FRAMEBIND: DIRECTIONAL GROUNDING FOR ALLOCENTRIC SPATIAL REASONING WITH VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have made rapid progress on spatial reasoning benchmarks, but it remains unclear whether these gains reflect grounded geometric understanding or semantic shortcuts. We study allocentric relative-direction tasks, where the models are prompted to determine the direction a target object is with respect to reference object or to determine the object in a given direction. On controlled scenes we find a pronounced answer-format asymmetry: models appear to be more accurate at selecting an object than at naming the direction, even when the image and underlying relation are held fixed. We trace this gap to a strong camera-centric bias; in our experiments, only 3.6% of a representative model's direction answers match the requested reference frame, while 95% match the camera frame. Component diagnostics show that VLMs localize objects reliably but cannot recover a reference object's orientation or extend partial orientation cues into a usable frame. Only a complete labeled coordinate display can restore performance, likely because it reduces the task to image-plane localization. Motivated by this diagnosis, we introduce *FrameBind*, a training-free, tool-assisted pipeline that overlays a reference-centered coordinate frame on the original RGB image rather than replacing it with a symbolic abstraction. Additionally, we introduce an *Object-Map Inversion* that converts direction queries into sparse, direction-to-object maps to exploit the object-answer preference. Across four benchmarks, FrameBind improves both answer formats over its Qwen3-VL-8B backbone, while object-map inversion adds further gains on controlled scenes.

1 INTRODUCTION

Vision-Language Models (VLMs) have made rapid progress on spatial reasoning benchmarks that span navigation, multi-view understanding, embodied planning, and 3D reasoning (Yang et al., 2025a; Du et al., 2024; Ma et al., 2025; Yang et al., 2026) making spatial understanding a central measure of multimodal intelligence . A single final-answer score, however, conflates several capabilities: recognizing and localizing entities, interpreting the query, reasoning about scene positions, and transforming them into the requested reference frame. A model can exploit object semantics and image-layout regularities to answer correctly without using the geometry that grounds the answer. This ambiguity is especially costly in embodied settings, where an agent must reason from the viewpoint of an object, another agent, or a future pose rather than from its current camera.

We study *allocentric perspective taking*, where spatial queries are posed from the perspective of a reference object (Fig. 1). VLMs struggle with such viewpoint-dependent judgments (Zhang et al., 2025b; Kong et al., 2026; Liu et al., 2022; Beckham et al., 2022): on our controlled 4COMFORT scenes, the model names the correct reference-relative direction for only 3.6% of queries (chance is 25%), while 95% of its answers match the camera's viewpoint instead. Prior remedies supply external cues, rendering the scene from the reference viewpoint (Lee et al., 2025) or replacing the image with a heading-aligned symbolic layout (Jang et al., 2026), but these reach only around chance on the same scenes (21.2% for APC-Vis, Table 2), when applied to our 4 direction tasks.

Answering allocentric query requires three reasoning steps: (1) localizing the reference and target objects, (2) inferring the reference orientation, and (3) expressing the displacement between both in the reference coordinate system. However, it remains unclear which of these three steps fails, and

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.

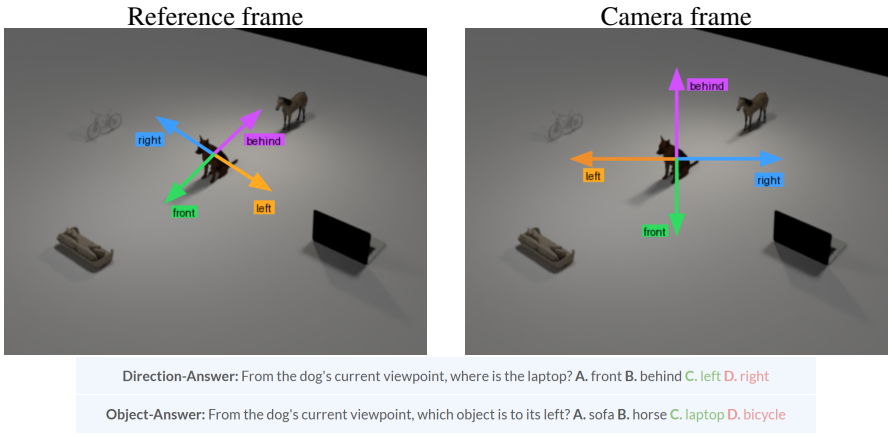


Figure 1: A 4COMFORT scene under the reference-frame and camera-frame interpretations. The query is posed to use the dog as a reference, but we find that most VLMs default to the camera frame, despite explicit prompts to adopt the reference frame.

why. We investigate where this process breaks down through matched object- and direction-answer queries and controlled tests of localization, orientation understanding, and reference-frame use.

Our first observation is an answer-format asymmetry. Given the same image and the same underlying relation, models are markedly more accurate at selecting the object in a requested direction than at naming the direction of a requested object. For example, Figure 1 pairs "From the dog's perspective, which object is to its left?" with "From the dog's perspective, where is the laptop?" The correct answers are "laptop" and "left", respectively. Although both questions concern the same spatial relation, success on the object-answer query does not reliably transfer to the direction-answer query. Higher object-answer accuracy therefore does not, by itself, establish reliable perspective taking. Our diagnostic experiments (Section 3) reveal a systematic pattern in the direction-answer errors: models often answer from the camera's viewpoint rather than the reference object's viewpoint. In Figure 1, the laptop appears to the right in the image but lies to the dog's left. Answering "right" follows the camera frame, whereas the requested answer, "left", requires adopting the dog's frame.

We then separate the components of reference-frame construction. The evaluated models localize named objects and identify objects inside supplied regions reliably, but cannot predict the reference object's orientation, and explicitly supplying a partial orientation cue (a front arrow, or front and left together) does not reliably resolve reference-relative selection even though the information is sufficient. Only a complete set of labeled axes yields a large improvement. We speculate the reason as being that the axes overlap reduces the query to an image-plane localization problem, the one component the models perform well.

These findings motivate FrameBind, a training-free pipeline that supplies a reference-centered coordinate scaffold while retaining the original RGB image. We use off-the-shelf foundation models as tools: SAM3 (Carion et al., 2025) to ground the relevant entities, Depth Anything 3 (Lin et al., 2025) to support perspective projection, and OrientAnything-v2 (Wang et al., 2026) to estimate the reference orientation when the question does not stipulate a viewpoint. A key advantage of FrameBind over competing symbolic-layout methods such as SymPL(Jang et al., 2026) is that it preserves the original image as a source of spatial evidence. When estimated object positions are incorrect, a symbolic layout can misrepresent the scene; FrameBind retains the visible objects, allowing the VLM to use their actual image locations alongside the supplied reference frame. Our comparisons with symbolic-layout baselines demonstrate this advantage when candidate positions are perturbed.

FrameBind also reformulates direction prediction as object selection through a process we call *object-map inversion*. The VLM assigns objects to the reference-relative directions, and the target object's assignment determines the final direction answer. Structural checks validate the predicted map, with direct direction answering used as a fallback. Across 4COMFORT, COMFORTer, Kubric, and 3DSR (Greff et al., 2022; Zhang et al., 2025b; Ma et al., 2025), FrameBind improves both answer formats over all our baselines. Our contributions are:

- Paired inverse queries that expose answer-format sensitivity and separate camera-aligned errors from other failures of reference-object reasoning.
- Controlled diagnostics of localization, orientation, and cue utilization, and a representation study showing that RGB with an overlay is more robust than symbolic layouts to candidate-detection errors.
- FrameBind, an RGB-preserving coordinate overlay with object-map inversion, together with an evaluation that identifies both the gains and the limits of converting object predictions into direction answers.

2 Related Work

Spatial benchmarks and reference-frame diagnostics. Recent benchmarks evaluate spatial reasoning from video (Yang et al., 2025a), embodied egocentric scenes (Du et al., 2024), multiple images (Yang et al., 2026), single natural images (Ma et al., 2025), top-view maps (Li et al., 2024), and controlled synthetic layouts (Kong et al., 2026). Closest to our setting, COMFORT shows that VLMs struggle to adopt alternative reference frames (Zhang et al., 2025b), and ISLE separates perspective-taking failures from detection (Góral et al., 2024). Representation-level studies expose entanglement between image height and distance (Min et al., 2026), limits on manipulating spatial information encoded in object tokens (MirroS Team, 2026), and the functional roles of attention heads in decomposed spatial operations (Ma et al., 2026b). We complement these with paired inverse queries that hold the relation fixed while varying the answer format. We also incorporate controls over localization, orientation, and coordinate cues that test which supplied evidence makes the requested relation accessible.

Learning spatial representations. A large body of work improves spatial understanding through data, geometry, and objectives: synthesized spatial supervision (Chen et al., 2024), depth inputs and depth-related supervision (Cai et al., 2025; Daxberger et al., 2025), progressive supervised and reinforcement training (Yang et al., 2025b; Li et al., 2025a), 3D positional features and geometry priors (Zhu et al., 2025a; Wu et al., 2026; Zheng et al., 2025), reconstruction objectives (Wang et al., 2025a), explicit 3D grounding (Cheng et al., 2026), dynamic and streaming settings (Zhou et al., 2025; Liu et al., 2026), and structured intermediate reasoning (Zhang et al., 2025a; Ma et al., 2026a; Shao et al., 2024). We evaluate several spatially specialized models and find that specialization alone does not close the object–direction gap. FrameBind proposes the Map Inversion to instead capitalize on this gap by transforming direction-answer prompts into object-answer version.

Inference-time scaffolding and perspective reformulation. Inference-time methods change the evidence available to the VLM through 3D foundation-model priors (Ma et al., 2024), keyframe and motion cues (Li et al., 2025b), coordinate-indexed anchors (Lei et al., 2024), or bird's-eye-view abstractions and RGB combinations (Zhu et al., 2025b); others obtain new evidence by imagining alternative views with world models (Yang et al., 2025d; Zhu et al., 2026; Zhang et al., 2026). The closest approaches to FrameBind are APC, which builds a scene abstraction from foundation-model outputs and transforms the perspective (Lee et al., 2025), SpatialPIN, which uses 3D axes visuals without RGB, and SymPL, which reformulates allocentric reasoning through orthogonal projection, symbolic abstraction, and localization queries (Jang et al., 2026). FrameBind shares their goal of making a difficult relation visually accessible, unlike these previous methods, which generate abstract representations, we retain RGB evidence for candidate localization. Instead, we add only the requested reference-centered axes as an overlay. Sec. 6 compares RGB overlays with symbolic layouts under matched scene information and controlled perturbations.

3 Diagnosing Reference-Frame Failures

We investigate object-answer preference through paired queries that prompt for direction and object prediction using the same underlying image. Our analysis probes the VLM's ability to identify the relevant objects, recover their orientations, and use the resulting coordinate frames. We use Qwen3-VL-8B-Instruct as a representative model for diagnosis with deterministic decoding. For these diagnostics, we construct **4COMFORT**, a controlled variant of COMFORT (Zhang et al.,



(a) Direction-answer query (b) Object-answer query

Figure 2: Paired scene examples and confusion matrices for the two query formulations. The red boxes are what the camera-frame frame interpretation for the query would be.

2025b) containing 500 rendered scenes. Each scene has a reference object and four candidate objects placed to its front, left, right, and behind (see Figure 1). The scenes keep the camera pose fixed relative to the reference object and place candidate objects in unambiguous orthogonal positions. 4COMFORT makes the reference-vs.-camera-frame distinction clearly identifiable, along with clear semantic "fronts" facing downwards-left for all reference objects. Our interventions use ground truth orientation and localization information to test the extent of the observed phenomenon.

3.1 Camera-frame bias

Prior spatial understanding benchmarks capture basic perspective-taking with two types of queries. A *direction-answer query* specifies a reference object and a target object, and asks for the target's direction in the reference object's frame. An *object-answer query* specifies a reference object and a direction, and asks which object lies in that direction in the same frame. More formally, consider a reference object r, a target t, and a direction d. The direction-answer query supplies (r, t) and asks for d, whereas the object-answer query supplies (r, d) and asks for t. Consider the example in Figure 1; the former asks "From the dog's perspective, which direction is the car?" while the latter asks "From the dog's perspective, which object is to its left?", when the car is to the dog's left.

Min et al. (2026) showed that VLMs misjudge depth and revert to image-plane left/right/up/down directions. We find a catastrophic failure along similar lines: only 3.6% of direction answers match the requested reference-object relation, whereas 95.4% match the camera-frame or image-plane L/R relation. From Fig. 2, we see that this effect is especially pronounced for reference-relative left and right answers, as they dominate the direction-answer queries. This phenomenon is observed to a lesser extent in object-answer queries, as it is not completely dominated by the L/R options. We observed this phenomenon on Qwen3-VL and other VLM baselines such as LLaVA-OneVision-2 and VST (see App. D). With the exception of GPT-5.6-Luna and InternVideo3, the baselines show the L/R concentration of responses in the direction-answer case as well. This concentration suggests that the models frequently substitute the camera-frame for the requested reference-frame. This finding motivates our study in pursuing interventions that make the intended frame more explicit, with a focus on the favorable object-answer format to avoid the L/R collapse.

3.2 Localization and Orientation

Locating and identifying objects. We begin our investigation by asking the most basic question of whether the VLM can find the reference and target objects in the scene. We apply two diagnostics, one to predict bounding boxes for named objects, and one to name objects of given bounding boxes. Unsurprisingly, the model performs both reliably (Fig. 3a). Visual inspection shows that the models tend to predict larger bounding boxes than supplied ground truth, which leads to lower IoUs at stricter thresholds. Taken together, these results indicate that the evaluated VLMs can reliably localize named objects and identify objects within supplied bounding boxes.

Predicting and applying orientation. We then probe the model's orientation understanding by asking it to predict the orientation in the image plane across 8 cardinal directions. All tested open-

Diagnostic	Metric	Result
Box prediction	Acc@IoU ≥ 0.50 (%)	70.00
	Acc@IoU ≥ 0.40 (%)	84.15
	Acc@IoU ≥ 0.30 (%)	89.56
	Mean IoU	0.5311
Boxed-object naming	Accuracy (%)	98.12

(a) BBox affinity

Relation	None	Front	Front+Left	Full
Left	0.00	0.00	19.56	67.59
Right	0.00	0.00	1.14	57.91
Front	41.01	78.48	83.23	98.04
Behind	66.46	64.30	1.08	96.52
Overall-Obj	26.87	35.70	26.25	80.02
Overall-Dir	3.59	2.74	12.17	41.47

(b) Partial versus complete cues

Figure 3: (a) The results of localization probing. The model shows reliable performance in both localizing and using localization inputs. (b) Reference-frame coordinate cues show an inability to infer a full frame from partial but sufficient cues.

Model	Acc. (%)	Modes	Mass (%)
InternVideo3-8B	43.31	2	86.17
Cambrian-S-7B	1.13	2	96.83
LLaVA-OneVision-2	0.00	1	92.52
InternVL3.5-8B	0.00	3	94.10

Model	Acc. (%)	Modes	Mass (%)
Qwen3-VL-8B	0.23	2	97.51
Qwen2.5-VL-SAT	0.00	2	100.00
SpatialLadder-3B	0.00	2	100.00
VST-7B-RL	0.00	2	100.00

Table 1: Image-plane facing-direction predictions for eight open VLMs on 4COMFORT. A dominant mode is defined operationally as a direction receiving at least 10% of predictions; mass is the summed probability of these modes. All examples have the same ground-truth direction, down-left, and all models achieve 100% parse success.

source models (Tab. 1) cannot consistently predict these orientations, even in the image-plane context. All models default to 1-3 options for all samples, with none of the modes falling on the correct option. InternVideo3 is an exception, but it still shows dual-modal behavior, as shown in Tab. 1. Supplying ground-truth orientation does not measurably help object selection in Fig. 3b. A front arrow provides limited assistance, and supplying front and left together fails help as well despite the information is sufficient for a complete coordinate system. This suggests difficulty extending partial orientation information into a usable complete frame.

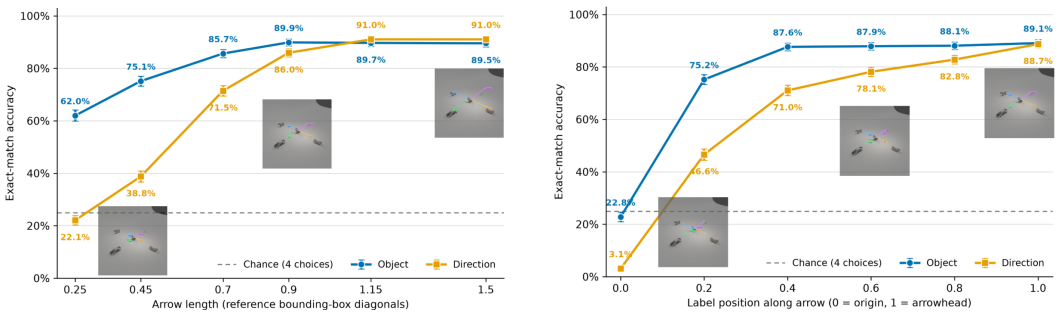
3.3 Explicit Frames Make Relations Visually Accessible

Displaying the complete frame. Replacing the partial cue with four labeled arrows substantially improves object selection (Fig 3b) across every queried relation. This distinguishes constructing a frame from using a suitable display of one. A complete overlay removes the need to recover the opposite axes and supplies a labeled visual path toward candidate regions. The model can consequently use its localization capabilities to associate a requested direction with an object. We continue to see the affinity of the models in predicting object-answer prompts here.

How is the VLM using the overlay? We further vary arrow length while preserving the directional labels and reference frame (Fig. 4a). Accuracy rises from 62.00% at scale 0.25 to 89.90% at 0.90, then approximately plateaus. Over the same interval, the mean distance from the queried arrow's endpoint to the correct object's box decreases from 9.46% to 1.59% of the image diagonal. At scale 1.50, some endpoints overshoot the object regions and the mean gap increases, while accuracy remains near its peak. This experiment tells us that the model's performance is tied to how easy we make the image-plane localization task of connecting the overlay direction to the target objects. Fig. 4b shows that more clearly as the same pattern is observed when the distance between the target objects and the labels are varied.

4 Method: FrameBind

FrameBind is a training-free pipeline that overlays a reference-centered coordinate frame on the original RGB image and optionally converts direction prediction into object-map retrieval (Fig. 5). Given an image and a question naming a reference r and target t, the task is to predict their relation



(a) Arrow-length (b) Label-position

Figure 4: (a)As the arrow length scales, the tip and the label get closer to the centroid of the correct object for each direction. (b) The position of the label along an arrow of fixed length also moves closer to the correct object for the label. Both graphs show that the model is exploiting image-plane localization.

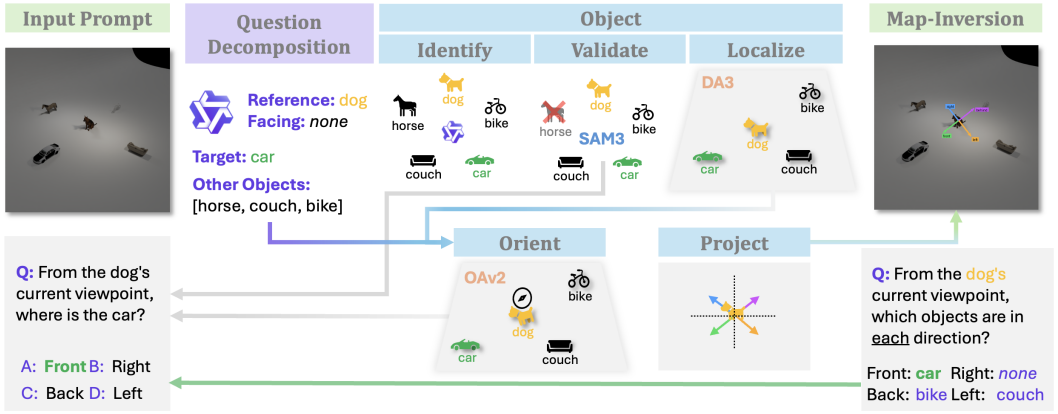


Figure 5: FrameBind grounds scene entities, constructs the requested reference frame, and projects labeled axes onto the original image. Optional object-map inversion recovers the target's direction from its unique map assignment. Gray paths are direct-answer fallbacks used when intermediate outputs are unusable.

in $\mathcal{D} = \{\text{left}, \text{right}, \text{front}, \text{behind}\}$, or, given r and $d \in \mathcal{D}$, to predict t. The original image is retained throughout. Appendix B gives full construction, rendering, and fallback details.

Grounding and frame construction. Qwen3-VL extracts the reference, target, candidate entities, and any explicit viewpoint instruction from the question. Then SAM3 (Carion et al., 2025) grounds them with masks and boxes (the VLM proposes additional visible objects for SAM3 verification when the candidate set is insufficient). Missing distractors are tolerated, but failure to ground the reference prevents overlay construction. For the next step, Depth Anything 3 (Lin et al., 2025) supplies depth and intrinsics for the reference center and a supporting-plane normal $\mathbf{n}$, fitted by RANSAC with SVD refinement to lower-image depth points outside the reference mask.

OrientAnything-v2 (Wang et al., 2026) is used to get a camera-frame forward estimate $\mathbf{f}_0$ when intrinsic orientation is required (when given, explicit viewpoint instructions take precedence). The forward axis is $\mathbf{f}_0$ projected onto the plane and normalized, the right axis is given by $\mathbf{f} \times \mathbf{n}$, and back and left are their negatives. Axis endpoints are projected into the image, and usable non-end-on arrows keep their projected directions while being scaled to the reference-box diagonal, so the overlay adds the requested frame without removing RGB evidence for localization.

Object-map prediction and inversion. To exploit the object-answer advantage, a single VLM call receives the overlay, reference identity, candidate set $\mathcal{C}$, and target, and returns a sparse map

$M : \mathcal{D} \to \mathcal{C} \cup \{\perp\}$, where $\perp$ marks an unassigned direction. Each concrete candidate may appear at most once and the target exactly once; directions may be empty, and when several objects share a direction the map records one representative. The decoder checks that all four keys are present, each value is a listed candidate or $\perp$, and the target occurs exactly once; an invalid map receives one repair request naming the violated constraints. These checks establish structural validity, not spatial correctness. The final direction is

$$\hat{d} = \begin{cases} \text{unique}\{d \in \mathcal{D} : M(d) = t\}, & \text{map accepted,} \\ \hat{d}_{\text{direct}}, & \text{otherwise,} \end{cases} \tag{1}$$

where $\hat{d}_{\text{direct}}$ answers the original question from the same overlay, or from the original image when no usable overlay exists. Accepted maps are inverted without consulting the direct answer or gold labels, and native object-answer questions are never inputs to map inference, so the inverse question cannot reveal the requested direction. Direct answering and inversion are reported separately; map-only accuracy, valid-map coverage, and fallback counts are given in Appendix B.4.

5 Evaluation

5.1 Datasets

We evaluate paired object-answer and direction-answer questions on four settings (details in Appendix A). **4COMFORT** is a controlled COMFORT-derived collection (Zhang et al., 2025b) with a clearly oriented reference, four targets in its principal directions, and a fixed reference–camera relationship, suited to diagnosing frame errors with little visual ambiguity. **COMFORTer** keeps the four-target structure but varies object appearance, target orientation, and camera viewpoint, testing generalization beyond 4COMFORT's regular layouts. **Kubric MOVi-A** (Greff et al., 2022) provides varied arrangements and visibility; we use the subset where the question stipulates the reference as facing the camera, so it tests applying a specified viewpoint rather than recovering intrinsic orientation. **3DSR** (Ma et al., 2025) extends evaluation to natural-image object-relative questions; we pair native direction questions with constructed object-answer counterparts whose distractors mix same-image annotations and Qwen-generated names, so its object-answer results are exploratory.

5.2 Baselines

We compare against three groups. *General-purpose models* LLaVA-OneVision2 (An et al., 2026), InternVL3.5-8B (Wang et al., 2025b), InternVideo3-8B (Yan et al., 2026), and Qwen3-VL-8B-Instruct (Bai et al., 2025) test whether answer-formulation sensitivity spans model families; Qwen3-VL is also FrameBind's backbone and hence the primary comparison for inference-time assistance, and GPT-5.6-Luna (OpenAI, 2026) is a proprietary reference point. *Spatially specialized models* Qwen2.5-VL-SAT (Ray et al., 2025), SpatialLadder-3B (Li et al., 2025a), VST-7B-RL (Yang et al., 2025b), and Cambrian-S-7B (Yang et al., 2025c) test whether specialization alone yields reliable reference-relative answers. *Tool-assisted* APC-Vis (Lee et al., 2025) answers from a generated abstract reference-view image, contrasting viewpoint reformulation with FrameBind's original-image overlay. SymPL's symbolic representation (Jang et al., 2026) does not natively support general relative-direction queries; Sec. 6 justifies not adopting a similar abstraction.

5.3 Protocol and Results

We report four-choice accuracy separately for object and direction answers (chance is 25%). The FrameBind row gives direct answers with estimated overlays; the map-inversion row gives directions from accepted maps with direct-answer fallback. Unavailable predictions count as incorrect. These end-to-end results are distinct from the ground-truth-cue diagnostics of Sec. 3.

Answer-formulation sensitivity. Every unassisted open model is more accurate at object than direction answering on Kubric and 4COMFORT (Table 2); for Qwen3-VL the gaps are 26.6 and 23.3 points. On COMFORTer and 3DSR the gaps shrink to a few points and some models favor directions, so the asymmetry is pronounced on controlled scenes rather than universal. Because the paired formulations change both the supplied information and the candidate domain, these gaps measure query-formulation sensitivity rather than an isolated output-format effect.

Group	Method	COMFORTer		Kubric		4COMFORT		3DSR	
		Obj	Dir	Obj	Dir	Obj	Dir	Obj	Dir
Random	Random	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Proprietary	GPT-5.6-Luna	56.15	50.95	55.50	33.03	33.18	31.68	38.66	41.18
General Purpose	LLaVA-OneVision2	25.20	24.65	35.32	30.73	25.16	2.47	28.57	22.69
	InternVL3.5-8B	26.00	25.50	46.33	30.28	18.94	11.85	<u>36.13</u>	19.33
	InternVideo3-8B	26.15	25.00	39.91	**36.70**	<u>30.74</u>	26.46	34.45	22.69
	Qwen3-VL-8B-Instruct	<u>27.55</u>	24.15	<u>46.79</u>	20.18	26.87	3.59	32.77	30.25
Spatial Specialized	Qwen2.5-VL-SAT	23.40	24.25	37.16	31.19	22.72	15.46	30.25	28.57
	SpatialLadder-3B	25.00	24.55	38.07	12.84	22.18	16.66	25.21	32.77
	VST-7B-RL	25.65	24.05	41.28	22.94	21.88	2.26	<u>36.13</u>	31.09
	Cambrian-S-7B	25.25	24.50	32.57	32.11	18.15	11.26	31.93	31.09
	APC-Vis*	23.25	24.50	31.40	27.48	25.39	21.17	27.73	23.52
Ours	**FrameBind**	**67.75**	<u>59.90</u>	**51.83**	33.03	**66.71**	<u>30.38</u>	**38.66**	<u>36.13</u>
	+ Map Inversion	–	**67.75**	–	<u>35.32</u>	–	**52.28**	–	**36.97****

Table 2: Object-answer (Obj) and direction-answer (Dir) accuracy (%) on paired reference-relative tasks. **Bold** and <u>underlined</u> values indicate the highest and second-highest distinct scores in each column, excluding GPT-5.6-Luna. *APC-Vis evaluates questions using a generated reference-view image. **The 3DSR object-answer evaluation uses constructed distractors whose generation overlaps with the FrameBind pipeline.

Reference-frame assistance. FrameBind improves both formats over its Qwen3-VL backbone on all four datasets, most strongly on COMFORTer (27.6→67.8 object, 24.2→59.9 direction) and 4COMFORT (26.9→66.7 and 3.6→30.4), with smaller gains on Kubric and 3DSR. The larger object gains show that supplying a frame does not by itself close the object–direction gap. FrameBind attains the highest object accuracy among open and tool-assisted systems on every dataset, but not every direction task: InternVideo3 leads Kubric direction answering, and GPT-5.6-Luna remains stronger on 3DSR directions.

Map inversion. Inversion adds 7.9 direction points on COMFORTer and 21.9 on 4COMFORT, but only 2.3 on Kubric and 0.8 on 3DSR, so object-mediated inference helps but its benefit is strongly dataset-dependent. The final score combines accepted-map predictions and fallbacks; it does not show that predicted maps are globally correct, and the main comparison does not isolate ground-plane alignment from other components. All FrameBind results use a single backbone, Qwen3-VL-8B-Instruct.

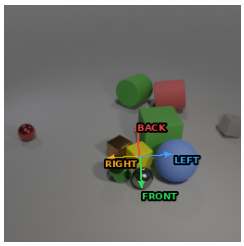
6 Analysis: RGB Overlays and Symbolic Representations

The representation supplied to a VLM determines which perception errors can remove or distort its spatial evidence. FrameBind keeps scene appearance and adds a coordinate overlay; a symbolic map instead represents candidates as explicit entity symbols. We compare the two on 218 matched Kubric object-answer queries with Qwen3-VL-8B-Instruct (Fig. 6). Both keep the source question, options, and reference identity; the RGB input adds four labeled axes, while the symbolic input replaces pixels with entity symbols and a text legend, laid out with the reference heading pointing image-up. This compares the two visual representations, not the full FrameBind and SymPL pipelines.

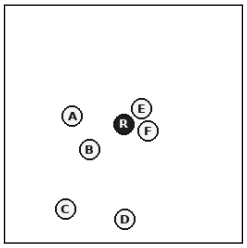
Candidate evidence. On clean inputs, RGB axes reach 55.5% versus 49.5% for the symbolic map (the clean direction-answer comparison favors the symbolic map, so this is an object-selection result). Removing non-reference symbols or perturbing their centroids affects only the symbolic input: symbolic accuracy falls to 23.9% at 75% dropout and 31.2% at the highest centroid noise, while RGB is unchanged at 55.5%. The widening gap is the cost of making candidate evidence depend on detected symbols and estimated positions; it does not show that RGB inference repairs corrupted estimates.

Reference location and heading. Reference-centroid noise and heading rotation corrupt the scaffold used by both representations. Under reference noise, RGB stays ahead at every level but drops

Intervention	Level	RGB axes	Symbolic-up	Δ
Clean	–	55.50	49.54	+5.96
Detection dropout	25%	55.50	40.37	+15.14
	50%	55.50	33.03	+22.48
	75%	55.50	23.85	+31.65
Centroid noise	10%	55.50	47.25	+8.26
	25%	55.50	38.53	+16.97
	50%	55.50	34.86	+20.64
	100%	55.50	31.19	+24.31
Reference-centroid noise	10%	55.05	51.38	+3.67
	25%	53.21	46.79	+6.42
	50%	47.25	42.66	+4.59
	100%	44.50	38.53	+5.96
Heading rotation	15°	54.13	47.25	+6.88
	30°	49.54	45.41	+4.13
	45°	44.50	43.58	+0.92
	90°	32.57	34.86	−2.29
	180°	25.69	27.52	−1.83



RGB Overlay



Symbolic

Figure 6: Object-answer accuracy (%) on 218 matched Kubric queries per condition. Δ is RGB minus symbolic accuracy in percentage points, calculated before rounding. Perturbation suites are independent. Detection dropout removes non-reference symbols, and candidate-centroid noise perturbs non-reference positions; neither changes the RGB input. Reference-centroid noise perturbs the reference location used by both representations. Centroid perturbations use isotropic Gaussian noise, scaled by reference-to-object distance for candidate noise and camera-to-reference distance for reference noise. Heading corruption rotates the RGB axes or changes the symbolic heading-up alignment. All conditions retain the original answer labels.

to 44.5% at the highest, so retaining appearance does not remove dependence on a well-placed frame. Under heading rotation the RGB advantage narrows to 0.9 points at 45° and reverses at 90° and 180°, where both are near chance; because original labels are retained, this measures sensitivity to frame corruption rather than correctness within the corrupted frame.

Scope. These results support retaining RGB for candidate appearance and location while supplying the frame externally, which reduces reliance on candidate detections but leaves reference localization and heading as the main dependencies. The comparison uses one VLM, one subset, and one deterministic corruption per condition, without paired uncertainty estimates, and supports a representation choice for this object-selection setting rather than general superiority of RGB over symbolic reasoning.

7 Conclusion

We show that VLMs' object-answer advantage on reference-relative tasks reflects camera-frame substitution rather than a recovered allocentric frame, and that models localize reliably but cannot recover or extend orientation cues into a usable frame. FrameBind supplies that frame at inference time as a coordinate overlay on the original RGB image and improves both answer formats over its Qwen3-VL backbone on four datasets, with object-map inversion adding further direction gains on controlled scenes. Retaining RGB makes the pipeline robust to candidate-detection errors, but reference location and heading remain consequential. The main limitations are a single answering backbone, controlled synthetic layouts, constructed natural-image distractors, and dependence on estimated geometry; evaluating additional backbones and isolating the contributions of frame estimation, representation, and map inference are the natural next steps.

AI USE STATEMENT

Generative AI tools were used to assist with phrasing and revising sections of the manuscript, reviewing intermediate drafts, rapidly prototyping code, and providing preliminary cross-checks of related-work coverage and descriptions. These uses supported writing and development; AI suggestions were used complementarily with independent verification of citations, implementations, and experimental conclusions. The appendix was collated with the use of coding agents to ensure conformity with the actual implementation.

ETHICS STATEMENT

This work raises no ethical concerns that we are aware of. All evaluation data are derived from publicly available sources: COMFORT, Objaverse-XL, Kubric MOVi-A, and 3DSRBench, under their respective licenses. The 4COMFORT and COMFORTer collections are synthetic renders of object assets, and no human subjects, personal data, or personally identifiable information are involved at any stage of data construction, model evaluation, or analysis. All evaluated models are used for inference only, with no additional training. FrameBind is an inference-time diagnostic and assistance method for spatial reasoning and does not introduce capabilities that we consider to carry a foreseeable risk of misuse beyond those of the underlying vision-language models.

REPRODUCIBILITY STATEMENT

We have taken the following steps to make our results reproducible. Appendix A specifies the construction of every evaluation set: scene geometry, asset inventories, camera sampling distributions, rendering settings, and ambiguity filtering for 4COMFORT and COMFORTer (App. A.1); frame sampling, grounding thresholds, coordinate-frame definitions, and relation-retention criteria for Kubric MOVi-A (App. A.2); and the candidate-proposal and option-construction procedure for the 3DSR diagnostic (App. A.3). Appendix A.4 states the evaluation units and the exact subsets used for each component diagnostic. Appendix B documents the FrameBind pipeline in full, including the depth and camera conventions, the ground-plane estimation procedure and its thresholds, the frame-construction and projection equations, the object-map constraints and validation rules. Appendix C provides the prompt templates and diagnostic conditions for all paired questions, visual-assistance variants, localization and orientation probes, and representation perturbations. All experiments use deterministic decoding, and the evaluated models and vision tools are publicly available.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. Llava-onevision-2: Towards next-generation perceptual intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Christopher Beckham, Martin Weiss, Florian Golemo, Sina Honari, Derek Nowrouzezahrai, and Christopher Pal. Visual question answering from another perspective: Clevr mental rotation tests, 2022. URL https://arxiv.org/abs/2212.01639.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models, 2025. URL https://arxiv.org/abs/2406.13642.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

An-Chieh Cheng, Yang Fu, Yatai Ji, Ligeng Zhu, Guanqi Zhan, Zhuoyang Zhang, Zhaojing Yang, Song Han, Yao Lu, Pavlo Molchanov, Vidya Nariyambut Murali, Jan Kautz, Xiaolong Wang, Hongxu Yin, and Sifei Liu. Grounded 3d-aware spatial vision-language modeling, 2026. URL https://arxiv.org/abs/2605.30307.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. Mm-spatial: Exploring 3d spatial understanding in multimodal llms, 2025. URL https://arxiv.org/abs/2503.13111.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects, 2023. URL https://arxiv.org/abs/2307.05663.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models, 2024. URL https://arxiv.org/abs/2406.05756.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti, Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator, 2022. URL https://arxiv.org/abs/2203.03570.

Gracjan Góral, Alicja Ziarko, Michal Nauman, and Maciej Wołczyk. Seeing through their eyes: Evaluating visual perspective taking in vision language models, 2024. URL https://arxiv.org/abs/2409.12969.

Jaeyun Jang, Seunghui Shin, Taeho Park, and Hyoseok Hwang. Keep it sympl: Symbolic projective layout for allocentric spatial reasoning in vision-language models, 2026. URL https://arxiv.org/abs/2602.19117.

Fei Kong, Jinhao Duan, Kaidi Xu, Zhenhua Guo, Xiaofeng Zhu, and Xiaoshuang Shi. Lrr-bench: Left, right or rotate? vision-language models still struggle with spatial understanding tasks, 2026. URL https://arxiv.org/abs/2507.20174.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation, 2025. URL https://arxiv.org/abs/2504.17207.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners, 2024. URL https://arxiv.org/abs/2406.02537.

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, 2025a. URL https://arxiv.org/abs/2510.08531.

Pengteng Li, Pinhao Song, Wuyang Li, Weiyu Guo, Huizai Yao, Yijie Xu, Dugang Liu, and Hui Xiong. See&trek: Training-free spatial prompting for multimodal large language model, 2025b. URL https://arxiv.org/abs/2509.16087.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025. URL https://arxiv.org/abs/2511.10647.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-ttt: Streaming visual-based spatial intelligence with test-time training, 2026. URL https://arxiv.org/abs/2603.12255.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *ArXiv*, abs/2205.00363, 2022.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors, 2024. URL https://arxiv.org/abs/2403.13438.

Weijian Ma, Shizhao Sun, Tianyu Yu, Ruiyu Wang, Tat-Seng Chua, and Jiang Bian. Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation, 2026a. URL https://arxiv.org/abs/2601.01984.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark, 2025. URL https://arxiv.org/abs/2412.07825.

Xueqi Ma, Shuo Yang, Yanbei Jiang, Shu Liu, Zhenzhen Liu, Jiayang Ao, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. Attention in space: Functional roles of vlm heads for spatial reasoning, 2026b. URL https://arxiv.org/abs/2603.20662.

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, 2026. URL https://arxiv.org/abs/2605.30161.

MirroS Team. S-space: Exploring spatial workspace in multimodal models, September 2026. URL https://mirros.ai/blog/s-space. Blog post.

OpenAI. GPT-5.6 Large Language Model. Computer software, 2026. URL https://chatgpt.com. Accessed: 2026-09-26.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. URL https://arxiv.org/abs/2412.07755.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024. URL https://arxiv.org/abs/2403.16999.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness, 2025a. URL https://arxiv.org/abs/2504.01901.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Zehan Wang, Ziang Zhang, Jiayang Xu, Jialei Wang, Tianyu Pang, Chao Du, HengShuang Zhao, and Zhou Zhao. Orient anything v2: Unifying orientation and rotation understanding, 2026. URL https://arxiv.org/abs/2601.05573.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence, 2026. URL https://arxiv.org/abs/2505.23747.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. Internvideo3: Agentify foundation models with multimodal contextual reasoning, 2026. URL https://arxiv.org/abs/2606.12195.

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, 2025a. URL https://arxiv.org/abs/2412.14171.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning, 2025b. URL https://arxiv.org/abs/2511.05491.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video, 2025c. URL https://arxiv.org/abs/2511.04670.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence, 2026. URL https://arxiv.org/abs/2505.23764.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning, 2025d. URL https://arxiv.org/abs/2507.12508.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data, 2025a. URL https://arxiv.org/abs/2506.03642.

Xuejun Zhang, Aditi Tiwari, Zhenhailong Wang, and Heng Ji. Predicting camera pose from perspective descriptions for spatial reasoning, 2026. URL https://arxiv.org/abs/2602.06041.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities, 2025b. URL https://arxiv.org/abs/2410.17385.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors, 2025. URL https://arxiv.org/abs/2505.24625.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4d: Dynamic spatial understanding for vision language models, 2025. URL https://arxiv.org/abs/2512.20557.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness, 2025a. URL https://arxiv.org/abs/2409.18125.

Chenming Zhu, Jingli Lin, Yilin Long, Peizhou Cao, Tai Wang, Jiangmiao Pang, and Xihui Liu. Thinking with imagination: Agentic visual spatial reasoning with world simulators, 2026. URL https://arxiv.org/abs/2606.06476.

Fangrui Zhu, Hanhui Wang, Yiming Xie, Jing Gu, Tianye Ding, Jianwei Yang, and Huaizu Jiang. Struct2d: A perception-guided framework for spatial reasoning in mllms, 2025b. URL https://arxiv.org/abs/2506.04220.

A Dataset Construction and Evaluation Subsets

We evaluate paired direction-answer and object-answer questions on two COMFORT-derived collections, Kubric MOVi-A, and 3DSRBench. Each pair describes the same spatial relation but changes the supplied information and answer choices. All questions have four options. Table 3 distinguishes images, relations, and prompt instances.

Dataset / subset	Images	Relations	Direction	Object
4COMFORT, filtered	395	1,580	6,320	6,320
COMFORTer	500	2,000	2,000	2,000
MOVi-A, both frames	58	437	437	437
Camera-relative	–	219	219	219
Object-facing-camera	–	218	218	218
3DSR diagnostic	32	119	119	119

Table 3: Evaluation sizes. Each 4COMFORT relation has four option-order variants per formulation. MOVi-A images span 29 videos; the main evaluation uses the object-facing-camera subset. The 3DSR image count denotes original source images, some reused through horizontal reflection.

A.1 4COMFORT and COMFORTer

Both collections extend COMFORT (Zhang et al., 2025b), reusing its base scene and original assets. Each initially contains 500 rendered scenes.

Scene geometry and relation labels. A scene contains a reference at the horizontal world origin and four distinct target assets sampled without replacement from the remaining library. Distinct asset identities need not have distinct natural-language names. The canonical reference basis is right along $+X$, front along $+Y$, and up along $+Z$. For $d \sim \mathcal{U}[5.0, 5.8]$, target positions are

$$\begin{aligned} \mathbf{p}_{\text{left}} &= (-d, 0, 0), \qquad \mathbf{p}_{\text{right}} = (d, 0, 0), \\ \mathbf{p}_{\text{front}} &= (0, d, 0), \qquad \mathbf{p}_{\text{behind}} = (0, -d, 0). \end{aligned} \tag{2}$$

Reference yaw rotates the arrangement about world $+Z$. Labels follow the reference orientation rather than image-plane position.

Assets are scaled using post-rotation horizontal bounds. Nominal reference and target widths are 2.2 and 1.7 scene units, respectively, with asset-specific multipliers. Floor clearances are 0.02 units in 4COMFORT and 0.04 in COMFORTer. Every scene supplies four relations, each queried once as direction selection and once as object selection. Direction choices are left, right, front, and behind; object choices are the four target names.

4COMFORT sampling and ambiguity filtering. The reference is sampled uniformly from six eligible assets, with yaw in $\{0°, 90°, 180°, 270°\}$. Targets share its yaw. The camera position $(-10, 12, 10)$ rotates with the reference yaw and points toward $(0, 0, 1)$. Thus world-yaw variation preserves the reference's camera-relative heading. This coupling supports diagnosis of alternative frame interpretations but does not test broad viewpoint generalization.

We remove 105 scenes containing repeated object instances that make names ambiguous, retaining 395. All questions belonging to an excluded scene are removed. Each retained question has four cyclic option orders, placing the correct answer once at each option index. The filtered collection contains 1,580 relations and 12,640 prompts, with 3,160 correct answers at each option index across both formulations. This filtering targets naming ambiguity rather than all possible visual ambiguities.

COMFORTer viewpoint and appearance variation. COMFORTer adds 20 manually orientation-calibrated Objaverse-XL assets (Deitke et al., 2023), yielding 32 assets and 16 eligible references. Reference yaw uses the same four cardinal angles, while targets receive independent yaws. Camera sampling follows

$$\begin{aligned} \phi_{\text{cam}} &= \theta_r + \alpha + \epsilon, & \alpha &\in \{45°, 135°, 225°, 315°\}, \\ \epsilon &\sim \mathcal{U}[-15°, 15°], & \rho_{\text{cam}} &\sim \mathcal{U}[14.5, 17.5], \\ z_{\text{cam}} &\sim \mathcal{U}[8.0, 11.5], & h &\sim \mathcal{U}[0.8, 1.8]. \end{aligned} \tag{3}$$

The camera looks toward $(0, 0, h)$. This varies camera azimuth relative to the reference and decouples target orientation from target relation.

Each question receives one shuffled option order, producing 4,000 prompts. Correct option-index counts are 960, 989, 1,005, and 1,046.

Asset group	Models
Original references (6)	bicycle mountain, car sedan, dog, duck, HorseL, HorseR
Original target-only (6)	bed, bench, chair, laptop, sofa, sophia
Added references (10)	tank, angler fish, fox, teddy bear, plane, chicken, rescue helicopter, giraffe, lion, bear
Added target-only (10)	red sofa, office chair, dresser, boat, table fan, fire extinguisher, piano, basketball, lamp, alarm clock

Table 4: Asset inventory. 4COMFORT uses the original 12 assets; COMFORTer uses all 32. Reference-capable assets may also appear as targets.

Calibration and rendering. Each asset's forward direction is aligned with canonical $+Y$ before semantic yaw is applied. Asset-specific rotations correct source poses. Added assets are inspected through four-view calibration images and example scenes; only assets with a visually meaningful calibrated front serve as references.

Images are rendered at 512×512 using Blender 3.6.0, Cycles with 64 samples and denoising, a 35 mm focal length, and a 32 mm sensor width. The scene contains a floor and three lights. Rendering uses the Standard color transform, zero exposure, gamma 1.0, and an opaque background.

COMFORTer additionally rejects globally invalid images. On RGB images downsampled to at most 128×128, mean intensity must be at least 5.0, mean per-channel standard deviation at least 2.0, and fewer than 95% of grayscale pixels may have intensity below 8 on a 0–255 scale. Failure triggers camera resampling with scene content fixed, for at most eight attempts. Of the 500 accepted scenes, 448 pass on the first attempt, 44 on the second, and eight on the third.

Annotations and limits. Annotations include asset identities, reference yaw, relations, projected boxes, and camera-frame poses. Boxes are obtained by projecting mesh bounding-box corners and clipping to the image; they are amodal bounds rather than visible segmentation boxes. Coordinates are normalized to $[0, 1]$ with a top-left origin. All 2,500 boxes in the original 4COMFORT collection are non-degenerate and avoid the clipped boundary; three COMFORTer boxes touch it. Neither collection applies complete per-object occlusion, visibility-fraction, or collision filtering. Both are evaluation collections without separate training and validation partitions.

A.2 KUBRIC MOVI-A

We derive paired questions from the 256×256 MOVi-A validation data (Greff et al., 2022). The source comprises 250 videos of 24 frames each, with three to ten objects varying in shape, size, color, and material. The diagnostic retains 437 relations across 58 frames from 29 videos: 219 camera-relative and 218 object-facing-camera relations.

Sampling and grounding. We sample at most four frames per video from indices 15–23. Eligible objects require a frame-specific box, at least 300 visible segmentation pixels, and normalized box area of at least 0.01. Frames require five eligible objects. Sampling is capped at four relations per coordinate frame per image and eight per coordinate frame per video.

Descriptions follow size, color, material, and shape. Reference grounding is checked with SAM3 (Carion et al., 2025) using a shorter color–shape description. Detections require confidence at least 0.25; the selected instance must have box IoU at least 0.50 and mask IoU at least 0.25 against simulator annotations. Selection prioritizes box IoU, then mask IoU, then confidence. Only references receive this grounding check; spatial labels remain simulator-derived.

Coordinate frames. Let $\boldsymbol{\delta} = \mathbf{p}_t - \mathbf{p}_r$. If R is the camera rotation derived from its scalar-first quaternion, with columns $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$, then

$$\mathbf{r}_{\text{cam}} = \mathbf{R}_1, \qquad \mathbf{u}_{\text{cam}} = \mathbf{R}_2, \qquad \mathbf{f}_{\text{cam}} = -\mathbf{R}_3. \tag{4}$$

Camera-forward points into the scene. For the object-facing-camera frame, camera position $\mathbf{c}$ defines

$$\mathbf{f}_{\text{obj}} = \frac{(c_x - p_{r,x}, c_y - p_{r,y}, 0)}{\|(c_x - p_{r,x}, c_y - p_{r,y}, 0)\|_2}, \qquad \mathbf{r}_{\text{obj}} = (f_{\text{obj},y}, -f_{\text{obj},x}, 0). \tag{5}$$

Horizontal reference–camera coincidence is rejected. This stipulated frame does not use intrinsic orientation, which is important for symmetric MOVi-A objects.

For either basis, set $x = \boldsymbol{\delta}^\top \mathbf{r}$ and $y = \boldsymbol{\delta}^\top \mathbf{f}$. The signed dominant component determines the relation. We retain only relations satisfying

$$\max(|x|, |y|) \geq 0.35, \qquad \big||x| - |y|\big| \geq 0.10. \tag{6}$$

Vertical displacement does not determine the four-way label.

Choices and ambiguity. Direction options are fixed as A=left, B=right, C=front, and D=behind. Object choices contain the target and three distractors, with distinct colors different from the reference color. Every distractor has a relation different from the queried relation; object-option order is shuffled.

These restrictions select for visibility, groundability, and distinguishable candidates, so performance does not estimate accuracy on unfiltered MOVi-A. Full descriptions are not required to be unique across the entire image: 16 relations have duplicated reference descriptions and 15 have duplicated target descriptions. These 31 potentially ambiguous relations belong to the combined 437-relation collection, not necessarily the 218-relation main subset.

Subset	Left	Right	Front	Behind	Total
MOVi-A camera-relative	49	64	51	55	219
MOVi-A object-facing-camera	54	50	54	60	218
3DSR	34	33	30	22	119

Table 5: Counts of paired relations by direction. 3DSR uses the term back instead of behind.

A.3 3DSRBENCH

The source 3DSRBench collection (Ma et al., 2025) contains 2,772 (5,157 augmented) questions, including 343 multi-object questions about a viewpoint toward another object. We preserve native direction questions and construct inverse object questions about the same relation. Each retained

relation requires at least one distractor from other queried objects in the same image. This and the remaining construction requirements exclude 224 candidates, leaving 119 paired relations across 32 source images. There are 61 unflipped and 58 horizontally flipped relations; the corresponding image transformation is preserved.

Image-conditioned candidate proposals. Qwen3-VL-8B-Instruct proposes concise visible-object names from foreground, middle ground, and background. Known annotated names are excluded. Instructions prohibit colors, materials, directions, and spatial relations, while permitting structures, fixtures, vegetation, and surfaces. Generation receives the image and known-object exclusions, not the requested inverse relation or its answer.

Proposals cover 143 sparse images identified before relation filtering, producing 1,819 retained labels. Each image has 8–64 labels, with mean 12.72 and median 12. Generation uses deterministic decoding and at most three attempts per image. Eight truncated responses are recovered by retaining complete label strings and discarding incomplete endings.

Option construction. Choices are assembled from the annotated target, distinct same-image queried terms, and then generated proposals until four options are available. The reference and aliases of existing choices are excluded. Names are lowercased and whitespace-normalized; equality or substring overlap is treated as duplication. Option order is deterministically shuffled per relation. Insufficient distinct proposals cause construction failure; the image-conditioned variant has no thematic fallback.

Of 357 distractor slots, 146 contain same-image queried terms and 211 contain Qwen proposals. Four questions need no proposals, 19 need one, and 96 need two. Thus 115 of 119 inverse questions contain generated distractors. All inverse questions have four normalized distinct choices.

Interpretation. Generated proposals are not human-verified or detector-confirmed labels. They may contain hallucinations, synonyms, or inconsistent granularity. Using Qwen3-VL for candidate generation may also favor vocabulary or visual distinctions shared by Qwen-family evaluators. The resulting diagnostic should be distinguished from the original thematic-padding variant. Its object-answer accuracy measures performance under this candidate-construction procedure, rather than an isolated answer-format effect.

A.4 Evaluation Units and Diagnostic Subsets

Matched relations are the units for comparing answer formulations. Relations may share an image, MOVi-A images may share a video, and 4COMFORT includes repeated option orders. Uncertainty estimates should preserve paired formulations and account for image-level dependence, or video-level dependence for MOVi-A.

Component diagnostics can use different eligibility rules. The unboxed facing-direction probe uses 441 images selected for unique normalized reference names from the original 4COMFORT collection. All have the same projected ground-truth heading, down-left. The single-axis utilization diagnostic contains 1,849 valid reference–target queries after excluding ambiguous normalized target labels. These are diagnostic subsets, not additional datasets, and their denominators should not be confused with the 395-scene filtered paired evaluation.

B FrameBind Technical Details

FrameBind preserves the original RGB image while adding a reference-centered coordinate scaffold. Direct answering uses this overlay with the authored question; object-map inference additionally predicts a sparse assignment of candidates to directions.

B.1 Grounding and Reference Geometry

Question decomposition identifies the reference, target, candidate entities, and requested viewpoint. Available entity annotations take precedence over generated names. SAM3 provides masks and

boxes. The reference mask supports center estimation and exclusion of its surface from plane fitting. Additional VLM-proposed candidates are retained only after successful grounding. Missing candidate detections need not prevent inference, but reference grounding is required for the overlay.

Intrinsic heading is estimated by OrientAnything-v2. Explicit viewpoint instructions override intrinsic orientation: a stipulated camera-facing reference uses that heading, while camera-relative questions retain the camera frame.

Depth and camera coordinates. Depth Anything 3 supplies depth and intrinsics. Coordinates follow $[x_{\text{right}}, y_{\text{forward}}, z_{\text{up}}]$. For forward depth $d(u, v)$,

$$\mathbf{p}(u, v) = \begin{bmatrix} (u - c_x)d(u, v)/f_x \\ d(u, v) \\ -(v - c_y)d(u, v)/f_y \end{bmatrix}. \tag{7}$$

The final sign converts downward-increasing image rows to upward-positive camera coordinates. Valid masked points provide the reference center $\mathbf{c}_r$ and candidate centers $\mathbf{c}_i$. These summarize observed surfaces and inherit depth errors; they need not equal complete-object centers.

B.2 Supporting Plane and Object Frame

Plane proposals come from finite positive depths in the lower 65% of the image, excluding the reference mask. At most 12,000 points are retained by deterministic subsampling. RANSAC fits $\mathbf{n}^\top \mathbf{p} + \beta = 0$ with $\|\mathbf{n}\|_2 = 1$ over 256 trials, using threshold

$$\tau = \max\left(10^{-4}, 0.015\,\text{median}_{\mathbf{p}\in\mathcal{P}}\,\|\mathbf{p}\|_2\right). \tag{8}$$

The best inlier set is centered and refined by SVD; the smallest-singular-value direction gives its normal. The normal is flipped if necessary to agree with the camera-up hint $\mathbf{u}_{\text{cam}} = (0, 0, 1)^\top$. Fewer than 50 usable points, invalid geometry, or a failed fit trigger an explicit camera-up fallback.

This procedure estimates a supporting plane heuristically. Other objects and non-floor surfaces can remain among the proposals. Normal-sign selection assumes physical up lies in the same hemisphere as camera-up. The plane offset does not set the overlay origin.

Horizontal frame. Let $\mathbf{f}_0$ be the valid forward estimate in camera coordinates and $\mathbf{n}$ the selected vertical. We construct

$$\begin{aligned} \tilde{\mathbf{f}} &= \mathbf{f}_0 - (\mathbf{f}_0^\top \mathbf{n})\mathbf{n}, \quad \mathbf{f} = \frac{\tilde{\mathbf{f}}}{\|\tilde{\mathbf{f}}\|_2}, \\ \mathbf{r} &= \frac{\mathbf{f} \times \mathbf{n}}{\|\mathbf{f} \times \mathbf{n}\|_2}, \qquad \mathbf{b} = -\mathbf{f}, \quad \mathbf{l} = -\mathbf{r}, \quad \mathbf{u} = \mathbf{n}. \end{aligned} \tag{9}$$

The basis $B = [\mathbf{r}\ \mathbf{f}\ \mathbf{u}]$ is checked for unit norms, orthogonality, and right-handedness. Degenerate forward estimates parallel to the normal are rejected. Ground alignment removes vertical components but does not repair an incorrect semantic front or in-plane heading.

The camera-up control uses $\mathbf{n} = \mathbf{u}_{\text{cam}}$ with the same construction. It isolates the substitution of an estimated supporting normal for fixed camera vertical. Plane fallback is distinct from answer fallback: a failed plane estimate may still yield a usable camera-up frame.

B.3 Projection and Candidate Coordinates

For $\mathbf{p} = (x, y, z)^\top$ with $y > 0$,

$$\pi(\mathbf{p}) = \begin{bmatrix} f_x x/y + c_x \\ c_y - f_y z/y \end{bmatrix}. \tag{10}$$

Axes originate at $\pi(\mathbf{c}_r)$. For $\mathbf{a} \in \{\mathbf{f}, \mathbf{r}, \mathbf{b}, \mathbf{l}\}$,

$$\mathbf{e}_\mathbf{a} = \mathbf{c}_r + 0.18\|\mathbf{c}_r\|_2\mathbf{a}, \qquad \mathbf{v}_\mathbf{a} = \pi(\mathbf{e}_\mathbf{a}) - \pi(\mathbf{c}_r). \tag{11}$$

For reference-box dimensions w_r, h_r in pixels, usable non-end-on arrows are rescaled to

$$\tilde{\mathbf{v}}_\mathbf{a} = 0.8\sqrt{w_r^2 + h_r^2}\,\frac{\mathbf{v}_\mathbf{a}}{\|\mathbf{v}_\mathbf{a}\|_2}. \tag{12}$$

This preserves projected direction while normalizing visible extent. Arrow length does not encode object distance. Near-zero projected displacements cannot be normalized by this expression. The semantic axes lie in the local horizontal plane; their projections need not be evenly spaced in the image.

Optional geometric prior. Candidate coordinates in the reference frame are

$$\mathbf{q}_i = B^\top(\mathbf{c}_i - \mathbf{c}_r). \tag{13}$$

They can supply an optional prior for map inference. With geometry coefficient zero, candidate coordinates do not score or rerank assignments, although reference geometry still determines the visual overlay. Geometry-assisted mapping additionally depends on candidate grounding and center estimation. It is therefore a separate intervention from ground-plane alignment.

B.4 Object Maps, Inversion, and Evaluation

The VLM receives the overlay, reference identity, candidate set $\mathcal{C}$, and target t. It predicts

$$M : \mathcal{D} \to \mathcal{C} \cup \{\perp\}, \qquad \mathcal{D} = \{\text{left}, \text{right}, \text{front}, \text{behind}\}. \tag{14}$$

All direction keys are required. Each concrete candidate may occur at most once, the target must occur exactly once, and only the unassigned value $\perp$ may repeat. The map is sparse and target-conditioned: it need not describe every object or populate every direction.

Validation checks keys, candidate membership, uniqueness, and target presence. At most one repair request is allowed. Structural acceptance does not establish spatial correctness. Given a usable frame, the final answer is

$$\hat{d}_{\text{final}} = \begin{cases} \text{unique}\{d : M(d) = t\}, & \text{accepted map,} \\ \hat{d}_{\text{direct}}, & \text{otherwise.} \end{cases} \tag{15}$$

Direct answering uses the same overlay, or the original image if no usable overlay exists. Accepted maps are not overridden by disagreement with the direct answer. Gold labels are used only for evaluation.

C Prompting and Diagnostic Conditions

The following templates summarize the task instructions; italicized placeholders vary by example. Equivalent templates are grouped to avoid repetition. Images accompany the visual tasks. Direction terminology is normalized so that back and behind denote the same relation.

C.1 Paired Spatial Questions

The common instruction is:

> Answer this spatial-reasoning question using the image. Select one answer option and respond with its letter.

The question is followed by four labeled options. The requested response is a bare A–D letter; conventional enclosing punctuation and answer prefixes are also recognized. Letters refer to the displayed option order.

COMFORT-derived questions.

> **Direction:** From the *reference*'s current viewpoint, where is the *target*?
> **Object:** Which object is in the direction that the *relation* side of the *reference* points toward?

Kubric questions. Camera-relative questions begin "From the camera's perspective." Object-facing-camera questions instead begin "From the *reference*'s perspective while it faces the camera." They continue with:

> **Direction:** Where is the *target* relative to the *reference*?
> **Object:** Which object is to the *relation* of the *reference*?

3DSR questions. Both formulations begin "Consider the real-world 3D locations and orientations of the objects." The native question is preserved, with the form:

> **Direction:** Which side of the *reference* is facing the *target*?
> **Object:** Which object is in the direction that the *relation* side of the *reference* points toward?

C.2 Visual Assistance

Assistance conditions retain the authored question, options, and letter response. The unassisted condition has no prefix. Other conditions prepend:

> **Front arrow:** Image aid: the cyan arrow indicates the front of the reference object (*reference*).
> **Four axes:** Image aid: the four labeled arrows superimposed on the reference object (*reference*) show its own left, right, front, and behind directions.
> **Two axes:** Image aid: the two short labeled arrows superimposed on the reference object (*reference*) show its own front and left directions. Infer behind as opposite front and right as opposite left.

The arrow-length sweep uses scales $\{0.25, 0.45, 0.70, 0.90, 1.15, 1.50\}$ times the reference-box diagonal and the prefix:

> Four labeled reference-relative arrows are overlaid on the reference object. Their length is *scale* times the reference bbox diagonal.

Thus the sweep varies both visible length and its textual description.

C.3 Localization, Orientation, and Axis Use

Image-coordinate outputs use a top-left origin, x increasing rightward, and y increasing downward. Point and box coordinates are normalized to $[0, 1000]$.

Localization and recognition.

> **Box prediction:** Locate the *object* in the plain image. Return exactly one bounding box as $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ in [0,1000], where (0,0) is top-left.
> **Box naming:** A magenta box labeled BOX is overlaid on one object. What object is inside it? Return exactly one name from: *candidate names*.

Projected facing direction. The boxed condition identifies the reference with "The reference *reference* is boxed." The unboxed condition uses "The reference object is the *reference*." Both continue:

> In which image-plane direction is its own front facing? Return exactly one of: right, down-right, down, down-left, left, up-left, up, up-right.

The horizontal-mirror control keeps this prompt fixed and mirrors the horizontal component of the gold label. These are eight-compass-label responses, not option letters.

Axis production. For *axis* equal to front or left:

> The reference *reference* is boxed. Predict a 2D arrow pointing toward the object's own *axis*. Start at the object's center.

The requested structured response contains named start and end points in normalized image coordinates.

Axis reading. For a supplied ground-truth front or left arrow:

> The reference *reference* is boxed. The overlaid arrow shows the reference object's ground-truth *axis* axis. Using the reference object's own perspective, where is the *target* relative to it? Return exactly one of: left, right, front, back.

Symbol grounding. Abstract markers support two inverse queries:

> **Symbol to object:** Each surrounding object has an abstract letter marker. Which object is marked *symbol*? Return exactly one name from: *candidate names*.
>
> **Object to symbol:** Each surrounding object has an abstract letter marker. Which marker is on the *object*? Return exactly one of: A, B, C, D.

Long- and short-arrow symbol-selection conditions share:

> The reference object's labeled arrows show its own directions and the letters mark surrounding objects. Which letter marks the object *direction* of the reference? Return exactly one of: A, B, C, D.

Here a letter identifies a scene marker rather than a position in an answer list.

C.4 Coordinate and Projected-Axis Tasks

Coordinate conversion. Both conversion tasks state: "In the reference object's coordinates, positive axes are front, up, and right." Vector-to-direction supplies named component values and requests one of front, back, above, below, right, or left. Direction-to-vector supplies a direction and requests a structured unit vector with named front, up, and right components. Names determine interpretation; the presentation order does not replace the canonical positional convention $[\mathrm{right}, \mathrm{front}, \mathrm{up}]$.

Projected-axis prediction.

> The reference *reference* is boxed. Predict its screen-projected front, up, and right unit directions in image coordinates, where x points right and y points down.

The response associates each named axis with a two-dimensional unit direction.

Numerical and visual axis assistance. The numerical condition states the projected front, up, and right vectors in the prompt. The visual condition instead states:

> The image overlays arrows labeled front, up, and right on the reference object.

Both ask where the target lies relative to the named reference and require left, right, front, or back. The comparison changes how the basis is supplied, while retaining the relation-classification task.

C.5 Representation Perturbations

The main representation analysis uses 218 matched Kubric object-answer queries. RGB inputs contain four labeled reference axes without candidate markers or a symbol legend. Symbolic inputs replace appearance with entity symbols and a text legend, aligning the reference heading with image-up without an explicit heading arrow. Questions, options, and original answer labels are retained.

Independent perturbation suites include non-reference detection dropout at 25%, 50%, and 75%; candidate-centroid Gaussian noise at 10%, 25%, 50%, and 100% of the median reference-to-object distance; reference-centroid noise at the same fractions of camera-to-reference distance; and heading rotations of $15°$, $30°$, $45°$, $90°$, and $180°$. Candidate dropout and candidate-centroid noise change only symbolic inputs. Reference-centroid and heading errors affect both representations. These are separate sweeps, not factorial combinations.

An earlier heading-arrow control instead places object markers and a red heading arrow in both representations. Its instruction explains the symbol-to-object legend and identifies the red arrow as the reference heading. That instruction does not describe the main four-axis RGB and heading-up symbolic comparison.

C.6 FrameBind Map Instructions

Direct answering retains the authored question and choices. Map inference receives the reference, candidate list, target, and selected overlay. Its instruction constraints can be summarized as:

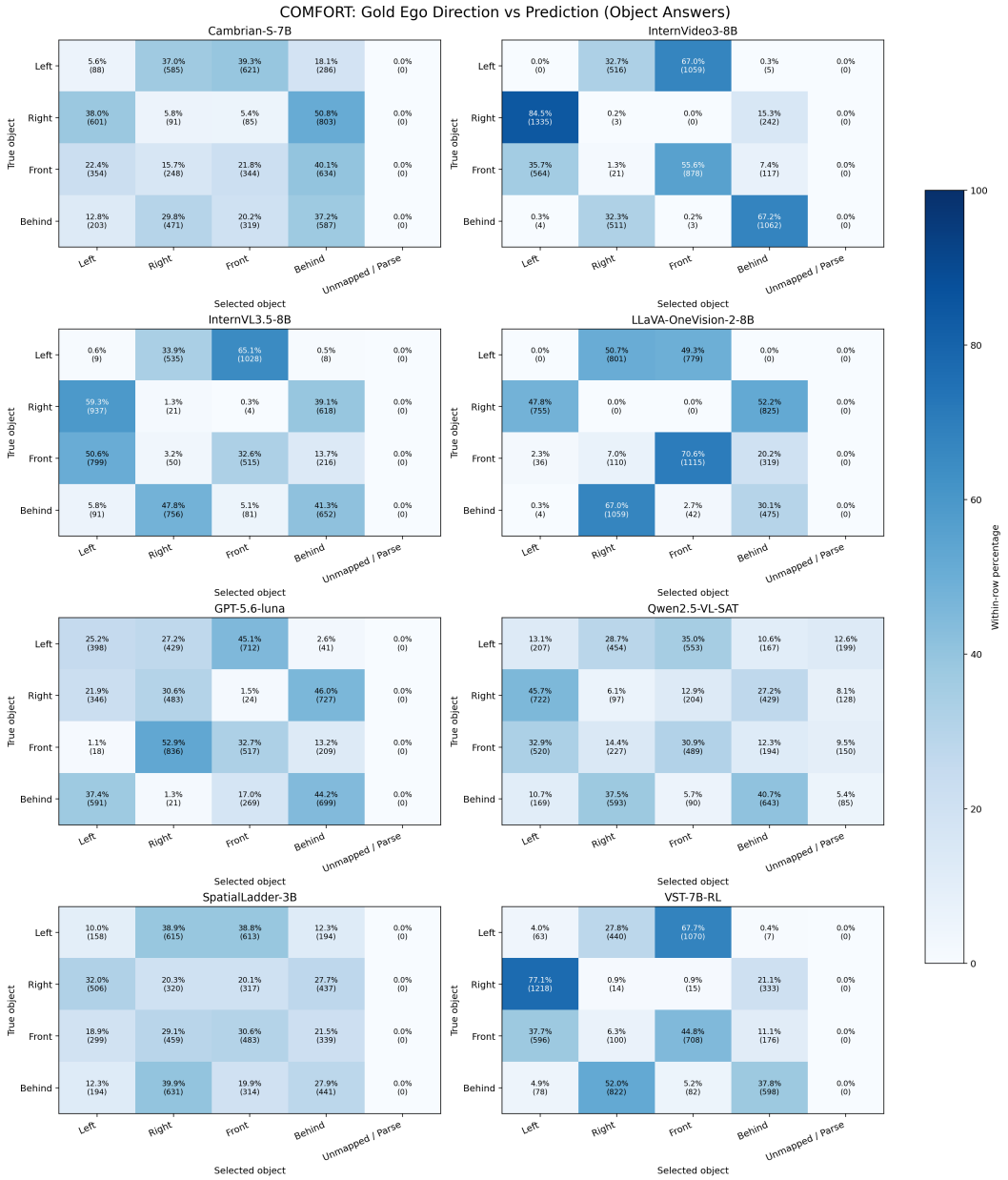
> From the reference object's perspective, identify which listed object lies in each direction. Return a structured map with keys left, right, front, and behind. Use a listed candidate or none for each value. Include the target exactly once and do not repeat any concrete object.

This is a constraint summary rather than a verbatim prompt transcription. Only the unassigned value may repeat. Repair requests specify violated structural constraints without supplying the gold relation. Ground-plane alignment changes the overlay, not the map's requested structure.

D CAMERA-FRAME ALIGNMENT ACROSS BASELINES

The following matrices compare predictions with camera-relative interpretations for the object-answer and direction-answer formulations. They supplement the reference-frame diagnostics in the main text. Camera-aligned errors characterize observable behavior; they do not directly identify an internal coordinate representation.

D.1 Object-Answer Confusion Matrices



D.2 Direction-Answer Confusion Matrices

