SOLVABLE, YET STILL WRONG: AN AXIS-WISE PERSPECTIVE ON SPATIAL REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite significant progress in Vision-Language Models (VLMs), spatial reasoning, which involves inferring the spatial relationships between objects, remains challenging. Existing benchmarks typically report a single aggregated accuracy per task, making it difficult to determine whether a spatial relation is solvable from the available geometric evidence or how models fail. In practice, different spatial axes (*e.g.*, horizontal, vertical, and depth) require different geometric evidence. We therefore introduce an axis-wise analysis that examines, for each relation, what information makes it solvable and how models fail when relevant information is supplied. Our analysis reveals a gap between what VLMs can spatially infer and what they ultimately answer: models can possess the relevant geometric information yet select a relation from the wrong axis. Through controlled analysis on real-world datasets, we reveal three key findings: i) Horizontal, vertical, and depth relations differ substantially in their geometric solvability and in how accurately VLMs can solve them from the available evidence, showing that spatial reasoning is not uniform across axes. ii) We observe that, when VLMs make mistakes, they often select a relation from an incorrect axis, even when at least one geometrically valid spatial relation is among the options. Interestingly, restricting the multi-axis options to the correct axis substantially improves accuracy, highlighting axis selection as an important error source. iii) This axis-wise failure can be alleviated by supplying axis-specific geometric evidence. Providing geometry as the relation implied by grounded objects significantly improves accuracy, whereas directly providing object coordinates sometimes even hurts performance, indicating that how geometric evidence is represented matters. Based on these findings, we introduce AxisCue, a training-free, axis-aware inference method that allows axis-conditioned spatial reasoning. AxisCue identifies the queried axis, grounds the reference objects, and converts them into axis-specific geometric evidence, eliciting relevant geometric information from the model to improve accuracy. Overall, our analysis demonstrates that an axis-wise perspective exposes failure modes hidden by aggregated spatial accuracy and provides a principled basis for deriving simple, targeted improvements to spatial reasoning.

1 INTRODUCTION

VLMs have evolved rapidly in recent years, demonstrating strong capabilities in vision-language tasks (Wu & Xie, 2024; Chen et al., 2024b; Liu et al., 2024). Nevertheless, spatial reasoning, which requires models to infer spatial relations between objects, remains challenging. Addressing this challenge is important because spatial reasoning is fundamental to a wide range of applications, including visual question answering (Liu et al., 2024; Chen et al., 2024a), embodied agents (Singh et al., 2023; Chang et al., 2025), and robotic navigation (Anwar et al., 2025), where models must interpret object configurations and translate visual information into spatially grounded decisions.

Most existing spatial reasoning benchmarks (Kamath et al., 2023; Fu et al., 2024; Du et al., 2024; Tong et al., 2024; Ma et al., 2025) report a single aggregated accuracy per task, treating different spatial relations across axes as a homogeneous problem. However, spatial axes (*e.g.*, horizontal, vertical, and depth) can depend on fundamentally different geometric evidence, making aggregated

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.

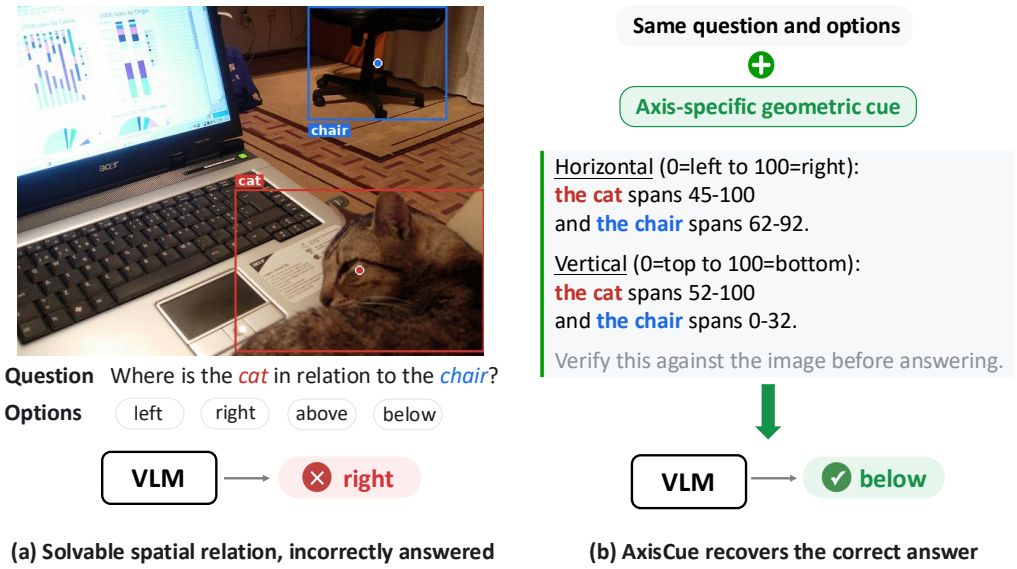


Figure 1: **VLMs answer incorrectly even when the spatial relation is geometrically solvable from the camera perspective.** Example from Qwen3-VL-8B: (a) The model wrongly answers "*right*" (horizontal relation), although it can localize the cat and chair correctly when explicitly asked to ground them. (b) Providing these grounded object boxes as an axis-specific geometric cue recovers the correct answer, *below*.

accuracy hard to interpret. To provide a more fine-grained evaluation beyond a single accuracy score, a recent benchmark (Liu et al., 2025) additionally reports performance across spatial axes, revealing substantial accuracy gaps between axes. Nevertheless, the axis is primarily used as a reporting dimension. Such performance differences do not reveal whether a spatial relation is actually solvable from the available geometric evidence, nor do they explain the sources of model failures.

To address these gaps, we propose an axis-wise analysis that examines, for each spatial relation, what geometric information makes the relation solvable and how VLMs fail even when relevant information is available. Rather than treating spatial reasoning as a homogeneous task, we decompose it into horizontal, vertical, and depth axes for analysis. Through controlled experiments across three real-world datasets (*i.e.*, WhatsUp (Kamath et al., 2023), EmbSpatial (Du et al., 2024), and CV-Bench (Tong et al., 2024)), we uncover three key findings: **i) Spatial solvability and VLM accuracy differ across axes.** Horizontal and vertical relations are largely determined by 2D object geometry, whereas depth relations may require 3D evidence. VLMs show substantial performance gaps even when 2D geometry largely resolves the relation, whereas their depth-axis performance is closer to what the available 2D evidence supports. Therefore, an aggregated accuracy obscures whether failures arise from reasoning errors or insufficient geometric evidence. **ii) Spatial reasoning errors often arise from incorrect axis selection.** Even when at least one geometrically valid relation is provided, VLMs often select an option corresponding to a different spatial axis in error cases. Interestingly, restricting the multi-axis options to the ground-truth axis (*e.g.*, limiting the choices to the vertical axis with above/below options) substantially improves accuracy, highlighting axis selection is an important error source. **iii) Axis-specific geometric evidence can alleviate these failures, and its format matters.** Supplying geometry as the relation implied by grounded objects significantly improves accuracy, whereas directly providing object coordinates sometimes even hurts performance. This indicates that geometric evidence is more effective when represented in a form that is aligned with the queried spatial axis.

Motivated by these findings, we propose AxisCue, a training-free, axis-aware inference method that enables axis-conditioned spatial reasoning. AxisCue first identifies the spatial axis queried by the question, grounds the relevant objects, and then converts their locations into axis-specific, relation-oriented geometric evidence. By explicitly aligning the provided evidence with the queried axis, AxisCue directly targets the axis-selection errors revealed by our analysis without modifying the model. Fig. 1 shows an example of AxisCue. Although the spatial relation is geometrically

solvable, the model initially produces an incorrect answer. AxisCue grounds the relevant objects using the same model and converts the object boxes into an axis-specific geometric cue. With this targeted geometric evidence, the model correctly identifies the spatial relation and answers correctly. Overall, our analysis demonstrates that an axis-wise perspective uncovers failure modes hidden by aggregated spatial accuracy. More importantly, it provides a principled basis for deriving simple, targeted improvements to spatial reasoning. The main contributions are summarized as follows:

- We propose an axis-wise perspective on spatial reasoning, revealing the underlying sources of spatial reasoning errors and providing a principled basis for improving spatial reasoning.
- We introduce AxisCue, a training-free, axis-aware inference method that explicitly supplies axis-specific geometric evidence to guide spatial reasoning.
- Extensive experiments across four models and multiple datasets demonstrate that AxisCue consistently improves spatial reasoning accuracy.

2 Related Work

Failures in Spatial Reasoning. Despite the strong capabilities of advanced VLMs, even simple spatial concepts such as left-right, above-below, and close-far remain challenging (Kamath et al., 2023; Du et al., 2024; Liu et al., 2025; Min et al., 2026). Previous studies have revealed several sources of spatial reasoning errors. A common finding is that VLMs can rely on language priors or textual shortcuts rather than sufficiently grounding their answers in visual evidence (Kamath et al., 2023; Tong et al., 2024; Fu et al., 2024; Liu et al., 2025). This limitation can arise at different stages of spatial reasoning: models may fail to recognize or ground the relevant objects (Liu et al., 2025; Yang et al., 2026), attend to irrelevant visual regions while underutilizing visual tokens (Chen et al., 2025), or fail to reliably bind objects to their relations and represent the underlying geometry (Liu et al., 2023; Yang et al., 2026). While these findings highlight the importance of visual grounding and geometric representation, they do not explicitly examine whether models use the *geometric evidence directly relevant to the queried spatial relation*. In this work, we provide an axis-wise perspective by disentangling spatial reasoning along the horizontal, vertical, and depth axes. We show that the models frequently select an incorrect spatial axis even when the correct relation is geometrically supported, exposing axis selection as a distinct source of spatial reasoning errors.

Improving Spatial Reasoning in VLMs. Existing approaches for improving spatial reasoning in VLMs can be broadly categorized into training-based and training-free approaches. Training-based methods build spatial datasets and fine-tune models with explicit spatial supervision (Hong et al., 2023; Chen et al., 2024a; Cheng et al., 2024; Li et al., 2026). While effective, these methods require additional data and training, and their improvements may not generalize well to unseen data. Training-free methods (Shiri et al., 2024; Izadi et al., 2025; Chen et al., 2025; Lee et al., 2025) instead leverage the inherent capabilities of pretrained VLMs to improve reasoning without modifying model parameters, such as attention scaling (Chen et al., 2025) and visual prompting (Izadi et al., 2025). However, these approaches primarily guide *where to look* and provide general visual guidance rather than the specific geometric evidence required to determine a spatial relation. In this work, we propose AxisCue, a simple training-free method that directly provides axis-specific geometric evidence for the queried relation. By supplying the geometric cues relevant to each spatial axis, AxisCue enables VLMs to more reliably identify spatial relations without additional training.

3 Axis-Wise Analysis of Spatial Reasoning

Existing spatial reasoning benchmarks typically aggregate different spatial relations into a single accuracy score, making it difficult to diagnose the error sources. Since the physical world is inherently three-dimensional, we therefore decompose spatial relations into three axes according to **camera perspective**: X-axis (horizontal, *e.g.*, left/right), Y-axis (vertical, *e.g.*, above/below), and Z-axis (depth, *e.g.*, front/behind). This decomposition provides a structured framework for analyzing spatial reasoning along axes with distinct geometric properties, including (1) whether different axes are geometrically solvable from the available evidence (Sec. 3.1), (2) how VLMs fail across axes when relevant geometric evidence resolves the relation (Sec. 3.2), and (3) what forms of geometric evidence improve spatial reasoning for different axes (Sec. 3.3).

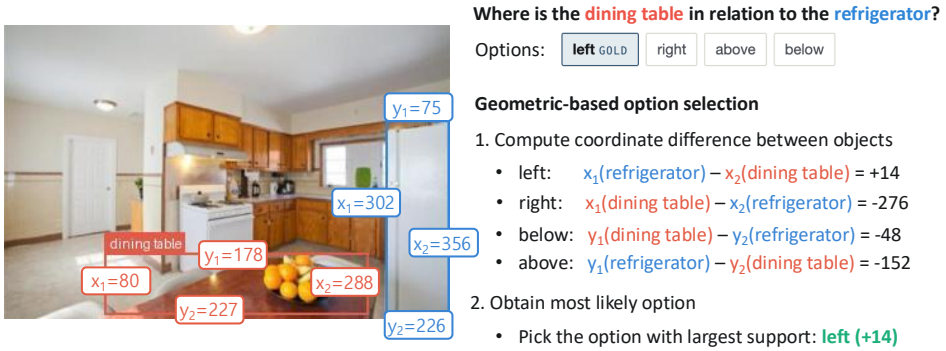


Figure 2: **Illustration of box-conditioned solvability**, where 2D boxes determine spatial relations.

Table 1: **Box-conditioned solvability and VLM accuracy across axes.** Solvability is computed from relations derived from geometric evidence, while VLM accuracy is averaged across 4 models.

Geometric Evidence	WhatsUp			EmbSpatial			CV-Bench
	X-axis	Y-axis	Z-axis	X-axis	Y-axis	Z-axis	Z-axis
2D Object Box	91.1	88.7	99.6	99.5	98.8	67.3	83.8
2D Object Box + Depth	—	—	—	—	—	—	97.0
VLM (4-model averaged)	82.3 $_{-8.8}$	67.3 $_{-21.4}$	78.3 $_{-21.3}$	85.1 $_{-14.4}$	80.4 $_{-18.4}$	62.2 $_{-5.1}$	93.2
Items	2,451	1,660	228	1,236	1,198	1,206	800

3.1 Are Spatial Relations Equally Solvable Across Axes?

We first examine whether each spatial relation can be determined from object geometry alone. We define **box-conditioned solvability** as whether the relation can be inferred from the available object boxes using a fixed geometric rule. It measures how much the available geometric evidence alone determines the relation, without VLM reasoning. As illustrated in Fig. 2, for X- and Y-axis relations, we compare object coordinates along the queried axis and select the most supported option. For Z-axis relations, we use either perspective cues from image position when only 2D annotations are available (*i.e.*, lower objects are nearer) or directly compare 3D coordinates if available.

Experiment Setup. We conduct experiments on three widely used spatial reasoning datasets, including WhatsUp (Kamath et al., 2023), EmbSpatialBench (Du et al., 2024), and CV-Bench (Tong et al., 2024). For ground-truth object boxes, we obtain 2D object annotations for WhatsUp from Microsoft COCO (Lin et al., 2014) and GQA (Hudson & Manning, 2019). EmbSpatial provides the object bounding boxes directly. For CV-Bench, we use its 3D subset and collect 3D object annotations based on SUN RGB-D (Song et al., 2015). Note that we only use images with available ground-truth annotations from each dataset for analysis, so the number of evaluated samples is smaller than that of the full dataset. More details about the datasets are provided in Appendix A.1.

For model evaluation, we consider four recent open-source VLMs: Qwen2.5-VL-7B (Bai et al., 2025), Qwen3-VL-8B (Yang et al., 2025), InternVL3.5-8B (Wang et al., 2025), and LLaVA-OneVision-2-8B (An et al. (2026); LLaVA-OV-2-8B for short).

Results. Table 1 shows that 2D geometry resolves most X- and Y-axis relations ($88.7 - 99.5$), while Z-axis relations depend on the scene: perspective cues are sufficient for tabletop scenes in WhatsUp (99.6) but less reliable for indoor scenes in EmbSpatial (67.3). Notably, Z-axis accuracy improves substantially when depth information is provided in CV-Bench, highlighting the importance of 3D geometry for reliably solving depth relations. Comparing these geometric references with VLM accuracy reveals two distinct bottlenecks: i) on the X and Y axes, VLMs fall well below the reference (*e.g.*, 67.3 vs. 88.7 on the WhatsUp Y-axis), indicating that their errors may not be attributed solely to insufficient geometric evidence; ii) on the EmbSpatial Z-axis, VLM accuracy is already close to the reference (62.2 vs. 67.3), suggesting a bottleneck beyond 2D geometry.

Insights. These results show that spatial solvability varies across axes with the available geometric evidence. More importantly, box-conditioned solvability provides context for interpreting VLM accuracy by showing how well each relation can be determined from the available geometry alone.

Table 2: **Spatial reasoning error distribution.** Each row shows one model's errors on a benchmark, including cross-axis error (*e.g.*, answering left for an above/below problem), same-axis error (*e.g.*, answering above for a below problem), and others (invalid or unrelated responses).

Model	WhatsUp (4-6 Options)			EmbSpatial (4 Options)	
	Cross-Axis	Same-Axis	Others	Cross-Axis	Same-Axis
Qwen2.5-VL-7B	**80.4%**	6.8%	12.8%	**78.6%**	21.4%
Qwen3-VL-8B	**95.7%**	2.7%	1.6%	**83.1%**	16.9%
InternVL3.5-8B	**88.4%**	10.3%	1.3%	**78.4%**	21.6%
LLaVA-OV2-8B	**96.2%**	2.6%	1.2%	**79.8%**	20.2%
Random Choice	71.1%	28.9%	0.0%	61.8%	38.2%

Table 3: **Accuracy with oracle-axis options and cross-axis distractors.** Oracle-axis restricts choices to the ground-truth axis (*i.e.*, only 2 options, such as left/right). Cross-axis distractor also uses 2-option, retaining true answer and sampling an option from another axis (*e.g.*, left/above).

Model	Setting	WhatsUp			EmbSpatial	
		X-axis	Y-axis	Z-axis	X-axis	Y-axis
Qwen2.5-VL-7B	Baseline	83.6	58.4	91.2	85.4	72.0
	Oracle-axis	**97.9** +14.3	**97.0** +38.6	**100.0** +8.8	**92.6** +7.2	**94.0** +22.0
	Cross-axis distractor	68.7	90.4	97.7	92.6	74.1
Qwen3-VL-8B	Baseline	76.9	68.3	87.7	88.3	78.5
	Oracle-axis	**98.3** +21.4	**98.4** +30.1	**95.2** +7.5	**95.6** +7.3	**95.7** +17.2
	Cross-axis distractor	75.1	94.0	95.1	93.2	79.9
InternVL3.5-8B	Baseline	83.0	75.2	88.2	79.6	85.6
	Oracle-axis	**97.5** +14.5	**98.5** +23.3	**96.1** +7.9	**92.9** +13.3	**95.7** +10.1
	Cross-axis distractor	71.0	94.2	96.6	84.7	87.3
LLaVA-OV2-8B	Baseline	85.5	67.2	46.1	86.9	85.4
	Oracle-axis	**98.9** +13.4	**99.0** +31.8	**97.4** +51.3	**96.0** +9.1	**97.8** +12.4
	Cross-axis distractor	54.7	97.9	72.0	93.5	83.9
Chance Accuracy (Baseline)		21.5	23.0	24.1	25.0	25.0
Items		2,451	1,660	228	1,236	1,198

3.2 How Do VLMs Fail When the Relation Is Solvable?

Sec. 3.1 examines whether spatial relations are geometrically solvable with object annotations. In this section, we further investigate how VLMs fail when relevant geometric evidence is available.

Experiment Setup. We continue to use WhatsUp and EmbSpatial for analysis. For each sample, we compare the standard multi-axis setting with an oracle-axis setting, where we remove the possibility of considering irrelevant axes by removing answer options irrelevant to the ground-truth axis from answer candidates, *i.e.*, reduction to an axis-conditional test with 2 options. To isolate the effect of cross-axis competition, we also consider a cross-axis distractor setting, which retains the option from the ground-truth axis and samples a distractor option from another spatial axis.

Results. Table 2 shows the error distribution under the standard multi-axis setting. Compared with random choice, VLM errors are biased toward other axes, with cross-axis errors substantially more frequent than same-axis errors on both WhatsUp and EmbSpatial. This suggests that axis selection is the first immediate error source. Notably, Table 3 indicates that restricting the options to the oracle axis substantially improves accuracy across all models and axes, with accuracy exceeding 90% in every setting. In contrast, the cross-axis distractor setting underperforms the oracle-axis setting, highlighting that competing axis choices can be an important error source. Overall, these results demonstrate the importance of axis selection and its varying impact across spatial axes.

Insights. These results reveal that a substantial portion of VLM spatial reasoning errors stem from wrong-axis selection rather than incorrect directional reasoning.

Table 4: **Axis-specific evidence prompting results on WhatsUp and EmbSpatial.**

Model	Setting	WhatsUp			EmbSpatial		
		X-axis	Y-axis	Z-axis	X-axis	Y-axis	Z-axis
–	2D object box (Table 1)	91.1	88.7	99.6	99.5	98.8	67.3
Qwen2.5-VL-7B	Baseline	83.6	58.4	91.2	85.4	72.0	55.4
	+ Visual prompting	82.9	65.6	69.3	81.9	73.0	51.8
	+ Coordinate	90.1	63.7	72.8	84.9	72.2	55.8
	+ Axis-position	**93.6** +10.0	62.3	68.4	86.2	69.9	56.5
	+ Axis-relation	86.4	**80.6** +22.2	**97.8** +6.6	**94.4** +9.0	**89.4** +17.4	**65.1** +9.7
Qwen3-VL-8B	Baseline	76.9	68.3	87.7	88.3	78.5	67.2
	+ Visual prompting	75.9	72.0	64.5	91.7	87.1	67.6
	+ Coordinate	77.9	69.5	86.0	89.6	78.6	68.2
	+ Axis-position	76.9	72.5	87.3	86.9	79.5	69.0
	+ Axis-relation	**80.7** +3.8	**77.1** +8.8	**98.2** +10.5	**93.8** +5.5	**92.3** +13.8	**69.7** +2.5
InternVL3.5-8B	Baseline	83.0	75.2	88.2	79.6	85.6	59.3
	+ Visual prompting	88.7	**83.7** +8.5	92.5	79.0	91.5	58.0
	+ Coordinate	93.6	73.6	85.1	81.4	85.9	59.6
	+ Axis-position	**95.4** +12.4	79.3	90.4	79.7	87.1	61.3
	+ Axis-relation	85.6	81.3	**99.6** +11.4	**88.6** +9.0	**96.7** +11.1	**68.2** +8.9
LLaVA-OV2-8B	Baseline	85.5	67.2	46.1	86.9	85.4	67.0
	+ Visual prompting	87.5	71.9	53.5	88.3	88.9	66.4
	+ Coordinate	92.2	64.4	61.8	88.3	82.6	68.3
	+ Axis-position	**96.2** +10.7	**88.7** +21.5	**95.6** +49.5	89.8	82.1	68.2
	+ Axis-relation	84.7	78.1	83.8	**94.8** +7.9	**94.0** +8.6	**69.8** +2.8
Items		2,451	1,660	228	1,236	1,198	1,206

3.3 What Form of Geometric Evidence Helps VLMs?

Building on the finding that wrong-axis selection is an important error source, we further investigate whether explicitly providing geometric evidence along the queried axis can improve accuracy.

Experiment Setup. We construct axis-specific evidence using ground-truth boxes and incorporate it into the input prompt to improve spatial reasoning. For comparison, we consider visual prompting and three text prompting settings, including coordinate, axis-position, and axis-relation. All settings provide the same ground-truth boxes and differ only in how the geometric information is represented. Each piece of evidence is appended after the question and its answer options. For the EmbSpatial question involving *cabinet* and *bag*, the three text prompting settings are illustrated below:

+ Coordinate

```
Spatial hint (from detected object positions, may be imperfect):
Object locations (x1, y1, x2, y2) normalized to 0-1000:
cabinet: [562, 631, 1000, 1000]; bag: [456, 852, 546, 977].
```

+ Axis-position

```
Spatial hint (from detected object positions, may be imperfect):
Horizontal (0=left to 100=right):
the cabinet spans 56-100 and the bag spans 46-55.
Vertical (0=top to 100=bottom):
the cabinet spans 63-100 and the bag spans 85-98.
```

+ Axis-relation

```
Spatial hint (from detected object positions, may be imperfect):
the cabinet is to the right of the bag.
```

Results. Table 4 compares different forms of evidence. We observe that: i) Axis-aligned evidence generally outperforms visual prompting and raw coordinates. ii) Coordinates provide inconsistent gains and sometimes hurt accuracy, likely because they require numerical comparison and translation into relation words, which VLMs may handle unreliably. iii) Axis-relation cues often achieve the best accuracy on the Y- and Z-axes, while axis-position performs comparably or better on the WhatsUp X-axis. Notably, axis-specific prompting can exceed the geometry-based reference (*i.e.*, 2D object box) in some settings, indicating that VLMs exploit visual cues beyond the 2D geometry.

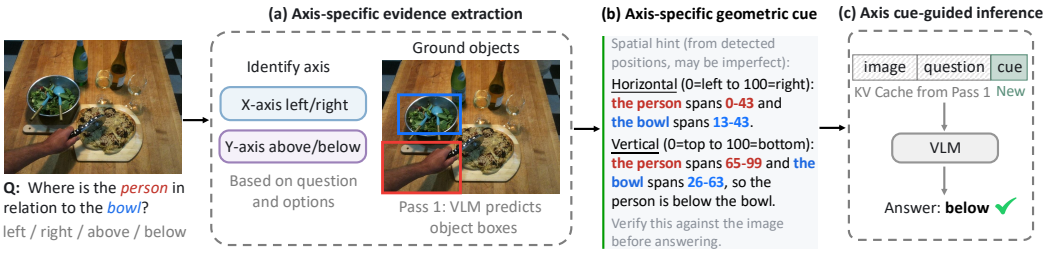


Figure 3: **Overview of axis-aware inference.** (a) AxisCue identifies the queried axis, grounds the referenced objects with the model itself, and projects each box onto the corresponding axis to form (b) an axis-specific geometric cue. (c) The cue is appended to the question, and the model produces the answers in a second pass while reusing the KV Cache of the image and question.

Insights. These results show that the representation of geometric evidence matters for spatial reasoning. Axis-specific cues are more effective as they directly align with the queried axis. Comparing VLM accuracy with 2D object box setting also reveals the reasoning gap: *performance below the reference suggests underutilization of available geometry, while exceeding it indicates the use of visual cues beyond the given geometry evidence.*

4 Axis-Aware Inference

Overview. Based on our axis-wise analysis, we introduce AxisCue, a training-free axis-aware inference method that explicitly conditions spatial reasoning on the queried axis. As shown in Fig. 3, given an input image and question, AxisCue first extracts axis-specific geometric evidence from the reference objects and then incorporates axis cues into the input prompt for inference. By converting visual geometry into cues aligned with the queried axis, AxisCue targets the wrong-axis selection errors and improves spatial reasoning, without modifying model parameters.

Axis-Specific Geometric Evidence Extraction. Given a spatial reasoning question, we first identify the queried spatial axis from the question and answer options (*e.g.*, left/right for the X-axis, above/below/on/under for the Y-axis, front/behind/closer/farther for the Z-axis). Then, we ground the reference objects in the question and obtain their bounding boxes. An example of the grounding prompt is *"<question> Before answering, locate the cart and the green trees in this photo..."*. For each queried axis, we project the grounded boxes onto that axis and convert their relative positions into text. As shown in Fig. 3(b), instead of supplying raw image coordinates, we provide axis-specific geometric cues along the queried axes, enabling axis-aware spatial reasoning.

Axis Cue-Guided Inference. The extracted axis cue is appended to the original question as an additional textual prompt. The model then performs a second inference pass using the original input together with the axis-specific cue. To reduce the computational overhead, we reuse the KV cache computed during the first grounding pass and only process the additional prompt tokens. In this way, AxisCue provides explicit axis-conditioned geometric evidence without changing the model architecture or parameters More details and computational cost can be found in Appendix B.

5 Experiment

5.1 Main Results

Table 5 shows the results on WhatsUp and EmbSpatial. AxisCue consistently improves the overall accuracy of all models on both datasets. The improvements are particularly notable on the Y-axis and Z-axis, with gains up to $+14.1$ and $+37.8$ on WhatsUp, respectively. Table 6 also reports results on CV-Bench. Notably, both the 2D-Relation and 3D subsets use a 2-option setting, with baselines already near the ceiling. Nevertheless, AxisCue still yields reasonable improvements, showing that its benefits extend beyond resolving axis-selection errors. Overall, axis-specific geometric cues enhance spatial reasoning across models and datasets, with larger gains when axis selection is required.

We also compare our method with a training-free method AdapVis (Chen et al., 2025) in Table 7. AdapVis applies attention scaling, using 20% of the data for hyperparameter tuning and the remaining 80% for testing. We follow the same protocol for a fair comparison. Notably, our method consistently outperforms AdapVis, demonstrating the effectiveness of axis-aware inference.

Table 5: **Results on WhatsUp and EmbSpatial datasets.** GPT-5-mini (OpenAI, 2025) and Claude-Sonnet-4.5 (Anthropic, 2025) results are from Qwen3.5 (Qwen Team, 2026).

Model	WhatsUp				EmbSpatial			
	X-axis	Y-axis	Z-axis	Overall	X-axis	Y-axis	Z-axis	Overall
GPT-5-mini	–	–	–	–	–	–	–	80.7
Claude-Sonnet-4.5	–	–	–	–	–	–	–	71.8
Qwen3.5-122B-A10B	–	–	–	–	–	–	–	83.9
Qwen2.5-VL-7B	85.6	62.1	91.3	77.0	85.4	72.0	55.4	71.0
+ AxisCue (**Ours**)	86.9	76.2	95.2	83.2	89.1	84.6	58.2	77.3
Improvement	+1.3	+14.1	+3.9	+6.2	+3.7	+12.6	+2.8	+6.3
Qwen3-VL-8B	80.1	71.7	87.8	77.3	88.3	78.5	67.2	78.1
+ AxisCue (**Ours**)	83.4	78.1	92.6	81.8	91.8	90.9	67.4	83.4
Improvement	+3.3	+6.4	+4.8	+4.5	+3.5	+12.4	+0.2	+5.3
InternVL-3.5-8B	86.4	77.7	88.3	83.2	79.6	85.6	59.3	74.9
+ AxisCue (**Ours**)	88.7	80.0	93.9	85.7	88.4	90.6	59.5	79.6
Improvement	+2.3	+2.3	+5.6	+2.5	+8.8	+5.0	+0.2	+4.7
LLaVA-OV-2-8B	87.5	70.8	46.1	79.2	86.9	85.4	67.0	79.8
+ AxisCue (**Ours**)	87.9	78.6	83.9	84.2	91.0	90.8	67.7	83.2
Improvement	+0.4	+7.8	+37.8	+5.0	+4.1	+5.4	+0.7	+3.4
Items	2,861	1,867	230	4,958	1,236	1,198	1,206	3,640

Table 6: **Results on CV-Bench.** 2D Rel. and 3D subsets use a 2-option setting.

Model	2D Rel.	3D
Qwen2.5-VL-7B	88.3	82.7
+ AxisCue (**Ours**)	88.9 +0.6	85.4 +2.7
Qwen3-VL-8B	91.4	92.8
+ AxisCue (**Ours**)	92.3 +0.9	92.8
InternVL-3.5-8B	87.7	86.0
+ AxisCue (**Ours**)	89.8 +2.1	88.8 +2.8
LLaVA-OV-2-8B	93.1	93.1
+ AxisCue (**Ours**)	93.4 +0.3	93.1

Table 7: **Comparison with AdapVis on WhatsUp, with LLaVA-1.5-7B as the baseline.**

Subset	LLaVA-1.5	AdapVis	AxisCue (**Ours**)
Controlled_A	60.3	84.9	**86.1**
Controlled_B	73.1	83.8	**93.0**
COCO_one	53.0	53.6	**63.7**
COCO_two	58.2	59.9	**73.3**
VG_one	35.9	42.7	**66.9**
VG_two	40.8	48.1	**80.3**
Average	53.6	62.2 +8.6	**77.2** +23.6

Comparison with Alternative Methods. Table 8 compares AxisCue with visual prompting (object box and set-of-mark (Yang et al., 2023)), coordinate text prompting, standard Chain-of-Thought (CoT) (Wei et al., 2022), and Compositional CoT (Mitra et al., 2024). All prompting methods use the same grounded boxes for a fair comparison. AxisCue consistently outperforms these alternatives. Since the grounded boxes are identical, the performance gap isolates the effect of geometric representation, highlighting the importance of how geometry is presented.

5.2 Ablation Study

Is Object Grounding Reliable? To assess the quality of the grounded object boxes used for axis-specific evidence extraction (Fig. 3(a)), we use the predicted boxes to derive the spatial relation and select the corresponding option from the given choices, following the procedure in Sec. 3.1. Table 9 shows that directly using the relation derived from grounded boxes achieves reasonable accuracy against the baseline. For example, Qwen2.5-VL-7B achieves 72.5 and 67.1 on WhatsUp and EmbSpatial, respectively, compared with 77.0 and 71.0 for the baseline. This demonstrates that the grounded object boxes provide reliable geometric information for deriving spatial relations. More results on grounding reliability can be found in Appendix C.1.

Do VLMs Read the Axis Cue? We further investigate whether VLMs treat the axis cue as evidence to be interpreted or simply follow the provided relation. Specifically, we intentionally flip the relation in AxisCue while keeping other settings unchanged, *e.g.*, only swapping the position of two objects in the cue prompt in Fig. 3(b). As shown in Table 9, despite the flipped relation cue,

Table 8: **Comparison with alternative methods,** including visual prompting, coordinate text prompting, and CoT. All prompting methods use the same grounded boxes for fair comparisons.

Model	Setting	WhatsUp	EmbSpatial
Qwen2.5-VL-7B	Baseline	77.0	71.0
	+ Visual prompting (Object Box)	77.8	66.5 $_{-4.5}$
	+ Visual prompting (Set-of-Mark)	80.1	69.0 $_{-2.0}$
	+ Coordinate	79.6	71.0
	+ CoT	68.2 $_{-8.8}$	70.9 $_{-0.1}$
	+ Compositional CoT	73.8 $_{-3.2}$	67.5 $_{-3.5}$
	+ AxisCue (**Ours**)	**83.2** $_{+6.2}$	**77.3** $_{+6.3}$
Qwen3-VL-8B	Baseline	77.3	78.1
	+ Visual prompting (Object Box)	76.8 $_{-0.5}$	79.9
	+ Visual prompting (Set-of-Mark)	74.6 $_{-2.7}$	81.5
	+ Coordinate	76.7 $_{-0.6}$	78.2
	+ CoT	76.3 $_{-1.0}$	78.5
	+ Compositional CoT	72.2 $_{-5.1}$	80.5
	+ AxisCue (**Ours**)	**81.8** $_{+4.5}$	**83.4** $_{+5.3}$

Table 9: **Evaluation of grounded spatial relations from AxisCue.** For *Grounding boxes derived answer* , we derive relations from the grounded object boxes and use them to select answer options. "AxisCue, relation flipped" uses the reversed derived relation as the inference cue.

Model	Setting	WhatsUp	EmbSpatial
Qwen2.5-VL-7B	Baseline	77.0	71.0
	+ AxisCue	**83.2** $_{+6.2}$	**77.3** $_{+6.3}$
	+ AxisCue, relation flipped	77.1 $_{+0.1}$	65.2 $_{-5.8}$
	Grounding boxes derived answer	*72.5*	*67.1*
Qwen3-VL-8B	Baseline	77.3	78.1
	+ AxisCue	**81.8** $_{+4.5}$	**83.4** $_{+5.3}$
	+ AxisCue, relation flipped	76.2 $_{-1.1}$	67.6 $_{-10.5}$
	Grounding boxes derived answer	*72.3*	*72.1*

Table 10: **Cross-axis and same-axis errors removed by AxisCue.** Actual error counts (baseline $\rightarrow$ +AxisCue) with the percentage of each error type removed, excluding other errors.

Model	WhatsUp		EmbSpatial	
	Cross-Axis	Same-Axis	Cross-Axis	Same-Axis
Qwen2.5-VL-7B	917 $\rightarrow$ 732 $_{-20\%}$	77 $\rightarrow$ 43 $_{-44\%}$	285 $\rightarrow$ 173 $_{-39\%}$	80 $\rightarrow$ 71 $_{-11\%}$
Qwen3-VL-8B	1075 $\rightarrow$ 867 $_{-19\%}$	32 $\rightarrow$ 31 $_{-3\%}$	251 $\rightarrow$ 118 $_{-53\%}$	51 $\rightarrow$ 46 $_{-10\%}$

Qwen2.5-VL-7B achieves 77.1 and 65.2 on WhatsUp and EmbSpatial, respectively. This suggests that VLMs do not blindly follow the supplied axis cue, but instead use it as a linguistic premise. Nevertheless, their ability to verify the premise against the image varies across models and datasets.

Does AxisCue Reduce Axis-Selection Error? Table 10 shows that AxisCue reduces both cross-axis and same-axis errors, with cross-axis errors reduced by up to 55% on EmbSpatial. This supports the effectiveness of AxisCue in resolving axis-selection errors.

6 Conclusion

We introduce axis-wise analysis to uncover the source of spatial reasoning errors and demonstrate that supplying axis-specific geometric evidence can improve accuracy. Extensive experiments across four models show that our axis-aware inference method consistently enhances performance across spatial axes. We hope our findings inspire new perspectives on spatial reasoning in VLMs.

AI USE STATEMENT

In this work, we used generative AI tools for [implement methods, support qualitative and thematic data analysis, interpret results]. We have not used generative AI tools for [propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation] and [generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, clean and reformat dataset] are not applicable to this work. Additionally, we used generative AI tools for [identify relevant literature]. We have reviewed all AI-assisted work. In particular, all LLM-generated code was verified and tested to ensure the correctness and reproducibility of the resulting experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided implementation details (Appendix B) to facilitate reproduction of our experiments, including grounding and axis-specific prompts. Our method is training-free and does not require additional model training. We evaluate multiple publicly available VLMs following the default evaluation settings from VLMEvalKit (Duan et al., 2024) and each dataset. All experiments use deterministic greedy decoding for reproducibility. We will release the code upon publication.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Anthropic. Claude sonnect. `https://www.anthropic.com/claude/sonnet/`, 2025.

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In *ICRA*, pp. 2838–2845, 2025.

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 preprint arXiv: 2502.13923*, 2025.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A benchmark for planning and reasoning in embodied multi-agent tasks. In *ICLR*, 2025.

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*, pp. 14455–14465, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, pp. 27056–27087, 2024b.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? An attention mechanism perspective on focus areas. In *ICML*, pp. 9910–9932, 2025.

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. *NeurIPS*, 37: 135062–135093, 2024.

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*, pp. 346–355, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, pp. 11198–11201, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, pp. 148–166, 2024.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, pp. 20482–20494, 2023.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. Visual structures help visual reasoning: Addressing the binding problem in LVLMs. In *NeurIPS*, 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, pp. 9161–9175, 2023.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Leonidas Guibas, Mikaela Angelina Uy, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *ICCV*, pp. 9241–9251, 2025.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *ICLR*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 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 *ACL*, pp. 620–632, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *ECCV*, pp. 216–233, 2024.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *ICCV*, pp. 6924–6934, 2025.

Cheolhong Min, Jaeyun Jung, Daeun Lee, Hyeonseong Jeon, Yu Su, Jonathan Tremblay, Chan Hee Song, and Jaesik Park. Why far looks up: Probing spatial representation in vision-language models. In *ECCV*, pp. 22–39, 2026.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional Chain-of-Thought Prompting for Large Multimodal Models . In *CVPR*, pp. 14420–14431, 2024.

OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *EMNLP*, pp. 21440–21455, 2024.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *ICRA*, pp. 11523–11530, 2023.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, pp. 567–576, 2015.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: a fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, pp. 87310 – 87356, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pp. 24824 – 24837, 2022.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, pp. 13084–13094, 2024.

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.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

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. In *ICLR*, pp. 157051–157088, 2026.

A More Details on Axis-Wise Analysis

A.1 Details About Spatial Solvability Analysis

Spatial Axes. We decompose spatial relations into three axes according to **camera perspective**: X-axis (horizontal, *e.g.*, left/right), Y-axis (vertical, *e.g.*, above/below), and Z-axis (depth, *e.g.*, front/behind). An image is considered an X-axis sample if its ground-truth answer corresponds to a horizontal relation, *e.g.*, left/right, and similarly for the Y- and Z-axes. This decomposition enables a structured analysis of spatial reasoning across axes.

Dataset Details. Table 11 shows the detailed datasets and subsets used for the spatial solvability analysis in Sec. 3.1. For WhatsUp, we include five subsets: `coco_qa_one_obj`, `coco_qa_two_obj`, `vg_qa_one_obj`, `vg_qa_two_obj`, and `controlled_B`, yielding 4339 samples with $2,451$ X-axis, $1,660$ Y-axis, and 228 Z-axis relations. The 2D bounding boxes are obtained from COCO annotations or GQA scene graphs, while no depth annotations are available for these subsets. Notably, `controlled_B` does not provide 2D annotations. To facilitate our analysis, we prompt Qwen3.5-122B-A10B to ground the queried objects and manually verify the results, ensuring that each grounded object correctly matches the corresponding object in the question. For EmbSpatial, we use the `ai2thor`, `mp3d`, and `scannet` subsets, comprising $3,640$ instances with

Table 11: **Datasets and subsets used in spatial solvability analysis.**

Subset	X-axis	Y-axis	Z-axis	Items	2D Box Source	Depth Source
WhatsUp						
coco_qa_one_obj	1,140	1,107	—	2,247	COCO annotations	None
coco_qa_two_obj	279	161	—	440	COCO annotations	None
vg_qa_one_obj	768	389	—	1,157	GQA scene graphs	None
vg_qa_two_obj	264	3	24	291	GQA scene graphs	None
controlled_B	—	—	204	204	VLM + manual check	None
Total	2,451	1,660	228	4,339	—	—
EmbSpatial						
ai2thor	416	412	411	1,239	Benchmark's boxes	None
mp3d	420	386	395	1,201	Benchmark's boxes	None
scannet	400	400	400	1,200	Benchmark's boxes	None
Total	1,236	1,198	1,206	3,640	—	—
CV-Bench						
3D_Depth	—	—	400	400	Benchmark's boxes	SUNRGBD
3D_Distance	—	—	400	400	Benchmark's boxes	SUNRGBD
Total	—	—	800	800	—	—

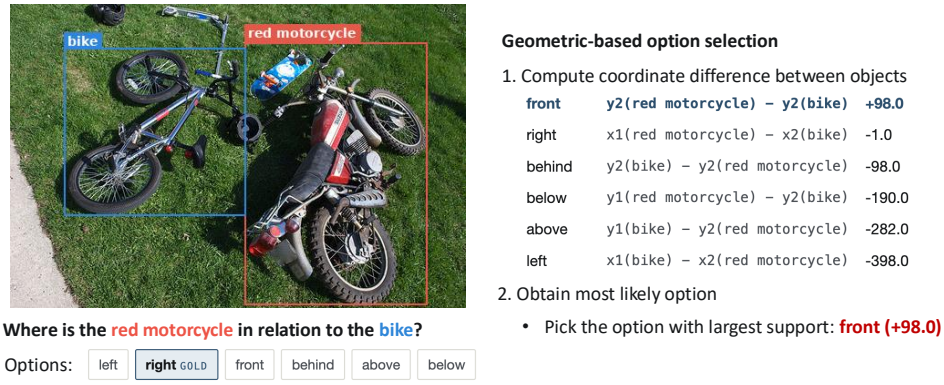


Figure 4: **Failure case of geometric rule.** Z-axis option overrides the correct X-axis option.

1,236 X-axis, 1,198 Y-axis, and 1,206 Z-axis relations. Their 2D bounding boxes are provided by the benchmark. For CV-Bench, we use the `3D_Depth` and `3D_Distance` subsets, which contain 800 Z-axis samples in total. Their 2D bounding boxes are provided by the benchmark, and depth information is obtained from SUNRGBD (Song et al., 2015). These subsets provide complementary sources of 2D and depth information for evaluating whether each spatial relation is geometrically solvable from the available evidence.

When Does Geometric Rule Fail? Table 1 shows that the X- and Y-axis relations in WhatsUp are not fully determined by the 2D object boxes. We observe that many failures arise from the interaction with depth options. Specifically, our perspective-based depth cue can assign substantial support to a depth option, overriding the correct X- or Y-axis option. An example is shown in Fig. 4. The red motorcycle is to the *right* of the bike. However, the perspective-based heuristic assigns a larger support to *front*, causing the Z-axis option to override the correct X-axis option. It is worth mentioning that failures of the heuristic rule do not necessarily imply that a spatial relation is not geometrically solvable, but may reflect limitations of the rule itself. Nevertheless, the heuristic performs well across most settings and still provides a useful reference for interpreting VLM accuracy when geometric evidence is available.

Table 12: **Inference cost of AxisCue with Qwen2.5-VL-7B.** Per-item latency on WhatsUp, measured on a single NVIDIA A100 GPU with an image size of 1024×768.

Model	Setting	Latency (ms)		
		Grounding Pass	Cue Pass	Total
Qwen2.5-VL-7B	Baseline	–	–	618
	+ AxisCue, w/o KV Cache	1280	598	1878
	+ AxisCue, with KV Cache	1280	**252**	**1532**

B More Implementation Details

We evaluate our approach on multiple spatial reasoning datasets, including WhatsUp (Kamath et al., 2023), EmbSpatialBench (Du et al., 2024), and CV-Bench (Tong et al., 2024). For model evaluations, we apply our approach to four recent VLMs, including Qwen2.5-VL Bai et al. (2025), Qwen3-VL-8B (Yang et al., 2025), InternVL3.5-8B (Wang et al., 2025), and LLaVA-OV-2-8B (An et al., 2026). We follow the evaluation setting of VLMEvalKit (Duan et al., 2024) and each dataset.

Grounding Prompt. Since different models use different output formats for object grounding, we follow their official guidelines when constructing the grounding prompts. Taking the question *"Where is the cart in relation to the green trees? Answer with left, right, front, behind, above or below."* as an example, the grounding prompts for different models are listed below:

Qwen2.5-VL, Qwen3-VL (Grounding format: JSON)

```
Where is the cabinet in relation to the bag? Answer with left,
right, front, behind, above or below.
Before answering, locate the cabinet and the bag in this photo.
Output only their bounding boxes as JSON: [{"bbox_2d": [x1, y1, x2,
y2], "label": "<name>"}, ...]
```

InternVL-3.5 (Grounding format: <box>[[x1,y1,x2,y2]]</box>)

```
Where is the cabinet in relation to the bag? Answer with left,
right, front, behind, above or below.
Before answering, provide the bounding box coordinates of
<ref>cabinet</ref> and <ref>bag</ref>.
```

LLaVA-OV-2 (Grounding format: JSON)

```
Where is the cabinet in relation to the bag? Answer with left,
right, front, behind, above or below.
Before answering, locate the cabinet in this photo. Output only
its bounding box as JSON: [{"bbox_2d": [x1, y1, x2, y2], "label":
"cabinet"}]
```

Axis-Specific Geometric Cue. We use the same axis cue across all models for axis cue-guided inference in Fig. 3(c). When the question contains options spanning horizontal, vertical, and depth relations, the prompt is formatted as follows:

```
Spatial hint (from detected object positions, may be imperfect):
Horizontal (0 = left to 100 = right):
<Object description>.
Vertical (0 = top to 100 = bottom):
<Object description>.
Depth (0 = top to 100 = bottom; lower = nearer):
<Object description>.
Verify this against the image before answering.
```

Computation Cost. Table 12 measures the inference cost of Qwen2.5-VL-7B on WhatsUp. AxisCue introduces modest additional inference cost when KV Cache is used. The grounding pass takes

Table 13: **Object grounding quality.** The reported metric is the mean IoU between the predicted and ground-truth boxes, averaged over all objects referred to in each question.

Dataset	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3.5-8B	LLaVA-OV-2-8B
WhatsUp	76.8	81.3	75.7	69.7
EmbSpatial	40.9	45.8	39.7	39.3

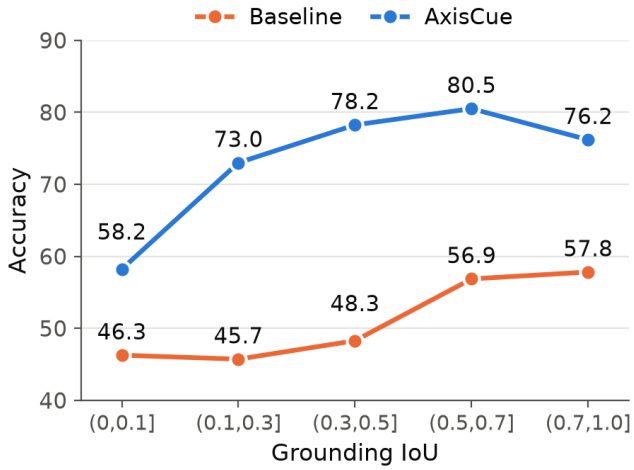


Figure 5: **Accuracy across grounding IoU on WhatsUp with LLaVA-1.5-7B**. AxisCue outperforms the baseline in every IoU range.

1280 ms, while the cue pass requires only 252 ms with KV Cache, resulting in a total latency of 1532 ms, compared with 1878 ms without caching. Therefore, KV Cache reduces the additional cost of the cue pass by more than half, making the second inference pass more efficient.

C More Results and Discussions

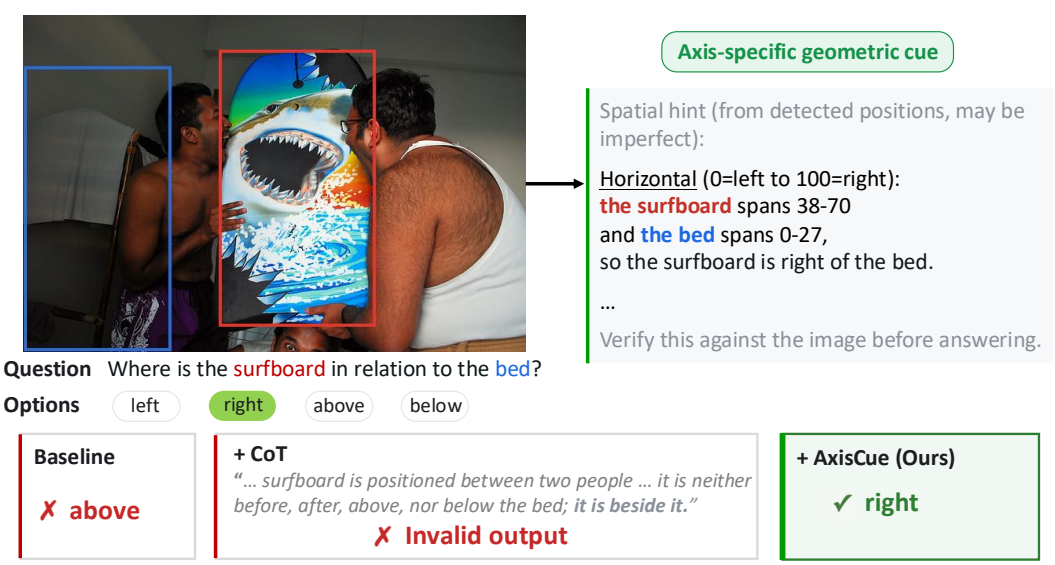
C.1 More Results on Object Grounding

Table 13 reports the Intersection over Union (IoU) between predicted and ground-truth bounding boxes. All four models achieve reasonable mIoU on WhatsUp and EmbSpatial. We further examine the relationship between grounding IoU and spatial reasoning accuracy on WhatsUp with LLaVA-1.5-7B in Fig. 5. Even with relatively low grounding IoU (*e.g.*, 0–0.1), the grounding boxes still preserve relative object order along the axis, yielding meaningful improvements over the baseline. This suggests that AxisCue does not require pixel-level accurate grounding. A coarse object localization can already provide useful geometric information for axis-conditioned reasoning.

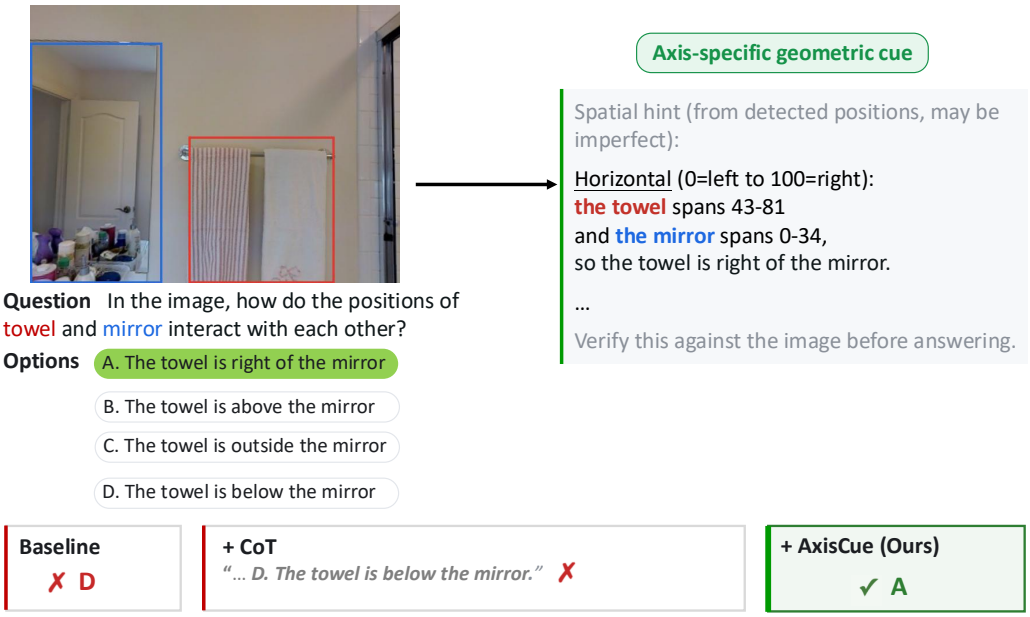
C.2 More Ablation Study

Generalization to Larger Models. We further evaluate AxisCue on larger models on WhatsUp, including Qwen3.5-35B-A3B and Qwen3.5-122B-A10B. AxisCue improves overall accuracy from 79.7 to 83.1 (+3.4) for Qwen3.5-35B-A3B and from 84.5 to 85.8 (+1.3) for Qwen3.5-122B-A10B. These consistent improvements further demonstrate the generalization capability of our method.

Qualitative Comparison. Fig. 6 compares AxisCue with baseline and CoT on WhatsUp and EmbSpatial For the WhatsUp example, the surfboard and bed are clearly separated along the horizontal axis, while their vertical positions do not support a clear above/below relation, making *right* the only geometrically supported option. The baseline answers *above*, selecting a relation from the wrong axis. CoT does not resolve this error, ultimately concluding that "it is beside" the bed, which does not match any option. In contrast, AxisCue provides explicit axis-specific geometric evidence from



(a) WhatsUp example.



(b) EmbSpatial example.

Figure 6: **Visualizations of the outputs of Qwen2.5-VL-7B.**

the grounded boxes, allowing the model to identify the horizontal separation and rule out the vertical options before answering *right*. Similarly, for the EmbSpatial example, the baseline selects the wrong-axis option, and CoT fails to correct the error. With axis-specific geometric evidence, AxisCue selects the correct option. These examples highlight the limitation of unguided reasoning: CoT adds reasoning steps without necessarily providing relevant geometric evidence, whereas AxisCue provides the evidence needed for axis selection.

C.3 Discussion

Limitation. While AxisCue consistently improves multiple VLMs across spatial axes, its effectiveness is constrained by the available 2D geometric evidence, which may be insufficient for depth

relations. A potential direction is to incorporate explicit 3D geometry, obtained either from the VLM itself or external models, to provide richer geometric evidence for spatial reasoning.