning in MLLMs. Spatial reasoning has emerged as a fundamental capability for multimodal large language models (MLLMs) (Chen et al., 2024; Cheng et al., 2024; Zhang et al., 2024b), and a growing body of work has proposed benchmarks to evaluate it (Yang et al., 2026; 2025a; Yu et al., 2026b; Daxberger et al., 2025; Wang et al., 2024b; Xu et al., 2025; Rajabi & Kosecka, 2024; Yang et al., 2025c; Ma et al., 2022). Early efforts focus on evaluating basic spatial relations such as relative relations, depth ordering, and size comparison using image/video-based question answering datasets. More recent benchmarks aim to provide broader and more systematic evaluations, including multi-view (Zhang et al., 2026a; Wang et al., 2025; Zhang et al., 2026b) and multi-task settings (Yin et al., 2023; Jia et al., 2025), which cover diverse spatial reasoning skills. Several works further emphasize the importance of controlled evaluation (Pothiraj et al., 2025; Liu et al., 2023; Johnson et al., 2017). For example, What'sUp (Kamath et al., 2023) constructs minimally varying image pairs to isolate spatial relations. Despite these advances, existing benchmarks primarily evaluate whether models produce correct answers, but do not explicitly assess whether a question is answerable given the observation. In contrast, we evaluate spatial reasoning under varying observation conditions (e.g., occlusion and perspective ambiguity), focusing on whether models can recognize when the available evidence is reliable for spatial questions.

Observational uncertainty and abstention. Uncertainty estimation and abstention are important for building reliable models. Classical work on calibration and selective prediction shows that neural networks can be overconfident, and that models should sometimes abstain when their predictions are uncertain (Guo et al., 2017; Hendrycks & Gimpel, 2016; Geifman & El-Yaniv, 2017; Whitehead et al., 2022). This problem has become especially important for large language models, which often generate fluent but unsupported answers (Kirichenko et al., 2026). Recent work therefore studies truthfulness, self-knowledge, confidence elicitation, hallucination detection, and calibrated expressions of uncertainty (Lin et al., 2022; Kadavath et al., 2022; Yin et al., 2023; Tian et al., 2023; Xiong et al., 2024; Manakul et al., 2023; Stengel-Eskin et al., 2024; Wen et al., 2025). While uncertainty and abstention have been widely explored in language models, they remain less studied in vision-language models. Related efforts examine object hallucination, unanswerable visual questions, and selective VQA, encouraging models to abstain rather than answer incorrectly (Rohrbach et al., 2018; Li et al., 2023; Sun et al., 2024; Guan et al., 2024; Gurari et al., 2018; Guo et al., 2024; He et al., 2024; Eisenschlos et al., 2024). However, these works primarily focus on factual uncertainty, object existence, or generic unanswerability, and typically assume that visual observations provide reliable evidence for reasoning. In contrast, we study observation-dependent uncertainty in spatial reasoning, where answerability is determined by the viewpoint.

3 SpatialUncertain: Controlled Evaluation Framework

We construct a controlled evaluation framework for studying viewpoint-dependent observational uncertainty in spatial reasoning, and the pipeline is shown in Fig. 2. Our key setting is that the underlying 3D indoor scene (Sec. 3.1) remains fixed, while the reliability of the available visual evidence changes with the observation viewpoint. We instantiate two complementary forms of uncertainty: occlusion (Sec. 3.2), where task-relevant evidence becomes missing from the current view, and perspective ambiguity (Sec. 3.3), where visual evidence remains available but its 2D projection may provide misleading cues about the underlying 3D geometry. All configurations undergo human validation to ensure quality (Sec. 3.5).

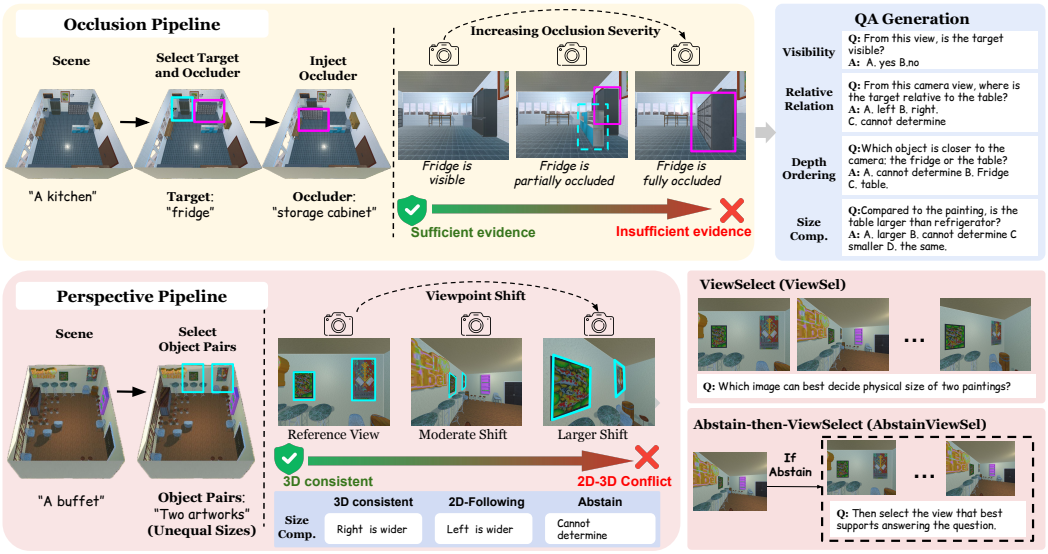


Figure 2: Overview of our evaluation framework, SPATIALUNCERTAIN. **(Top) Occlusion:** We select a target-occluder pair and progressively decrease target visibility from a clean reference view to full occlusion. We evaluate four spatial question types: visibility, relative position, depth ordering, and size comparison. **(Bottom) Perspective:** We select same-category object pairs and shift the camera viewpoint. We focus on horizontal or vertical size comparison and categorize model predictions as 3D-consistent, 2D-following, abstaining, or other. We further introduce **ViewSel**, which directly evaluates informative-view selection, and **AbstainViewSel**, jointly evaluates whether a model can either reason correctly from the current view or, after abstaining, select a more informative one.

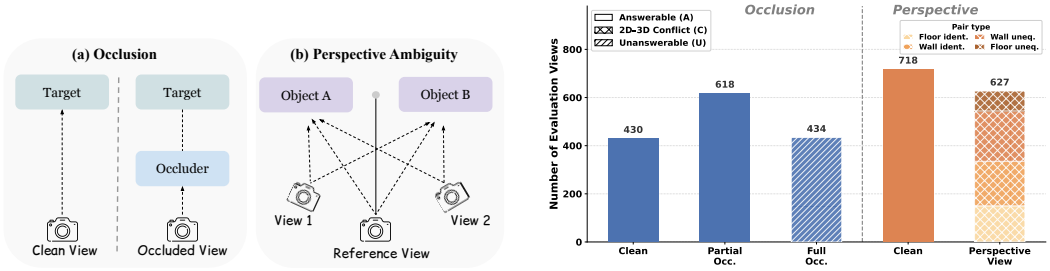
3.1 3D SCENE COLLECTION

We generate 3D indoor scenes using Holodeck (Yang et al., 2024), an LLM-based automated layout generation system. Given a natural language prompt (e.g., *"a bedroom"* or *"a kitchen"*), Holodeck uses a large language model (GPT-4o (OpenAI, 2024)) to plan object selection and placement, producing diverse and realistic room configurations. For each scene, Holodeck provides full 3D asset placement with object positions, orientations, and bounding box information, which we use to automatically derive ground truth annotations without the need for manual labeling. All scenes are rendered using AI2-THOR (Kolve et al., 2017), which supports controllable camera placement and produces photo-realistic RGB images. This controlled rendering environment is central to our framework: by fixing the 3D scene and varying only the camera viewpoint or object configuration, we can isolate how changes in observation reliability affect model reasoning.

3.2 OCCLUSION CONFIGURATIONS

Target-occluder selection. For each clean scene, we select target objects satisfying two criteria: (1) *visibility*: the object must be clearly visible from the camera viewpoint, measured by its projected bounding box area; and (2) *uniqueness*: the object must be the only instance of its category in the scene, ensuring unambiguous reference in generated questions. We retain the top-k (k=3) targets per scene ranked by visibility score (more details are in the Appendix A.1). In our setting, this procedure results in a diverse collection of target–occluder pairs, with targets spanning 225 unique object categories (e.g., bookshelf, coffee table, ottoman) and occluders spanning 286 categories (e.g., storage cabinet, bookshelf, armchair), covering a wide range of object types and occlusion scenarios.

Occlusion scene camera placement. Given a selected target-occluder pair, we place the occluder along the line of sight between the camera and the target, ensuring it is closer to the camera than the target (shown in Fig. 3(a) left). The placement is subject to several geometric constraints to ensure physical plausibility: (1) the occluder must not penetrate other objects or room boundaries, (2) a minimum depth separation is maintained between the occluder and target to avoid clipping, and (3) the occluder must remain within the same room area as the target.



(a) Camera placement under different conditions.

(b) Distribution of types.

Figure 3: Camera placement under different observation conditions (a) and question distributions for occlusion and perspective ambiguity (b), showing answerable/unanswerable cases under occlusion and 2D-3D conflict cases under perspective ambiguity.

Quantifying target visibility. For each target-occluder pair, we measure the fraction of the target that remains visible under occlusion using instance masks rendered from the same camera pose. Let $M_{\text{occ}}(T)$ denote the target mask in the original occluded scene, and $M_{\text{ref}}(T)$ the target mask from a reference rendering in which only the designated occluder is hidden. We define the target visibility ratio as $r_{\text{vis}} = \frac{|M_{\text{occ}}(T)|}{|M_{\text{ref}}(T)|}$, where $|\cdot|$ denotes the number of target pixels in the instance mask. A smaller visibility ratio indicates stronger occlusion caused by the designated occluder, and we use this measure to group configurations by occlusion severity. We categorize an occluded view as *partial* when the measured target visibility satisfies $r_{\text{vis}} \geq 0.05$, and as *full* when $r_{\text{vis}} < 0.05$.

3.3 Perspective Ambiguity Configuration

Object pair selection. To study how viewpoint changes affect perceived object geometry, we select pairs of same-category objects whose physical dimensions can be compared unambiguously. We consider two spatial configurations. *Floor pairs* consist of two floor-standing objects of the same category (e.g., two chairs) that are spatially proximate and have similar orientations. *Wall pairs* consist of two wall-mounted objects (e.g., two paintings) placed on the same or adjacent walls. Within each configuration, we include both *identical* and *unequal* pairs. For identical pairs, the two objects are instantiated from the same 3D asset and therefore have the same physical size. For unequal pairs, we deliberately select two different assets from the same semantic category whose physical dimensions differ along the queried axis. Examples are shown in Fig. 4.

Perspective camera placement. As shown in Fig. 3(a) right, for each object pair, we generate a *reference view* and a *perspective view*. The reference view places the camera at an approximately equidistant position from the two objects. The perspective view shifts the camera toward one object while keeping both objects within the field of view. The resulting effect depends on the pair configuration. For *identical-size* pairs, these effects can make physically identical objects appear different in size. For *unequal-size* pairs, they can reverse the true physical size ordering, such that the physically smaller object appears larger in the image. Importantly, such perspective-induced disagreement does not necessarily make the underlying physical relation fundamentally unrecoverable from a single image. For example, objects lying on the same plane may provide projective cues that support geometric recovery (Criminisi et al., 2000). We therefore distinguish the physical 3D relation from the relation suggested by the 2D projection, and analyze whether model predictions are consistent with the physical relation, follow the projected appearance, or abstain.

Quantifying perspective ambiguity. For each object pair, we obtain metric object dimensions from the underlying 3D asset meshes and analytically project the compared dimension into the image plane using the renderer's pinhole camera model. Let $\ell_{\text{2D}}(O)$ denote the projected pixel length of the queried dimension (horizontal width or vertical height) for object O. For unequal-size pairs, we define the projected-size inversion ratio as $r_{\text{proj}} = \frac{\ell_{\text{2D}}(O_{\text{small}})}{\ell_{\text{2D}}(O_{\text{large}})}$, where O_{small} and O_{large} are determined from the physical 3D size of the corresponding object dimension. A ratio above 1 indicates that the physically smaller object appears larger in the image, yielding a reversal between the projected 2D relation and the physical 3D relation. Larger values indicate stronger 2D-3D conflict.

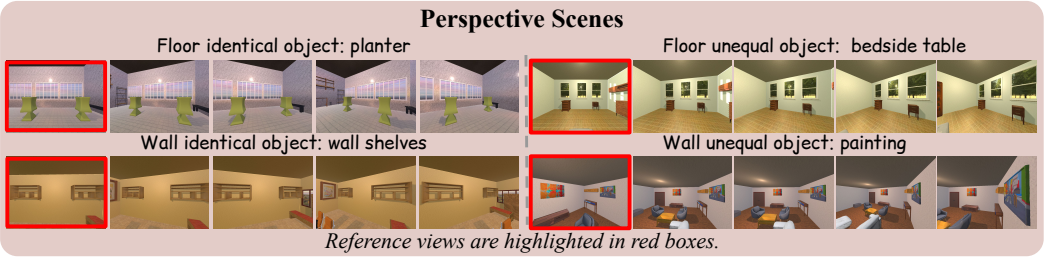


Figure 4: Examples of perspective scenes where camera shifts create conflicts between projected 2D appearance and the underlying 3D relation.

3.4 Task Design

Spatial question design. For occlusion, we consider four question types: *Visibility*, *Relative position*, *Depth ordering*, and *Size comparison*. Answerability is defined separately for each type. For partially occluded observations, partial visibility alone does not guarantee that a question remains answerable. We therefore retain only questions for which human reviewers can reliably recover the geometry-derived answer from the visible evidence. For fully occluded observations, visibility questions remain answerable, while questions requiring access to the hidden target are treated as unanswerable, with *Cannot determine* as the appropriate response. For the perspective setting, we focus specifically on physical size comparison. For each object pair, we query either the horizontal or vertical dimension. We define a *2D-3D conflict* when the size relation suggested by the projected 2D appearance disagrees with this physical 3D relation. In such cases, we characterize model predictions as *3D-consistent*, *2D Appearance-following*, *Abstain*. For 2D-3D conflict cases, we treat 3D-consistent predictions as correct, 2D-following predictions as errors induced by misleading appearance, and abstention as a conservative response that is reported separately rather than as correctness.

View selection under perspective ambiguity. A reliable model should also recognize when additional visual evidence would support a more reliable judgment. Therefore, we introduce two complementary tasks for viewpoint assessment under perspective ambiguity. **ViewSelect (ViewSel)**: the model is presented with five candidate views (one informative reference view and four ambiguous alternatives) and asked to identify the view that best supports answering a spatial reasoning question about physical size. This metric evaluates pure viewpoint selection ability in isolation, independent of abstention behavior. **Abstain-then-ViewSelect (AbstainViewSel)**: We introduce a two-stage evaluation that tests whether a model can either reason correctly from the current perspective-conflict view or recognize when an additional view is needed. In Stage 1, a 3D-consistent answer is counted as correct directly. If the model abstains, Stage 2 presents five candidate views and asks it to select the informative reference view. The prediction is counted as correct if either the Stage-1 answer is 3D-consistent or the model abstains and successfully selects the informative view.

3.5 Dataset Statistics and Human Validation

We collect 240 unique scenes spanning 43 room types and construct 1,222 occlusion units, including 618 partial- and 434 full-occlusion cases, yielding 5928 QA pairs across four spatial question types. For each unit, we additionally quantify occlusion severity using the target visible fraction computed from paired instance-mask renderings with and without the designated occluder. For perspective ambiguity, we focus on physical size comparison across four configurations: 150 floor-identical, 186 wall-identical, 83 floor-unequal, and 209 wall-unequal pairs. Each configuration includes a perspective-shifted view designed to create disagreement between projected 2D appearance and the underlying 3D size relation. In total, the perspective set contains 627 2D-3D conflict questions and 718 clean-view controls. More statistics and examples are shown in Fig. 3b(b).

All automatically generated configurations undergo human verification for quality control. Evaluation labels are derived from the underlying 3D geometry and camera projection rather than assigned by annotators. For occlusion, reviewers verify rendering quality and, for partially occluded views, additionally filter questions whose geometry-derived answers cannot be reliably recovered from the current observation. Occlusion severity is defined by the measured target visible fraction r_{vis}, with full occlusion corresponding to less than 5% visibility. For perspective ambiguity, reviewers verify

Table 1: Performance under occlusion and perspective ambiguity. **Ans.** and **Unans.** denote accuracy on answerable and unanswerable cases. For perspective conflict, **3D-Cons.**, **Abstain**, and **2D-Foll.** denote 3D-consistent, abstaining, and 2D-following predictions. **ViewS.** and **AbsViewS.** denote ViewSel and AbstainViewSel accuracy.

Model	Occlusion		Perspective Ambiguity					
	Ans.↑	Unans.↑	Ans.↑	2D–3D Conflict			ViewS↑	AbsViewS↑
				3D-Cons.↑	Abstain	2D-Foll.↓		
Random	32.3	23.3	25.0	25.0	25.0	25.0	20.0	30.0
Open-source								
Qwen2.5-VL-7B (Bai et al., 2025)	51.6	42.0	19.9	3.8	27.1	65.6	24.6	5.6
Qwen2.5-VL-32B (Bai et al., 2025)	52.0	45.5	65.7	22.8	16.4	55.2	20.7	3.8
InternVL3-8B (Zhu et al., 2025)	62.9	8.3	68.8	29.0	0.8	65.7	18.5	0.0
Closed-source								
GPT-4o (OpenAI, 2024)	43.7	**63.6**	34.4	15.0	**38.3**	**35.6**	39.3	26.0
GPT-5-mini (Singh et al., 2025)	**66.2**	8.4	**70.8**	26.6	8.9	63.5	53.7	**34.1**
GPT-5.4 (OpenAI, 2026)	59.0	21.1	65.7	15.6	16.7	61.1	**70.9**	28.5
Gemini-2.5-Flash (Deepmind, 2025a)	56.8	49.4	67.4	**32.5**	4.8	57.7	18.5	33.8
Gemini-3.0-Flash (Deepmind, 2025b)	63.3	51.4	65.9	22.6	6.2	68.4	50.3	24.8

object identifiability and rendering quality, while 2D-3D conflict labels are computed automatically from the projected-size ratio r_{proj}. Full annotation protocols and filtering statistics are provided in Sec. A.1.1 in the Appendix.

4 Experimental Results

4.1 Evaluation Models and Protocol

We evaluate eight vision-language models spanning both open- and closed-source families. Open-source models include Qwen2.5-VL-7B (Bai et al., 2025), Qwen2.5-VL-32B (Bai et al., 2025), and InternVL3-8B (Zhu et al., 2025). Closed-source models include GPT-4o (OpenAI, 2024), GPT-5-mini (Singh et al., 2025), GPT-5.4 (OpenAI, 2026), Gemini-2.5-Flash (Deepmind, 2025a), and Gemini-3.0-Flash (Deepmind, 2025b). All models are evaluated zero-shot using standardized multiple-choice prompts (Appendix A.2.1). For occlusion, models are evaluated on clean, partially occluded, and fully occluded observations. We report **Ans.**, accuracy on answerable questions, and **Unans.**, accuracy on cases where the appropriate response is *Cannot determine*. For perspective ambiguity, we separately characterize predictions on 2D-3D conflict cases as **3D-Cons.**, **2D-Foll.**, **Abstain**, depending on whether the prediction follows the physical 3D relation, the projected 2D appearance, or abstains. For viewpoint selection, we report **ViewSel** and **AbstainViewSel** accuracy. Additional metric details are provided in Appendix A.2.2. To rule out answer-format artifacts, we shuffle answer options, include a text-only baseline, and additionally verify on 600 open-ended questions that performance closely matches the multiple-choice setting (Appendix A.3.1).

4.2 Model Behavior under Observational Uncertainty

Failure to handle unreliable observations. As shown in Table 1, across nearly all models, performance on answerable questions exceeds random chance, indicating meaningful spatial reasoning when visual evidence is sufficient. However, model behavior changes substantially when observations become unreliable, revealing three recurring patterns. **(1) Answer-abstention trade-off.** Models that abstain more readily often sacrifice accuracy on answerable questions. For example, Gemini-2.5-Flash achieves relatively high Occ-Unans. (49.4) but lower Occ-Ans. (56.8), while GPT-5-mini achieves the highest perspective Ans. (70.8) but abstains on only 8.9 of 2D–3D conflict cases. GPT-4o illustrates the opposite pattern, achieving the highest Occ-Unans. (63.6) but the lowest Occ-Ans. among evaluated models (43.7). These results suggest that current models often improve abstention by becoming more conservative overall, rather than selectively recognizing unreliable evidence. **(2) Inconsistent uncertainty behavior across observation types.** The same model can respond very differently to different forms of unreliable evidence. For example, Gemini-2.5-Flash achieves Occ-Unans. 49.4 under occlusion but abstains on only 4.8 of perspective-conflict cases, while GPT-4o achieves Occ-Unans. 63.6 and an Abstain rate of 38.3 under perspective conflict. Such

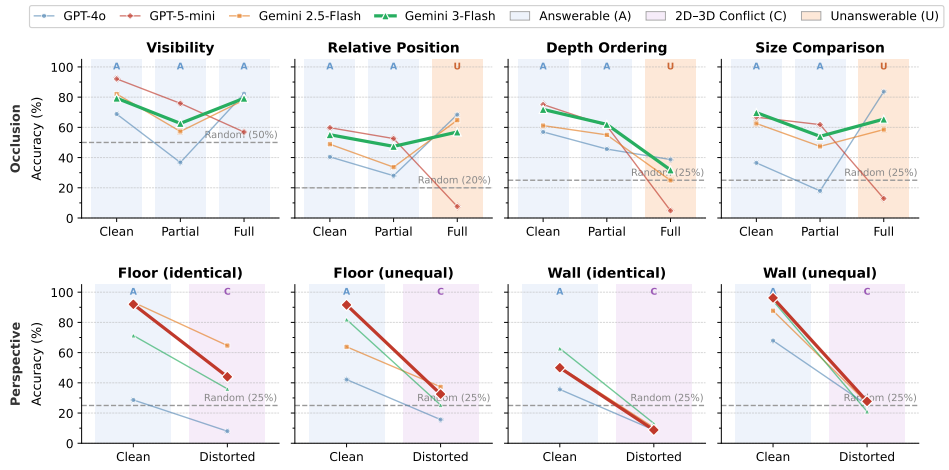


Figure 5: Model accuracy across question types under occlusion (top) and perspective ambiguity (bottom). Shaded regions indicate answerable (A), unanswerable (U), and 2D-3D conflict (C) conditions, with dashed lines showing random baselines. Bold lines highlight Gemini-3.0-Flash for occlusion and GPT-5-mini for perspective ambiguity.

variation suggests that abstention behavior does not transfer consistently across different sources of observational uncertainty. **(3) Sensitivity to observation reliability.** As shown in Fig. 5, performance declines as occlusion increases. Partial occlusion already hurts performance, while full occlusion reveals models' limited ability to recognize when the queried relation is no longer visually determined. Under perspective ambiguity, models answer clean reference views reliably, but consistently favor the misleading projection on 2D-3D conflict views. Together, these results show that models do not reliably adjust their behavior when visual evidence becomes incomplete or misleading.

Failure to identify informative viewpoints. Results on view selection tasks reveal that current models struggle not only to recognize unreliable observations, but also to identify which additional viewpoints would provide reliable evidence. On *ViewSel*, stronger models such as GPT-5.4 (70.9) and GPT-5-mini (53.7) achieve substantially above-random performance, indicating that models can often identify informative views when explicitly prompted. However, performance remains substantially lower on *AbstainViewSel*, which requires models to either produce a 3D-consistent answer from the current view or abstain and select a more informative viewpoint. This gap suggests that models can often identify an informative view when explicitly prompted, but still struggle to determine whether the current observation is reliable enough to answer directly or whether additional evidence is needed.

4.3 Diagnosing Observational Failures

Occlusion: abstention only weakly tracks missing evidence. Models perform well on visibility questions, suggesting that they can recognize when the target becomes occluded. Their abstention rates also tend to increase as the measured visible fraction decreases, but the response is weak and highly model-dependent (Fig. 6a). Even when less than 5% of the target remains visible, some models still rarely abstain. These results suggest that models can perceive degraded visual evidence, but do not consistently translate its severity into an appropriate decision to abstain.

Perspective: projected 2D cues provide a stronger internal signal. To isolate the source of perspective failures, we exploit our controlled construction to vary projected appearance while keeping the underlying physical 3D relation fixed. As the projected-size conflict becomes stronger, models increasingly follow the projected 2D relation (Fig. 6b), showing that their predictions systematically shift with viewpoint-dependent appearance rather than the unchanged physical relation. We further examine this behavior internally using Qwen2.5-VL-7B. While the projected 2D relation is readily decodable across model layers, the underlying physical 3D relation remains only weakly decodable and close to chance (Fig. 7b), suggesting that the model lacks a robust internal signal for recovering physical size under misleading projections. Answer-token attention is also biased toward the object that appears larger in projection (Fig. 7a). Interestingly, this attention bias is reduced after LoRA fine-tuning on our mixed occlusion and perspective training data, consistent with the substantially

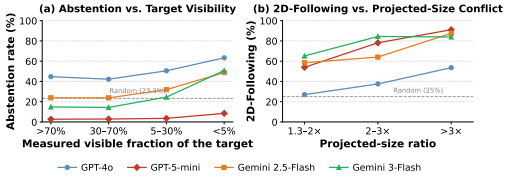


Figure 6: Effect of observation severity on model behavior. **(a)** Abstention rate under occlusion. **(b)** 2D-following rate with perspective shifts.

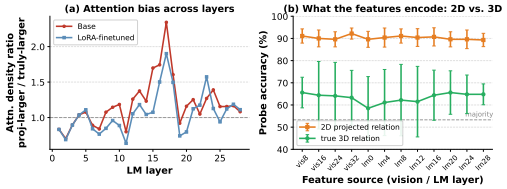


Figure 7: Internal analysis of projected 2D vs. physical 3D cues on Qwen2.5-VL-7B.

lower 2D-following behavior after fine-tuning. Together, these results suggest that when the physical 3D relation is weakly represented, models tend to rely on the more accessible projected appearance. The same behavioral pattern persists under open-ended evaluation (Appendix A.3.1 and Fig. 10) and transfers to real images. Among 35 real-image object pairs that do not appear to belong to a matched set, GPT-5-mini gives contradictory size judgments across viewpoints on 30 pairs (86%), and Gemini-2.5-Flash on 23 pairs (66%), each time favoring the object that projects larger. Because physical size is viewpoint-invariant, these contradictions provide a label-free certificate that at least one prediction is incorrect, showing that the effect is not specific to synthetic rendering.

4.4 Toward Improvement: What Helps and What Does Not

Prompting and fine-tuning. We first test whether simple interventions can mitigate these failure modes. As shown in Sec. A.2.1 and Table 5 in the Appendix, a structured prompt that explicitly asks models to assess target visibility and viewpoint reliability increases abstention under occlusion, but at the cost of answerable accuracy. For GPT-5-mini, Occ-Unans. increases from 8.4 to 30.4, while Occ-Ans. decreases from 66.2 to 54.7. In contrast, LoRA fine-tuning, shown in Table 4, produces broader gains. For Qwen2.5-VL-7B, it substantially improves both answerable and unanswerable performance under occlusion and nearly eliminates 2D-following under perspective conflict. However, most perspective predictions shift toward abstention (86.2) rather than 3D-consistent recovery (13.8). These results suggest that prompting primarily encourages more conservative behavior, while supervision can more effectively teach models to recognize when visual cues are unreliable.

Changing the viewpoint is more effective than adding depth information. To test what evidence can mitigate perspective-induced errors, we compare three inference-time interventions: ground-truth depth provided as text, a pixel-aligned depth map, and an alternative reference view. As shown in Fig. 8, depth information does not reliably resolve the 2D–3D conflict. Depth as text largely shifts predictions toward uncertainty, while the depth map increases both 3D-consistent predictions (25 to 33) and 2D-following predictions (57 to 65), indicating that it does not consistently steer the model toward the underlying physical relation. In contrast, an alternative viewpoint increases 3D-consistent predictions from 25 to 51 while reducing 2D-following. This suggests that changing the observation can be more effective than augmenting a misleading view with depth information. These results motivate an active-perception perspective: models should recognize when the current observation is unreliable and seek a more informative viewpoint rather than relying on additional cues from the same view.

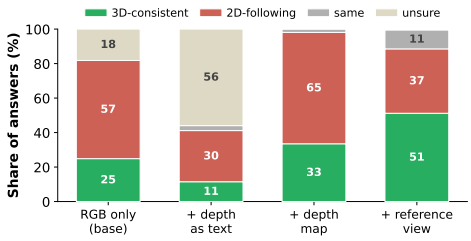


Figure 8: Effect of additional visual evidence on perspective reasoning.

5 Conclusion

We present SpatialUncertain, a controlled framework for evaluating whether VLMs can assess the reliability of visual observations under occlusion and perspective ambiguity. Across eight models, behavior often fails to track observation reliability, with models remaining insufficiently cautious under missing evidence and overly influenced by projected appearance. Fine-tuning reduces this reliance largely by increasing abstention rather than recovering the underlying 3D relation, motivating future work on spatial reasoning that can recognize unreliable observations and seek better evidence.

Artificial intelligence use statement

We used AI coding assistants to support implementation and evaluation, and generative AI tools to revise text, format LaTeX, and check spelling and grammar. As described in Section 3.1, Holodeck uses GPT-4o to generate scene layouts, while evaluation labels are derived from 3D geometry and camera projections, with human verification for quality control. Generative AI was not used for dataset cleaning or reformatting, research design or feedback, conceptual development, hypothesis formulation, or result analysis and interpretation. Theoretical modeling, mathematical claims, and proofs are not applicable to this work. We reviewed AI-assisted outputs through code inspection and manuscript proofreading and take responsibility for the final content.

Ethics statement

This work evaluates models on synthetic indoor scenes, supplemented by real indoor images from public listings and photographs captured by the authors. Human involvement is limited to dataset quality verification rather than experiments on human participants. External simulation assets and models are used under their respective licenses and API terms, as described in Appendix A.5. All authors have read and adhere to the ICLR Code of Ethics.

Reproducibility statement

Benchmark construction and human verification procedures are detailed in Appendix A.1. Evaluation prompts and metrics are provided in Appendix A.2, and fine-tuning hyperparameters in Appendix A.2.3. Evidence interventions and the supplementary real-image evaluation are described in Appendices A.4 and A.4.1, respectively. We include benchmark data in the supplementary to support reproducibility.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. URL https://api.semanticscholar.org/CorpusID:276449796.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

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 model. *ArXiv*, abs/2406.01584, 2024. URL https://api.semanticscholar.org/CorpusID:270215984.

Antonio Criminisi, Ian D. Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision*, 40:123–148, 2000. URL https://api.semanticscholar.org/CorpusID:2499410.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7395–7408, 2025.

Deepmind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. https://arxiv.org/abs/2507.06261, 2025a.

Deepmind. Gemini 3 flash: Frontier intelligence built for speed. https://blog.google/products/gemini/gemini-3-flash/, 2025b.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

Julian Eisenschlos, Hernán Maina, Guido Ivetta, and Luciana Benotti. Selectively answering visual questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4219–4229, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10284–10296, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Xingwei He, Qianru Zhang, A Jin, Yuan Yuan, Siu-Ming Yiu, et al. Tubench: Benchmarking large vision-language models on trustworthiness with unanswerable questions. *arXiv preprint arXiv:2410.04107*, 2024.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *Advances in Neural Information Processing Systems*, 38, 2026.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

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

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 620–632, 2025.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

Nishanth Madhusudhan, Vikas Yadav, and Alexandre Lacoste. Knowing when not to answer: Evaluating abstention in multimodal reasoning systems. *arXiv preprint arXiv:2604.14799*, 2026.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

OpenAI. Hello GPT-4o, 2024. URL `https://openai.com/index/hello-gpt-4o`.

OpenAI. Openai: Gpt-5.4 model, 2026. URL `https://developers.openai.com/api/docs/models/gpt-5.4`.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8001–8010, 2025.

Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *ArXiv*, abs/2406.13246, 2024. URL `https://api.semanticscholar.org/CorpusID:270619607`.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. Lacie: Listener-aware finetuning for calibration in large language models. *Advances in Neural Information Processing Systems*, 37:43080–43106, 2024.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024a.

Qineng Wang, Baiqiao Yin, 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 *The Fourteenth International Conference on Learning Representations*, 2025.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19757–19767, 2024b.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2022.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

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*, pp. 10632–10643, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2025b.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 2026.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025c.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16227–16237, 2024.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. *ArXiv*, abs/2602.08236, 2026a. URL https://api.semanticscholar.org/CorpusID:285452504.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. *arXiv preprint arXiv:2602.08236*, 2026b.

Hantao Zhang, Jinru Sui, Ed Li, Dirk Bergemann, and Zhuoran Yang. Multiview-bench: A diagnostic benchmark for world-centric multi-view integration in vlms. *arXiv preprint arXiv:2607.08970*, 2026a.

Yue Zhang and Parisa Kordjamshidi. Vln-trans: Translator for the vision and language navigation agent. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:257038436.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024a.

Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language models. *arXiv preprint arXiv:2410.03878*, 2024b.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 70072–70141, 2026b.

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.

A Appendix

A.1 SpatialUncertain Construction Details

Target occluder in occlusion configuration We score each object based on its visibility from the camera viewpoint, combining two factors: angular centrality (how close the object is to the camera's optical axis) and apparent size (the object's projected size relative to its depth). Objects that are partially occluded by other scene objects receive a penalty. We retain the top-k (k=3) objects per scene as target candidates, further requiring scene-level uniqueness.

Object pair selection in perspective configuration. To induce perspective ambiguity, we identify pairs of objects that are physically comparable but visually sensitive to viewpoint changes. Specifically, we select pairs of same-category objects with similar physical size, ensuring that they are expected to appear comparable under neutral viewpoints but can exhibit large appearance differences under perspective distortion. We consider two types of object pairs. *Floor pairs* consist of two floor-standing objects of the same type (e.g., chairs), whose centers are below 1.2m and are separated by at least 2.5m to allow for significant depth variation. *Wall pairs* consist of two wall-mounted objects (e.g., paintings) placed on the same or adjacent walls, separated by at least 1.2m, and matched in aspect ratio within a 20% tolerance to ensure similar physical proportions. For each scene, we sample up to 3 floor pairs and 2 wall pairs, promoting diversity while maintaining controlled geometric conditions.

A.1.1 Human Verification

All automatically generated configurations undergo human verification by seven reviewers for quality control. Annotators independently review each configuration, and approximately one third of generated candidates are removed during this process. Evaluation labels themselves are derived from the underlying 3D geometry and camera projection rather than assigned by annotators.

For occlusion, reviewers verify that the designated occluder is correctly rendered and that the resulting scene is visually plausible and free of obvious artifacts. Occlusion severity is determined by the geometrically measured target visible fraction r_{vis}, computed from paired instance-mask renderings with and without the designated occluder. We define partial occlusion as $r_{\text{vis}} > 5\%$ and full occlusion as $r_{\text{vis}} < 5\%$, where targets below this threshold are typically no longer visually identifiable. For partially occluded views, reviewers additionally assess question answerability, and we retain only questions for which the geometry-derived answer can be reliably recovered from the current observation. This step filters ambiguous questions but does not alter their geometry-derived labels.

For perspective ambiguity, reviewers verify that both objects are clearly identifiable and that the reference and perspective views are free of rendering artifacts or ambiguous framing. The 2D–3D conflict labels are computed automatically from the projected-size ratio r_{proj}. Human review therefore filters configurations and questions for quality and answerability, but does not assign any evaluation label.

A.2 Evaluation Setup

A.2.1 Prompt Templates

We use two prompt variants in our experiments: a standard multiple-choice prompt and a structured reasoning prompt.

Standard Prompt. The standard prompt instructs the model to select the best answer based on visible evidence, without any explicit guidance on assessing observation reliability. It permits abstention via "Cannot determine" but does not actively encourage it. This prompt serves as our primary evaluation setting.

Structured Reasoning Prompt. The structured reasoning prompt explicitly guides the model to assess observation reliability before selecting an answer. It decomposes the reasoning process into two explicit checks: whether the target is visible, and whether the viewpoint is reliable. Only if both checks pass does the model proceed to select a specific answer; otherwise, it defaults to "Cannot determine." This prompt is used in our prompting analysis to investigate whether explicit reasoning guidance can improve observational awareness.

You are answering a visual multiple-choice question about the provided image(s).
Look carefully at the image and choose the best answer.
Rules:

- You must choose exactly one option.
- Choose "Cannot determine" if the image lacks sufficient visual evidence to decide reliably.
- Reply with ONLY one letter.

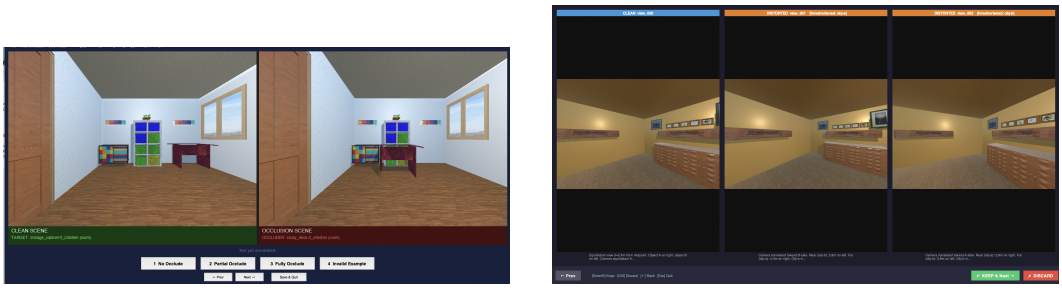
Question:
{question}
Options:
{options}
Reply with ONLY one letter from: A, B, C, ...

Structured Reasoning Prompt.

You are answering a visual multiple-choice question about the provided image(s).
Follow these steps before selecting your answer:
Step 1: Is there sufficient visible evidence to answer the question reliably? Answer Yes or No.
Step 2: Could the current viewpoint make the relevant spatial property misleading? Answer Yes or No.
Step 3: If Step 1 is Yes and Step 2 is No, select the correct option. Otherwise, select {cannot_letter}) Cannot determine.
Question:
{question}
Options:
{options}
Write your step-by-step reasoning, then on a new line write:
`Answer: <letter from A, B, C, ...>`

A.2.2 Evaluation Metrics

Models are presented with multiple-choice questions and required to select exactly one option, including *Cannot determine* where applicable. We report the following metrics: **Answerable Accuracy (Ans.)** is calculated by Ans. $= \frac{\#\text{ correct (ans)}}{\#\text{ total (ans)}}$. **Unanswerable Accuracy (Unans.)** is calculated by



(a) Occlusion annotation examples.

(b) Perspective ambiguity examples.

Figure 9: Annotation Interface for Occlusion and Perspective Scenes.



Figure 10: Examples from the real-image viewpoint-consistency probe. Each column shows the same object pair from two viewpoints, with each view closer to a different object so that the nearer one projects larger. Images are taken from Matterport walkthroughs of real staged apartments.

Unans. $= \frac{\#\text{ correct (unans)}}{\#\text{ total (unans)}}$. **Overall Accuracy (All)** is calculated by All $= \frac{\#\text{ correct (ans)}+\#\text{ correct (unans)}}{\#\text{ total (ans)}+\#\text{ total (unans)}}$, **ViewSel** is calculated by ViewSel $= \frac{\#\text{ correctly selected views}}{\#\text{ total view selection questions}}$, and **AbstainViewSel** is calculated by AbstainViewSel $= \frac{\#\text{3D-consistent Stage-1 predictions}+\#\text{successful abstain-and-select cases}}{\#\text{perspective-conflict questions}}$.

A.2.3 Implementation Details

We fine-tune Qwen2.5-VL-7B-Instruct with LoRA adapters on the occlusion and perspective training splits separately to study cross-condition abstention transfer. LoRA is applied to all linear projections in the language tower (r=16, α=32, dropout 0.05), while the vision encoder remains frozen. Training uses bf16, gradient checkpointing, cosine learning-rate scheduling, and a warm-up ratio of 0.03. Loss is computed only on assistant response tokens. The occlusion adapter is trained on 5.2K samples for 1 epoch with learning rate 3e$-$5, batch size 4, and gradient accumulation 2. The perspective adapter is trained on 3.0K samples for 2 epochs with learning rate 1e$-$4, batch size 2, and gradient accumulation 8. Training is conducted on 2$\times$A100 80GB GPUs per adapter. At evaluation, each adapter is tested on held-out scenes from both benchmarks to measure in-domain and cross-domain abstention transfer.

A.3 Extra Experiments

A.3.1 Open-Ended Robustness Probe

To rule out multiple-choice shortcuts, we re-evaluate a stratified sample of 600 questions using an open-ended protocol. The model receives no answer options and responds with a single constrained word (e.g., `left`, `right`, `same`, or `unsure`). The sample includes 300 2D-3D conflict questions and 300 occlusion questions, evenly split between answerable partial-occlusion and unanswerable full-occlusion cases. As shown in Table 2, Gemini-2.5-Flash shows nearly the same behavior under both protocols. Occlusion performance is unchanged, and under perspective conflict, the model still predominantly follows the projected 2D relation while rarely abstaining. 3D-consistent responses

Table 2: Protocol robustness for Gemini-2.5-Flash. MC denotes multiple-choice evaluation, OE denotes open-ended evaluation without answer options, and Text denotes text-only evaluation without the image.

Setting	Metric	MC	OE	Text
Perspective conflict	3D-Cons.↑	33.7	41.3	0.0
	Abstain	6.7	2.4	79.3
	2D-Foll.↓	53.4	51.0	12.1
Occlusion	Ans.↑	49.1	49.3	20.0
	Unans.↑	56.0	56.7	49.3

Table 3: Abstention precision and recall. For occlusion, the positive class is an unanswerable question (22.0% base rate); for perspective, it is a 2D–3D conflict view (46.6% base rate).

	Occlusion		**Perspective**	
Model	**Abs-P**	**Abs-R**	**Abs-P**	**Abs-R**
GPT-4o	41.9	63.6	46.8	38.3
GPT-5-mini	62.9	8.4	61.5	8.9
GPT-5.4	52.5	21.1	48.2	16.7
Gemini-2.5-Flash	50.7	49.4	68.2	4.8
Gemini-3-Flash	61.8	51.4	73.6	6.2
Qwen2.5-VL-7B	45.3	42.0	33.1	27.1
Qwen2.5-VL-32B	50.4	45.5	47.2	16.4
InternVL3-8B	54.0	8.3	83.3	0.8

are slightly higher in the open-ended setting, suggesting that the multiple-choice format does not inflate performance. We additionally shuffle answer options for every question, yielding a balanced gold-label distribution.

A.3.2 Selective-Prediction Analysis

We further analyze whether model abstention is aligned with the cases in which visual evidence is insufficient or potentially misleading. Sec. A.3.2 reports abstention precision and recall, treating unanswerable occlusion cases and 2D–3D conflict views as the positive classes, respectively. The results reveal markedly different abstention behaviors across models. Some models abstain relatively often and recover a larger fraction of the designated cases, but with limited precision, while others abstain more selectively and achieve higher precision at the cost of very low recall. No model achieves both precision and recall above 55% in either setting. These results suggest that current models do not consistently align their abstention behavior with challenging visual conditions. For occlusion, this reflects difficulty in identifying genuinely unanswerable observations; for perspective, it indicates that abstention is not systematically triggered by views exhibiting a 2D–3D conflict, rather than implying that all such views should necessarily be abstained on.

A.3.3 Details of the Representation and Attention Analyses

Data. Both analyses use the wall-unequal subset with Qwen2.5-VL-7B-Instruct. For 270 conflict views and 81 clean reference views, we analytically project the 3D mesh corners through the render camera to obtain image-plane object boxes, together with mesh-derived physical-size labels. Among the 270 conflict views, 198 are *strict reversals*, where the physically smaller object projects larger along the queried dimension; these are used for the attention analysis.

Linear probe. We freeze the model and extract features from vision-tower layers (vis8–vis32) and LM layers (lm0–lm28). Object features are mean-pooled within the left and right bounding boxes, and the probe input is $[\mathbf{f}_L; \mathbf{f}_R; \mathbf{f}_L - \mathbf{f}_R]$. A standardized logistic regression (C=0.05) predicts either the physical 3D relation or projected 2D relation using scene-level 5-fold cross-validation (majority baseline ≈53%). Because the two labels are complementary on strict reversals, probes are trained on the decorrelated conflict set. LoRA does not modify the vision tower, so vision-layer probe results

Table 4: Effect of LoRA fine-tuning under different observational uncertainty conditions. we report fine-tuning results on a held-out test split separate from the main benchmark evaluation.

	Occlusion		Perspective Ambiguity			
Model	**Ans.↑**	**Unans.↑**	**Ans.↑**	**3D-Cons.↑**	**Abstain**	**2D-Foll.↓**
Base (Qwen2.5-VL-7B)	53.1	69.3	15.9	3.4	37.9	58.6
LoRA-Fintuned	**74.6**	**94.7**	**79.4**	**13.8**	86.2	**0.0**

Table 5: **Effect of structured prompting on occlusion.** Structured prompting improves unanswerable accuracy, but often reduces accuracy on answerable cases.

Model	**Prompt**	**Occ-Ans.**	**Occ-Unans.**
GPT-5-mini	Standard	66.2	8.4
	Structured	54.7	30.4
Gemini-2.5-Flash	Standard	56.8	49.4
	Structured	50.4	52.7

remain unchanged after fine-tuning. Probe accuracy measures linear decodability and does not by itself imply that the model uses the encoded relation.

Answer-token attention. Using the standard multiple-choice prompt, we record the first answer token's attention to image tokens at each LM layer, averaged over heads with attention-sink tokens removed. For each object, attention density is defined as attention mass inside its bounding box divided by the box-area fraction, controlling for projected object size. We report the density ratio between the projected-larger and truly-larger objects for both the base and LoRA models. For projection-following answers (n=117), the bias reaches 1.84× in middle layers and remains 1.54× in later layers, whereas for 3D-consistent answers (n=13), late-layer attention returns to near parity (0.99×). This suggests, correlationally, that correcting the misleading projection may require a late-stage shift in attention.

A.4 Implementation Details of the Evidence Interventions

We evaluate all interventions on the same 209 wall-unequal reversal views using Gemini-2.5-Flash with temperature 0. The model answers each comparison with exactly one word from {`left`, `right`, `same`, `unsure`}. We compare four input conditions: **Base:**The distorted RGB view only. **Depth as text:** The same RGB view, augmented with the ground-truth camera distances of the left and right objects. **Depth map:** The RGB view together with a pixel-aligned depth map estimated using Depth-Anything-V2-Small. The depth map uses the `turbo` colormap with per-image normalization, and the prompt specifies that red indicates closer regions and blue indicates farther regions. **Reference view:** The distorted RGB view together with an additional reference view of the same object pair. The question remains about the physical size relation in the distorted view. All conditions use the same question wording and differ only in the additional evidence provided.

A.4.1 Real-Image Viewpoint-Consistency Probe

Image sources. We collect real indoor images from two sources: (i) Matterport-based virtual walkthroughs of real rental units from public apartment listings, and (ii) self-captured photographs of household objects. Matterport tours provide dense 3D scans of real interiors and allow controlled viewpoint changes similar to our simulated setting. Platform watermarks and UI overlays are cropped from all screenshots.

Collection protocol. Each example contains two same-category objects photographed from two viewpoints, with each view closer to a different object so that the nearer object projects substantially larger. The model answers the same physical-size comparison question independently from each view, with an abstain option. Because the physical size relation is invariant to viewpoint, committed answers that imply different physical relations across the two views constitute a logical contradiction and certify at least one error without requiring ground-truth size measurements. Abstentions are not

counted as contradictions, making the contradiction rate a conservative lower bound on error. When the two viewpoints reverse image left/right, predictions are mapped to physical object identity before comparison.

Results. On 35 valid pairs without an obvious matched-set prior, Gemini-2.5-Flash produces contradictory size judgments across viewpoints on 23 pairs (66%). In every contradictory case, the model favors the object that projects larger in the corresponding view. Because physical size is invariant to viewpoint, these contradictions certify that at least one prediction is incorrect, providing a conservative, measurement-free estimate of real-image failure.

A.5 Licenses and External Assets

We use AI2-THOR (Apache 2.0) for simulation, open-source models (e.g., Qwen2.5-VL, InternVL3) under their respective licenses, and proprietary models (e.g., GPT and Gemini) via official APIs in accordance with their terms of service.

A.6 Broader Impact

This work studies observational uncertainty in vision-language models and highlights their tendency to produce confident spatial reasoning under incomplete or misleading observations. Improving awareness of unreliable visual evidence may benefit reliability-critical applications such as embodied agents and robotic systems. At the same time, our findings suggest that current models can make overconfident decisions under ambiguous viewpoints, which may lead to unsafe behaviors if deployed without appropriate safeguards. We hope this work encourages future research on uncertainty-aware and more reliable multimodal reasoning systems.

A.7 Limitation

Our framework relies on controlled synthetic 3D environments, which enable systematic control of observational conditions but may not fully capture the complexity and diversity of real-world scenes. In addition, our work focuses on observational uncertainty arising from occlusion and ambiguous viewpoints. While these settings isolate important challenges in spatial reasoning, real embodied environments may involve more complex and dynamic sources of uncertainty, such as motion, temporal changes, or sensor noise. Besides, our evaluation focuses on single-step viewpoint assessment, rather than interactive exploration. Extending observational awareness to long-horizon embodied decision-making remains an important direction for future work.