EGOSR-BENCH: UNVEILING MULTI-HOP EGOCENTRIC SPATIAL REASONING IN PANORAMIC ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding complex spatial relations from an egocentric perspective is fundamental for embodied agents to navigate, reason, and interact within real-world panoramic environments. While recent advances in three-dimensional (3D) scene understanding and visual question answering (VQA) have improved object recognition and grounding, structured multi-hop spatial reasoning remains significantly underexplored, especially under single-view constraints such as 360° panoramic observations. In this work, we introduce EgoSR-Bench, a benchmark for multi-hop spatial reasoning from panoramic egocentric input, challenging models to infer fine-grained spatial relations solely from a single red-green-blue (RGB) 360° observation. We propose a structured question-generation framework that leverages 3D scene geometry, object-level attributes, and large language model (LLM)-based linguistic refinement to construct logically grounded questions. Based on this framework, we build OcPano, the released dataset and manifest underlying EgoSR-Bench, with **3,577** structured instances from 15 scenes, **multilingual variants**, and **dense spatial-semantic annotations**. We further define evaluation protocols for diagnosing spatial reasoning performance and use post-training to study benchmark learnability.

1 INTRODUCTION

Spatial intelligence is fundamental to an agent's ability to perceive, reason about, and interact with complex physical environments. For embodied agents, this capability requires constructing a coherent spatial mental model from egocentric observations. Recent benchmarks study omnidirectional understanding from complementary perspectives: OSR-Bench targets spatial reasoning Dongfang et al. (2025), ODI-Bench evaluates general and spatial understanding Liu et al. (2026), and PanoEnv grounds VQA in 3D panoramic geometry Lin & Zheng (2026). However, these evaluations provide limited evidence about whether a model can compose several observer-centered spatial relations into a multi-hop query and identify the correct target. This motivates a more focused study of multi-hop egocentric spatial reasoning from a single panoramic observation.

Panoramic imagery provides a wide-field representation for embodied agents, capturing the surrounding scene in a single equirectangular projection (ERP). Compared with narrow pinhole views, a panorama preserves broader spatial context in one observation, which is useful for assessing object locations and relative spatial relations across a scene. However, interpreting spatial relations in a 360° image introduces geometric challenges: ERP distortion changes the apparent scale and shape of objects across the image, while the 0°–360° boundary can separate spatially adjacent objects onto opposite edges of the image.

Despite recent progress in omnidirectional VQA Dongfang et al. (2025); Liu et al. (2026); Lin & Zheng (2026), existing benchmarks largely emphasize object counting, relative distance and direction, or general geometry-grounded spatial understanding. They therefore provide limited evidence about whether a model can compose multiple observer-centered spatial relations from a single panoramic observation. This raises a central question: can current multimodal large language models (MLLMs) reliably perform multi-hop spatial reasoning over a single ERP image and identify the correct target?

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.

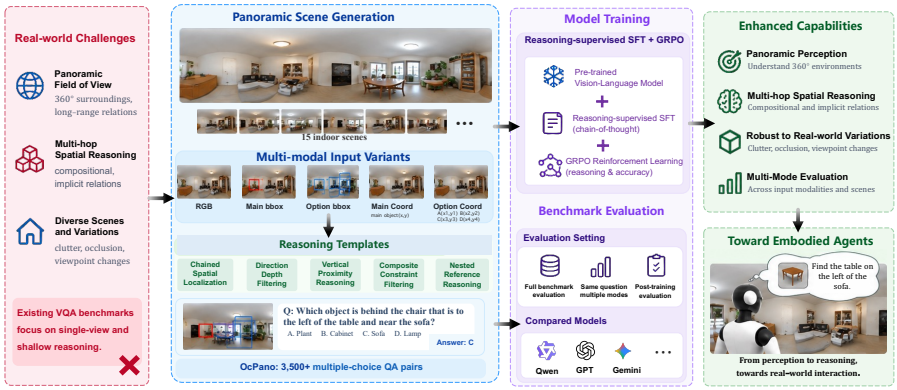


Figure 1: Conceptual overview of EgoSR-Bench. Top: an observer answers a logic-chain question from a single panoramic RGB observation; the example combines "on the bed," "near the lamp," and "below the picture" to identify a pillow. Bottom: predefined logic chains are grounded in object-level 3D geometry, and deterministic predicate execution retains a unique valid option as the computed label. The scene, predicate values, and displayed model prediction are illustrative.

Our dataset, **OcPano**, comprises 3,577 multiple-choice questions generated from 15 photorealistic indoor scenes. The questions are instantiated from five implemented reasoning templates: Chained Spatial Localization (CSL), Direction and Depth Filtering (DDF), Vertical and Proximity Reasoning (VPR), Composite Constraint Filtering (CCF), and Nested Reference Reasoning (NRR). We use these paper-level names consistently throughout the manuscript. LLM-based paraphrasing provides English and Chinese variants while preserving the underlying spatial constraints. Object-level 3D annotations provide geometry-derived labels for question generation and verification, enabling controlled analysis across reasoning templates, relation types, and scene configurations. Throughout the paper, EgoSR-Benchdenotes the benchmark, whereas OcPanodenotes its released dataset and manifest. Figure 1 illustrates the observation–question–answer setting.

We evaluate benchmark learnability by adapting MLLMs with reasoning-supervised fine-tuning (SFT) followed by Group Relative Policy Optimization (GRPO). The post-training split uses 80% of the questions for training and one held-out 857-question test set (20%), separate from the 3,577-question benchmark evaluation.

Our primary contributions are: **(I)** We formalize the task of **multi-hop panoramic spatial reasoning**, shifting the focus from simple recognition to structured geometric inference.

(II) We release **OcPano**, a 3D-grounded benchmark featuring 3,577 multilingual questions designed to diagnose compositional spatial reasoning.

(III) We evaluate benchmark learnability with a **reasoning-supervised SFT+GRPO setup** and define the corresponding scoring protocol.

Figure 2 contrasts conventional VQA-style settings with the multi-hop panoramic spatial reasoning task studied in EgoSR-Bench.

2 Related Works

Embodied question answering (QA) and Vision-Language Navigation: Classical visual question answering (VQA) established the image–language evaluation setting Antol et al. (2015); Wu et al. (2017), while compositional benchmarks such as CLEVR, Visual Genome, GQA, AGQA, NLVR2, and Winoground probe structured relations, dense grounding, or visio-linguistic composition Johnson et al. (2017); Krishna et al. (2017); Hudson & Manning (2019); Grunde-McLaughlin et al. (2021); Suhr et al. (2019); Thrush et al. (2022). Embodied QA and navigation benchmarks further require agents to connect language, egocentric perception, and actions. EmbodiedQA requires an agent to navigate before answering a question Das et al. (2018); Room-to-Room, Touchdown, and RxR study language-guided navigation in real, street-view, or multilingual environments Anderson et al. (2018); Chen et al. (2019); Ku et al. (2020); EmbSpatial-Bench evaluates spatial reasoning in embodied settings Du et al. (2024); and ALFRED evaluates long-horizon household instruc-

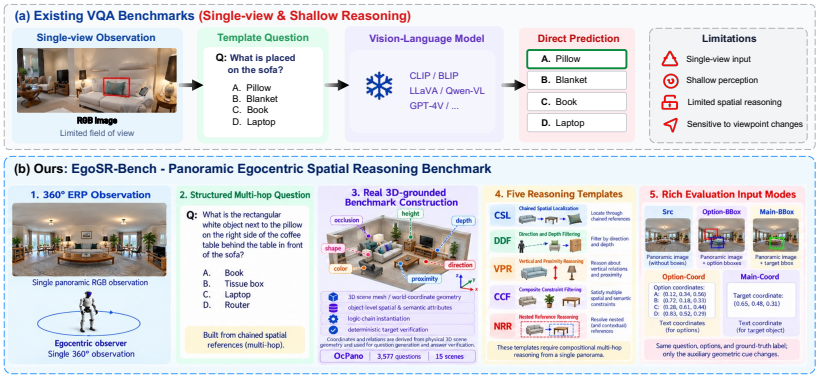


Figure 2: Comparison between conventional VQA-style tasks and EgoSR-Bench. Traditional VQA, attribute recognition, geometric VQA, and simple spatial reasoning typically require localized or single-step perception, while EgoSR-Benchemphasizes multi-hop egocentric spatial reasoning grounded in panoramic observations.

Table 1: Comparison with representative visual question answering, embodied, panoramic, and spatial-reasoning benchmarks. "Partial" indicates that a property is present only for part of the benchmark or is not its primary supervision. The final column marks whether a benchmark uses egocentric or embodied observations; it does not imply navigation or action execution.

Dataset	Input	Panoramic	3D Scene	Spatial	Multi-hop	Egocentric / Embodied setting
VQA v2 / GQA Goyal et al. (2017); Hudson & Manning (2019)	Perspective image	No	No	Limited	Partial	No
EmbodiedQA / ALFRED Das et al. (2018); Shridhar et al. (2020)	Embodied RGB	No	Partial	Partial	Partial	✓
VQA-360 / Pano-AVQA Chou et al. (2020); Hu et al. (2021)	360 image / video	✓	Partial	Limited	Limited	Partial
ScanQA / SQA3D Azuma et al. (2022); Ma et al. (2023)	3D scan / RGB-D	No	✓	✓	Partial	✓
VSI-Bench Yang et al. (2025)	Perspective image	No	Partial	✓	Partial	No
OpenEQA Majumdar et al. (2024)	Embodied RGB / language	Partial	Partial	✓	Partial	✓
SpatialVLM Chen et al. (2024)	Perspective image	No	Partial	✓	Partial	No
OSR-Bench / ODI-Bench Dongfang et al. (2025); Liu et al. (2026)	Omnidirectional image	✓	Partial	✓	Partial	Partial
PanoEnv Lin & Zheng (2026)	ERP panorama + 3D	✓	✓	✓	Partial	✓
OcPano	ERP panorama	✓	✓	✓	✓	✓

tions with egocentric vision Shridhar et al. (2020). These settings emphasize exploration and action grounding. In contrast, EgoSR-Bench removes the navigation policy from the loop and directly probes whether a model can infer multi-hop spatial relations from one static panoramic observation.

Panoramic and Omnidirectional Understanding: Panoramic perception has been studied for full-scene geometry and dense understanding. Matterport3D provides building-scale red-green-blue-depth (RGB-D) scans with panoramic viewpoints Chang et al. (2017), while HorizonNet and HoHoNet show that 360° panoramas can support room layout estimation, depth prediction, and semantic segmentation Sun et al. (2019; 2021). VQA-360° introduces question answering on omnidirectional images Chou et al. (2020), and Pano-AVQA further studies grounded audio-visual QA on 360° videos Hu et al. (2021). These works demonstrate the importance of spherical context, but they do not focus on 3D-grounded multi-hop object relations from a single egocentric ERP image.

Spatial Reasoning Benchmarks with MLLMs: Recent benchmarks for multimodal large language models show that spatial reasoning remains difficult even for strong general-purpose systems. VSR and BLINK probe visual spatial relations and broad multimodal perception Liu et al. (2023); Fu et al. (2024); OmniSpatial and Thinking in Space study spatial reasoning under broader or explicit spatial-intelligence settings Jia et al. (2026); Yang et al. (2024); and VSI-Bench Yang et al. (2025) highlights failures in global spatial layout understanding. OpenEQA Majumdar et al. (2024) evaluates open-vocabulary embodied QA with persistent gaps on spatial questions, while SpatialVLM Chen et al. (2024) introduces large-scale spatial QA supervision to improve metric reasoning in images. 3D grounding benchmarks such as ScanRefer, ReferIt3D, 3D-LLM, and Space3D-Bench Chen et al. (2020); Achlioptas et al. (2020); Hong et al. (2023); Szymanska et al. (2024) further establish language-conditioned reasoning over reconstructed scenes. However, these efforts do not isolate panoramic boundary continuity, ERP distortion, and observer-centered multi-hop reference chains in a single diagnostic benchmark.

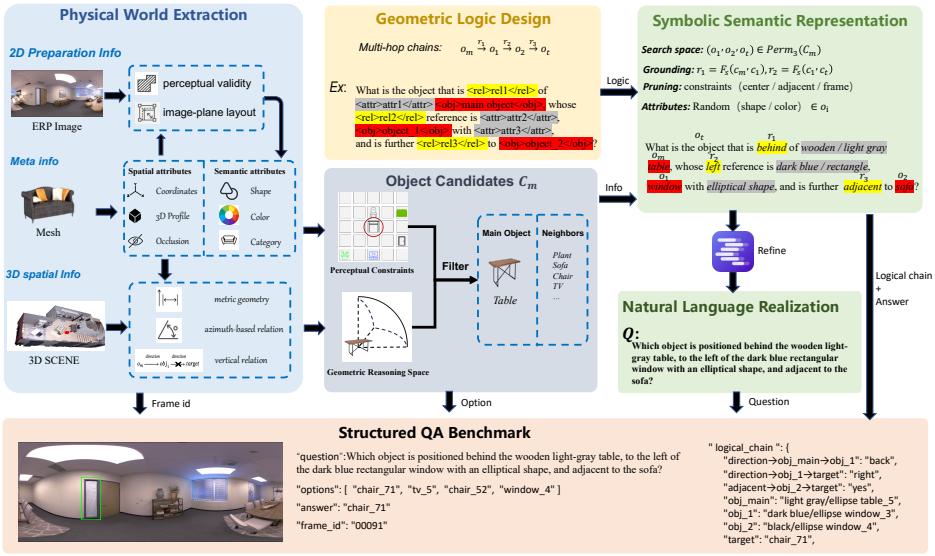


Figure 3: Overview of the EgoSR-Benchdataset construction pipeline. Spatial and semantic attributes are extracted from 3D object meshes and ERP images, aligned across modalities, used for logic design and hard distractor selection, and finally converted into structured question-answer pairs with deterministic answer selection.

3 DATASETS

To facilitate research in observer-centered spatial mental modeling, we introduce **OcPano**, an observer-centered panorama question answering dataset derived from ReplicaPano-style photorealistic indoor scenes Straub et al. (2019). Each sample consists of a natural language question, a panoramic RGB image, a set of answer choices, and a correct label. In contrast to traditional VQA datasets, EgoSR-Benchemphasizes high-level spatial reasoning from a single equirectangular panoramic image with a full 360° field of view. Figure 3 provides an overview of the dataset construction pipeline. Beginning from aligned 3D scene layouts and ERP images, we extract object-level spatial and semantic attributes, construct 2D-3D candidate groups, instantiate the five reasoning templates, sample hard distractors, and refine the final questions with LLMs. Ground-truth spatial relations and answer labels are determined through deterministic constraint evaluation using world-coordinate geometry and the associated object attributes.

3.1 QUESTION DEFINITION AND REASONING TEMPLATES

Each data sample in EgoSR-Benchis represented as a structured tuple $\langle Q, I, O, y \rangle$, where Q denotes the question, I the panoramic image, $O = \{o_1, \ldots, o_K\}$ ($K \in \{3, 4\}$ in the full manifest) the candidate answers, and y the correct label. The design evaluates an agent's ability to perform observer-centered spatial reasoning by integrating multi-hop spatial relationships with semantic attributes. The available visual-input configurations are documented in the supplementary material; they vary the auxiliary geometric annotations while keeping the question, candidate set, and ground-truth label fixed.

To illustrate the construction process, Figure 4 provides an example of a panoramic question–answer pair together with its underlying spatial reasoning process. Starting from a single egocentric observation, the question gradually narrows down the target through a sequence of spatial references, reflecting how a human observer would resolve complex spatial descriptions in everyday environments.

Each question in EgoSR-Bench is instantiated from one of five implemented reasoning templates that capture complementary forms of observer-centered multi-hop inference. *CSL* progressively narrows the target through a sequence of directional, depth, or proximity relations; *DDF* combines an egocentric sector such as front-left with near/far depth and attribute constraints; *VPR* couples height relations such as above or below with local proximity; *CCF* identifies an object by jointly satisfying several conditions, including direction, depth, occlusion, and support; and *NRR* resolves

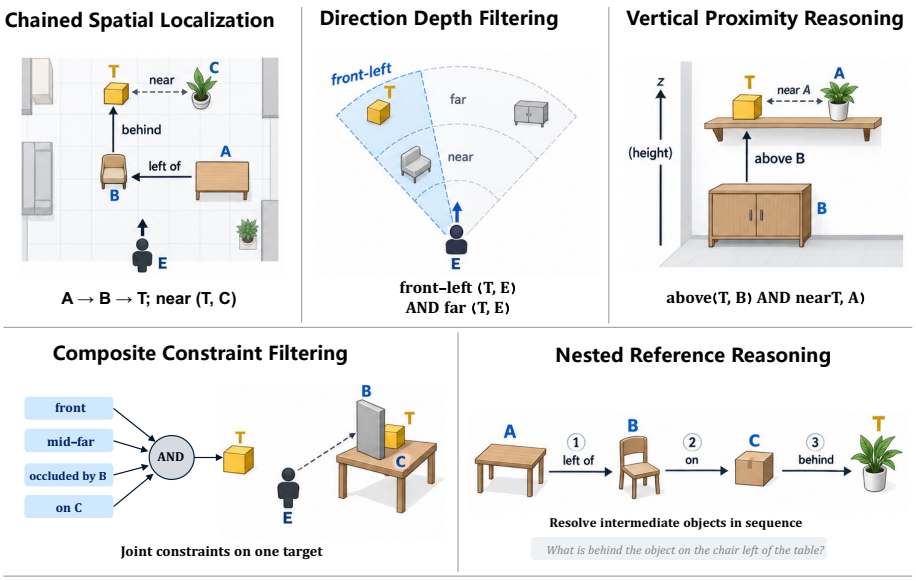


Figure 4: Five implemented reasoning templates in EgoSR-Bench. From left to right and top to bottom: (a) CSL, (b) DDF, (c) VPR, (d) CCF, and (e) NRR. Each panel illustrates the spatial constraints that must be composed to identify the target object from a single panoramic observation.

an indirect target by traversing a chain of references. Together, these names form the paper-level taxonomy and ground natural-language questions in physically consistent 3D layouts while testing distinct ways of composing spatial evidence.

3.2 Spatial-Semantic Attribute Annotation.

To support structured reasoning and controlled linguistic variation, each object in a scene is annotated with a set of spatial and semantic attributes. These attributes are automatically extracted from the underlying 3D geometry, world-coordinate positions, projected 2D bounding boxes, and image appearance. The resulting representation provides a common interface between logic-chain execution and natural-language question generation. A summary of the attribute roles used in question generation is provided in Supplementary Table S2.

For an object o_i, we denote its attribute representation as

$$\mathrm{attr}(o_i) = (\mathrm{direction}, \mathrm{depth}, \mathrm{height}, \mathrm{occlusion}, \mathrm{color}, \mathrm{shape}, \ldots).$$

The spatial components of this representation are used during symbolic execution to constrain the selection of valid intermediate and target objects. For example, a question instantiated from a direction–depth template retains only candidates that satisfy the corresponding directional and depth relations with respect to the reference objects. In contrast, semantic attributes such as color, shape, and occlusion are primarily used to diversify the linguistic realization of object references while preserving the underlying spatial logic.

Attribute injection into questions. After a valid logic chain has been instantiated, descriptive attributes are inserted into the corresponding object references according to

$$Q_{\mathrm{final}} = \mathrm{Template}\big(o_i, \mathrm{attr}(o_i)\big).$$

This operation changes the surface form of the question without modifying its symbolic relations or ground-truth answer. For example, the template question “What is the object behind the table near the chair?”may be realized as “What is the red object behind the table near the chair?”. Attributes are selected from a predefined priority order, such as color, shape, and occlusion, and are introduced only when they remain visually grounded in the observation. When multiple valid descriptors are available, different selections may generate linguistic variants of the same spatial reasoning instance.

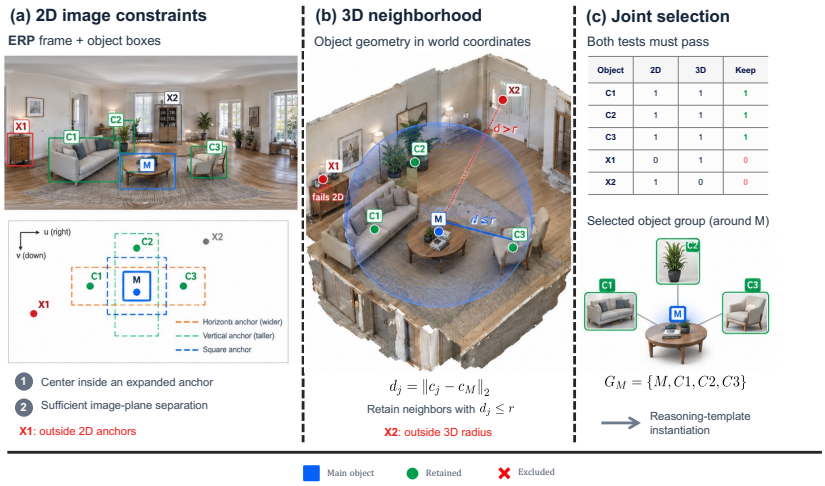


Figure 5: Joint 2D-3D candidate grouping. Image-plane anchors and world-coordinate neighborhoods are checked jointly, and only candidates that satisfy both constraints are retained for reasoning-template instantiation.

3.3 Object Grouping and Reasoning Template Instantiation

Each question is generated from a local object group centered on a main object o_m. The group construction combines image-plane evidence with world-coordinate geometry so that the resulting reasoning chain is both visually identifiable and physically consistent. The overall construction pipeline is summarized in Figure 3.

Joint 2D–3D candidate grouping. For every valid main object, we first identify objects that share the same scene frame and satisfy the geometric, visibility, and semantic requirements used by the dataset construction pipeline. A candidate is retained only when its 3D volume and metadata are valid, its projected region is sufficiently visible in the ERP image, and its category is compatible with the target-object set. The local 3D neighborhood is defined as

$$\mathcal{N}_m = \{o_j \in \mathcal{O} \setminus \{o_m\} \mid \|\mathbf{c}_j - \mathbf{c}_m\|_2 \le \delta_d,\ \mathrm{frame}(o_j) = \mathrm{frame}(o_m)\}, \tag{1}$$

which limits candidates to objects within a fixed radius of the main object. In parallel, the projected 2D boxes are examined to ensure adequate image-plane separation and compatibility with the expanded anchors around the main object. We additionally enforce a minimum distance between candidate centers and favor groups with semantic diversity, thereby reducing visually trivial or highly ambiguous configurations. Figure 5 illustrates how the image-plane and world-coordinate checks are combined to retain a consistent local candidate group.

Geometry-grounded template instantiation. Once a valid candidate group has been obtained, an abstract reasoning template is grounded in the scene through the geometry-driven mapping function $\mathcal{F}_s(\cdot)$. This function converts pairs of world-coordinate centroids into deterministic spatial predicates, such as direction, depth, height, and proximity. For a multi-hop template, the main object, intermediate references, and target object are assigned to an ordered candidate tuple (o_1, o_2, o_t), where o_2 is the auxiliary object used by the final adjacency constraint, according to

$$o_m \xrightarrow{r_1} o_1 \xrightarrow{r_2} o_t, \qquad r_1 = \mathcal{F}_s(\mathbf{c}_m, \mathbf{c}_1), \quad r_2 = \mathcal{F}_s(\mathbf{c}_1, \mathbf{c}_t). \tag{2}$$

The validity indicator

$$\Psi = \mathbb{I}(r_1, r_2 \in \mathcal{R} \ \wedge\ \|\mathbf{c}_2 - \mathbf{c}_t\|_2 \le \delta_{\mathrm{adj}}) \tag{3}$$

retains only those tuples for which all required predicates and the final adjacency constraint are satisfied. Because the answer label is selected by executing these predicates over the scene geometry, each retained question has a unique and deterministic target.

Hard distractor sampling. The final candidate set is completed with distractors sampled from the local neighborhood or, when necessary, from the wider scene. A distractor is required to remain

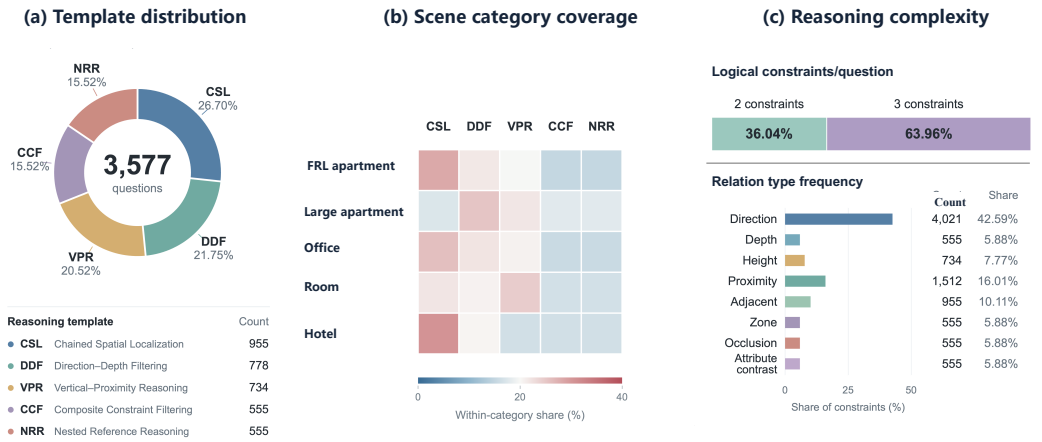


Figure 6: Composition and reasoning-structure statistics of OcPano. (a) Distribution of the 3,577 questions across the five reasoning templates. (b) Template shares within five scene categories, where cell colors encode the within-category share. (c) Distribution of questions with two versus three logical constraints and frequencies of the 9,442 relation instances.

plausible with respect to the target in terms of visual appearance, semantic category, or projected location, while violating at least one relation in the instantiated logic chain. Candidate groups are discarded when multiple objects satisfy the complete chain, when the target is insufficiently visible, or when the distractors can be rejected using a superficial visual cue. This procedure prevents models from solving the task through simple category or appearance matching and preserves the intended multi-hop spatial reasoning challenge.

3.4 Dataset Statistics and Diversity

Quantitative overview and source version. Our released benchmark contains 3,577 instances generated from 15 photorealistic indoor scenes and 119 distinct observations defined by a scene and frame. Each instance is linked to the processed evaluation manifest through its scene identifier, frame, raw template, logic chain, target answer, and ordered candidate options, providing a deterministic basis for the statistics and evaluations reported below. A separate historical export contains 6,157 entries. Because its generation lineage and correspondence with the current release have not been fully reconciled, we report it separately and exclude it from the statistics in this section. Figure 6 summarizes the distribution of the five implemented templates. The complete scene-level table and per-scene distribution are provided in Supplementary Section S3.

Implemented template and scene coverage. The benchmark contains five implemented reasoning templates: CSL, DDF, VPR, CCF, and NRR, with 955, 778, 734, 555, and 555 instances, respectively. NRR combines zone, occlusion, height, and attribute constraints, so its name does not imply that every instance uses the same primitive relation. The dataset contains 2,349 apartment instances, 464 room instances, 443 office instances, 174 large apartment instances, and 147 hotel instances. These counts describe dataset coverage only; they should not be interpreted as measures of scene complexity or evidence that object density determines instance yield.

4 Benchmarks

4.1 Evaluation Protocol

We evaluate every model on the 3,577-question benchmark under five input modes: Src, Opt. BBox, Main BBox, Opt. Coord, and Main Coord. The modes differ only in the geometric cue provided with the same panoramic question, candidate set, and answer label. We use exact-match accuracy as the evaluation metric. The post-training experiments use an 80% training split and a single 857-question held-out test set (20%). Dataset consistency checks and the clean-recomputation subset are documented in Supplementary Section S1. We also conduct a diagnostic human study using a separate presentation protocol; its design and results are given in Appendix S4.

Table 2: Accuracy (%) across five input modes. `Src` is the raw ERP panorama; `Opt. BBox` and `Main BBox` add boxes for candidate and main objects, respectively; `Opt. Coord` and `Main Coord` add their coordinates. Mean is the average across the five modes. Bold marks the highest value in each column and in Mean.

Group	Model	Src	Opt. BBox	Main BBox	Opt. Coord	Main Coord	Mean
GPT Official source	GPT-6-Astra	**37.35**	**35.20**	**37.29**	43.70	**37.85**	**38.28**
	GPT-4.1-mini	35.56	31.70	35.53	39.53	34.61	35.39
	GPT-5.6-Sol	35.67	31.34	34.83	38.80	35.20	35.17
	GPT-5.6-Terra	32.26	31.45	33.44	37.43	32.04	33.32
	GPT-5.6-Luna	32.09	31.34	32.82	35.62	31.17	32.61
	GPT-4o	30.11	29.86	30.81	35.06	30.19	31.20
Gemini Official source	Gemini-3.7-Flash	37.13	30.58	37.10	43.25	35.87	36.79
	Gemini-3.1-Pro	37.10	28.63	35.28	**45.26**	31.90	35.63
	Gemini-3-Flash	35.39	28.60	35.17	41.07	35.73	35.19
	Gemini-2.5-Pro	34.58	31.98	36.32	35.50	33.86	34.45
	Gemini-2.5-Flash	34.67	30.14	36.85	34.89	35.53	34.41
Qwen Official source	Qwen3.8-Max	35.64	33.69	34.30	40.76	35.03	35.88
	Qwen3-VL-Plus	36.06	32.07	36.65	37.55	36.09	35.68
	Qwen3.7-Flash	34.58	30.98	34.89	40.42	34.16	35.01
	Qwen3.7-Plus	33.44	31.09	34.58	41.88	33.52	34.90
Qwen local	Qwen3-VL-30B (BF16)	31.14	29.89	32.12	34.64	33.46	32.25
	Qwen2.5-VL-7B (Q8)	30.22	28.49	29.27	36.37	30.42	30.95
	Qwen3-VL-4B (8-bit)	30.70	29.05	30.67	31.90	30.61	30.58
	Qwen3-VL-8B (BF16)	27.93	28.38	29.77	34.33	32.04	30.49
	Qwen2-VL-2B (BF16)	24.63	22.84	23.29	26.00	24.07	24.17
InternVL local	InternVL3.5-2B (Q8)	28.26	27.73	28.18	29.97	27.34	28.30
	InternVL3.5-2B (BF16)	28.32	27.73	28.21	29.75	27.26	28.25
	InternVL3.5-8B (Q8)	30.16	29.10	29.58	35.45	30.72	31.00
	InternVL3.5-8B (BF16)	30.22	29.10	29.44	35.67	30.81	31.05
	InternVL3.5-14B (BF16)	32.21	32.35	31.56	32.51	33.02	32.33
Other local	Gemma-4-31B	34.58	33.07	35.92	37.77	34.44	35.16
	MiniCPM-V-4.5	32.37	30.78	33.77	36.43	31.00	32.87
	Kimi-VL-A3B	32.63	30.56	31.56	32.88	30.25	31.57
	Gemma-4-12B	31.06	28.46	30.22	32.71	31.73	30.84
	Llama-4-Scout-17B	27.34	27.57	26.78	38.80	27.48	29.59
	GLM-4.1V-9B	28.35	27.82	29.27	31.31	28.26	29.00
	MiniCPM-V-4.5-Reasoning	19.85	20.02	22.34	25.10	21.53	21.77

4.2 Overall Benchmark Results

Table 2 reports accuracy across model families and input modes. GPT-6-Astra obtains the highest mean accuracy (38.28%), followed by Gemini-3.7-Flash (36.79%) and Qwen3.8-Max (35.88%). The best InternVL3.5 configuration reaches 32.33%. The overall accuracy remains low, showing that the benchmark is challenging even when geometric cues are provided. Coordinate-based inputs generally provide stronger performance than bounding-box inputs, particularly for the candidate-object setting, although the gain varies across models. This pattern suggests that explicit geometric locations can facilitate spatial inference, but they do not eliminate the need to compose multiple relations across the scene. Detailed analyses by reasoning template and input configuration are provided in the supplementary material. We group models by provider and deployment: GPT, Gemini, and Qwen models from official sources, followed by locally deployed Qwen, InternVL, and other models. For local Qwen rows, the recorded runtime formats are shown in parentheses (BF16, Q8, or 8-bit); corresponding precision tags are retained for InternVL configurations when they distinguish model variants. Supplementary material provides the input-mode definitions, dataset statistics, human-study protocol, and additional post-training analyses.

5 Methods

5.1 Benchmark-Oriented Post-training

We use reasoning-supervised post-training to measure how readily open-source vision-language models learn the structured tasks in EgoSR-Bench. The training framework is `ms-swift`; supervised fine-tuning is followed by GRPO. This experiment evaluates benchmark learnability and does not introduce a new optimization method. During SFT, the model maps the panoramic image and task instruction to a structured answer. GRPO samples multiple responses for each input and

Table 3: Post-training accuracy (%) across the five input modes. All rows use the same 857-question held-out test set.

Model	Src	Opt. BBox	Main BBox	Opt. Coord	Main Coord
InternVL3.5-2B	36.82	39.05	40.30	70.36	71.76
InternVL3.5-8B	36.82	33.83	36.07	70.13	73.63
InternVL3.5-14B	38.56	41.79	37.81	66.28	67.91
Qwen2-VL-2B	30.60	37.81	31.09	73.98	75.38
Qwen2-VL-7B	33.83	37.06	34.08	74.02	75.85
Qwen2.5-VL-3B-Instruct	36.32	37.06	36.32	67.72	73.54
Qwen2.5-VL-7B-Instruct	36.82	40.05	38.31	67.68	67.69
Qwen3-VL-4B-Instruct	35.57	44.28	42.54	66.51	65.46
Qwen3-VL-8B-Instruct	36.82	48.01	46.77	68.11	68.63

optimizes them with a task-verifiable reward. The five task formats correspond to Src, Opt. BBox, Main BBox, Opt. Coord, and Main Coord, covering target recognition, bounding-box prediction, candidate selection, and coordinate localization.

5.2 Post-training Evaluation

We evaluate the adapted models on the corresponding tasks and compare them with the corresponding base configurations. Results are organized by model family and input mode. When a matched base result exists, we report the accuracy difference; otherwise, the comparison does not imply a causal improvement.All post-training evaluations use the same 857-question held-out test set. Accuracy is computed from responses that can be parsed into one of the candidate labels; missing or invalid responses are not counted as correct. The evaluation procedure and its limitations are described in Supplementary Section S2.Table 3 reports the post-training accuracy for the five input modes. A common-question comparison between the base and GRPO models is provided in Appendix S5; task-level accuracy differences are reported with their valid-response denominators.

6 Implications for Embodied Agents

The benchmark targets a capability that is directly relevant to embodied agents operating in structured environments. Consider an autonomous laboratory in which a mobile or humanoid agent must localize instruments, sample containers, work benches, robotic arms, and safety-constrained regions from a panoramic observation. A request such as "retrieve the vial beside the centrifuge behind the mobile workstation" requires the agent to compose the same directional, depth, proximity, support, and occlusion relations evaluated by EgoSR-Bench.

This example illustrates how the benchmark can serve as a diagnostic layer between panoramic perception and downstream planning. A model that resolves the target and its spatial relations can provide a grounded object identity and geometric description to a navigation or manipulation module. The laboratory scene is a motivating use case rather than a separately evaluated deployment; validating closed-loop action, collision avoidance, and multi-step task completion remains future work.

7 Conclusion

We introduced EgoSR-Bench, a benchmark for multi-hop egocentric spatial reasoning from a single panoramic observation. It contains 3,577 instances from 15 photorealistic indoor scenes and five reasoning templates that compose directional, depth, proximity, vertical, occlusion, and reference relations. By grounding question generation and verification in aligned 3D geometry, EgoSR-Bench evaluates spatial reasoning beyond object recognition and single-step relation prediction.Current multimodal models remain challenged by this setting: coordinate prompts are often stronger than bounding-box prompts, but performance varies across models and input modes, and post-training does not yield a uniform gain. The release is limited to static single-view indoor panoramas and object-level spatial relations; it does not evaluate temporal observation, active viewpoint selection, closed-loop manipulation, or navigation. Future work will extend the benchmark to interactive settings, broader scenes and objects, and fully paired adaptation evaluations.

AI Use Statement. In this work, generative AI tools were used in three limited ways. First, AI assisted in producing some visual assets used in the figures. The authors designed the underlying scientific concepts, layouts, and visual narratives; AI-generated outputs were treated as drafts or components and were manually selected, edited, arranged, and checked by the authors.

Second, generative AI assisted in generating and organizing a subset of the data used in this study. The authors defined the data-generation procedure, reviewed the generated outputs, performed manual validation and correction, and remained responsible for the final data and all subsequent analyses.

Third, generative AI tools were used to improve the language and readability of an early draft of the manuscript. The authors subsequently revised and humanized the text and manually checked the technical content, equations, citations, figures, data descriptions, and scientific claims.

Generative AI tools were not used to determine the scientific conclusions or to replace author judgment. The authors reviewed all AI-assisted work and take full responsibility for the final content of this paper, including all text, data, figures, analyses, and claims produced with the assistance of generative AI.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision*, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. arXiv preprint arXiv:2401.12168, 2024.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*, 2019.

Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Min Sun Hsu, and Min Sun. Visual question answering on 360 degree images. In *WACV Workshops*, 2020.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018.

Zihao Dongfang, Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Danda Pani Paudel, Luc Van Gool, Kailun Yang, and Xuming Hu. Are multimodal large language models ready for omnidirectional spatial reasoning? arXiv preprint arXiv:2505.11907, 2025.

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.

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. arXiv preprint arXiv:2404.12390, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *CVPR*, 2021.

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. In *NeurIPS*, 2023.

Di Hu, Rui Qian, Min Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Pano-avqa: Grounded audio-visual question answering on 360 degree videos. In *ICCV*, 2021.

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

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *ICLR*, 2026.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, 2020.

Zekai Lin and Xu Zheng. Panoenv: Exploring 3d spatial intelligence in panoramic environments with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9647–9657, 2026.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Yang Liu, Huiyu Duan, Ran Tao, Juntao Cheng, Sijing Wu, Yunhao Li, Jing Liu, Xiongkuo Min, and Guangtao Zhai. Odi-bench: Can mllms understand immersive omnidirectional environments? In *International Conference on Learning Representations*, 2026.

Xiaojian Ma, Sicheng Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2023.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, 2024.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Ming Yan, Brian Budge, Yang Yan, Xuehan Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.

Cheng Sun, Ching-Yao Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *CVPR*, 2019.

Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021.

Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark. In *ECCV Workshops*, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

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. arXiv preprint arXiv:2412.14171, 2024.

Zhengyuan Yang et al. Vsi-bench: Evaluating visual spatial intelligence of multimodal large language models. arXiv preprint, 2025.

S1 Evaluation Audit Tables

This section defines the five input modes and gives diagnostic breakdowns for the main benchmark comparison.

Input and failure provenance limits. The archived predictions do not preserve enough run-specific information to verify every per-question image binding or serialized prompt. Two records share the same visible question and ordered options but have different targets; both remain in the raw release and are excluded from clean recomputation rather than being resolved retrospectively. The two records are excluded only when recomputing statistics on a clean subset; the released benchmark remains unchanged.

Post-training split and held-out test set. The post-training protocol uses one held-out test set of 857 questions: 20% of the split is held out for testing and 80% is used for training. The post-training results are aligned to this same test set after format-only suffixes are removed; answers are scored by candidate identity while rationale text is ignored. The 3,577-row full benchmark remains the primary benchmark population and is reported separately from this post-training split. No second test-set count is introduced, and no unsupported cross-manifest pairing or held-out claim is made beyond the specified split.

Reproducibility status. No new inference or training run was performed for this revision. Claims about common decoding defaults, checkpoint precision, paired adaptation effects, or held-out independence require the original run configurations and input manifests and are therefore outside the verified scope of this study.

Table S1: Intended input modes. Run-specific image bindings and coordinate outputs remain to be verified.

Mode	Visual Input	Textual Spatial Cue
Src	Raw ERP	None
Opt. BBox	Candidate boxes	None
Main BBox	Main-object box	None
Opt. Coord	Raw ERP	Candidate coordinates
Main Coord	Raw ERP	Main-object coordinates

Reading the table. These five configurations differ only in the auxiliary geometric cue paired with the same panoramic question and candidate set. Src is the reference condition; the box and coordinate variants expose either candidate-level or main-object geometry.

Table S2: Object attributes used to construct grounded references and reasoning constraints.

Attribute	Role in question generation
Direction	Observer-centered horizontal relation between objects
Depth	Near/far relation along the viewing direction
Height	Above, below, or eye-level relation
Proximity	Local adjacency or neighborhood constraint
Occlusion	Visibility state used in grounded descriptions
Color, shape, category	Semantic descriptors for object references and distractors
Coordinates and boxes	Geometric grounding and optional model input cues

Reading the table. These attributes construct and verify grounded references during dataset generation; they are not all provided as model inputs.

Reading the table. The table stratifies Src accuracy by scene category while retaining every evaluated row in the denominator. The five configurations have scene-level predictions; newer pooled models do not.

Reading the table. The table measures template-specific difficulty for the same five configurations. The main benchmark table reports pooled mode-level accuracy; no per-template values are inferred for models without prediction records.

Reading the table. Each row selects the configuration with the highest mean accuracy within its family or series. The values summarize the five input modes for that configuration.

Table S3: Src accuracy (%) by scene category for five fixed model configurations. Empty or unparseable responses remain in the denominator.

Scene category	N	Gemini-2.5-Pro	GPT-4.1-mini	InternVL3.5-14B	Qwen-VL-Max	Qwen3-VL-30B
frl_apartment	2349	8.73	22.35	32.31	19.33	29.12
office	443	7.00	11.96	34.76	20.54	36.34
large_apartment	174	1.72	28.74	33.33	8.62	31.03
room	464	0.00	23.49	30.39	28.88	36.42
hotel	147	0.00	12.93	27.21	31.97	31.29
Pooled	3,577	6.68	21.14	32.21	20.72	31.14

Table S4: Src accuracy (%) by reasoning template for five fixed model configurations.

Template	N	Gemini-2.5-Pro	GPT-4.1-mini	InternVL3.5-14B	Qwen-VL-Max	Qwen3-VL-30B
CSL	955	8.17	22.93	36.86	27.54	35.39
DDF	778	6.17	19.02	33.93	19.02	30.85
VPR	734	5.86	23.98	32.43	22.48	34.33
CCF	555	6.13	16.40	27.75	15.50	22.88
NRR	555	6.49	21.98	25.95	14.23	28.29
Pooled	3,577	6.68	21.14	32.21	20.72	31.14

Table S5: Highest-mean configuration for each model family and series. For each series, the selected row has the highest mean accuracy in Table 2.

Family	Series	Best configuration	Src	Opt. BBox	Main BBox	Opt. Coord	Main Coord	Mean
GPT	GPT Official source	GPT-6-Astra	37.35	35.20	37.29	43.70	37.85	38.28
Gemini	Gemini Official source	Gemini-3.7-Flash	37.13	30.58	37.10	43.25	35.87	36.79
Qwen	Qwen Official source	Qwen3.8-Max	35.64	33.69	34.30	40.76	35.03	35.88
Qwen	Qwen local	Qwen3-VL-30B (BF16)	31.14	29.89	32.12	34.64	33.46	32.25
InternVL	InternVL local	InternVL3.5-14B (BF16)	32.21	32.35	31.56	32.51	33.02	32.33
Other	Other local	Gemma-4-31B	34.58	33.07	35.92	37.77	34.44	35.16
Other	Other local	MiniCPM-V-4.5	32.37	30.78	33.77	36.43	31.00	32.87
Other	Other local	Kimi-VL-A3B	32.63	30.56	31.56	32.88	30.25	31.57
Other	Other local	Llama-4-Scout-17B	27.34	27.57	26.78	38.80	27.48	29.59
Other	Other local	GLM-4.1V-9B	28.35	27.82	29.27	31.31	28.26	29.00

S2 Post-training Details

The main-text GRPO table reports task-level accuracy for Src, Opt. BBox, Main BBox, Opt. Coord, and Main Coord. Each cell is scored with the parser for its task format. All evaluations use the same 857-question held-out test set (20% test, 80% training); missing or invalid responses reduce the valid denominator. Parser implementations and checkpoint settings are not available for independent reproduction.

S3 Dataset Statistics

Reading the table. The counts are recomputed from the released 3,577-row manifest. Every row is assigned to one scene and one of the five paper-level templates, so the totals provide a direct check on manifest coverage and on the distribution used by the benchmark.

S4 Diagnostic Human Evaluation

Human-oriented grounding aids. For human participants, bounding boxes with object labels serve as a human-oriented grounding aid, functionally analogous to the coordinate cues provided to models. These representations share the goal of facilitating object localization and reference resolution, but are not treated as equivalent experimental conditions. In particular, candidate boxes with labels support candidate localization, corresponding in purpose to candidate-coordinate cues; a labeled main-reference-object box supports reference-object localization, corresponding in purpose to main-object-coordinate cues. The combined condition marks both the main reference object and the candidate objects, and thus has no direct counterpart among the two separate coordinate-cue conditions. This correspondence concerns the intended role of the information, rather than an empirically established equivalence in its content or usability.

Table S6: Instances across all 15 scenes, recomputed from the current release and matched to the evaluation manifest. The columns sum to 955, 778, 734, 555, and 555, with total $N = 3,577$. Percentages are rounded independently.

Scene	CSL	DDF	VPR	CCF	NRR	Total	Share (%)
frl_apartment_0	165	122	102	83	83	555	15.5
frl_apartment_2	151	106	102	70	70	499	14.0
frl_apartment_3	150	98	86	70	70	474	13.3
frl_apartment_4	101	100	90	69	69	429	12.0
frl_apartment_5	96	84	86	63	63	392	11.0
hotel_0	45	30	24	24	24	147	4.1
large_apartment_1	30	30	24	22	22	128	3.6
large_apartment_2	0	14	14	9	9	46	1.3
office_0	27	20	22	15	15	99	2.8
office_2	21	20	16	13	13	83	2.3
office_3	42	30	30	24	24	150	4.2
office_4	25	28	24	17	17	111	3.1
room_0	64	50	40	33	33	220	6.2
room_1	38	42	48	30	30	188	5.3
room_2	0	4	26	13	13	56	1.6
Total	955	778	734	555	555	3,577	100.0

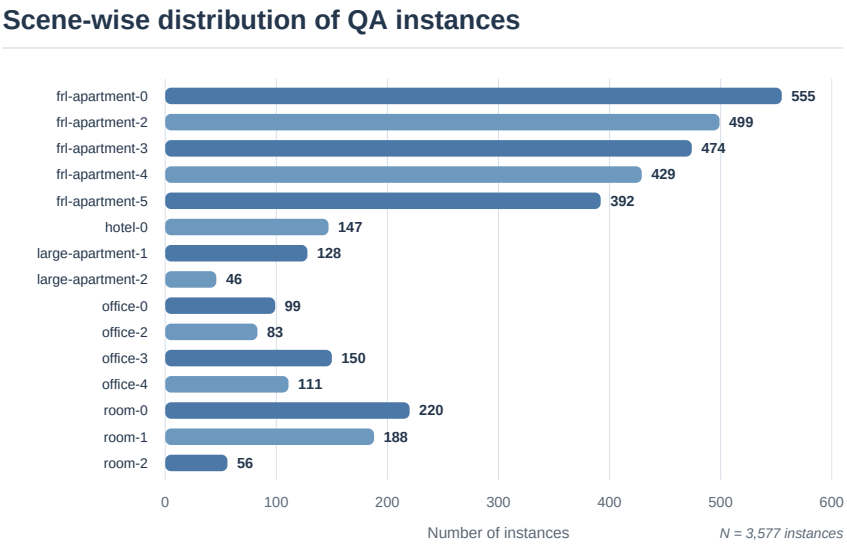


Figure S1: Number of instances in each scene, recomputed from the same current manifest as Table S6. All 15 scene identifiers are shown. This distribution is provided for completeness and does not establish the cause of differences across scenes.

Presentation conditions. All five conditions retain the full panoramic image. Presentations 1–10 use the unannotated image; 11–20 add candidate-object boxes without labels; 21–30 add candidate-object boxes with labels; 31–40 add labeled boxes for both the main reference object and the candidate objects; and 41–50 add a labeled box only for the main reference object. These human-questionnaire conditions are described functionally so that they are not confused with the five standardized model input modes.

Protocol and results. We analyze responses from 12 graduate students: six with a computer science background and six working on spatial and visual topics in art and design. Participants were compensated for their time. Six participants received a pre-task introduction to the panoramic setting, while the other six received no such introduction before answering. The available aggregate results do not separate responses by background or introduction status, so we do not estimate either subgroup effect or assume that these two partitions coincide. All participants answer a shared questionnaire containing 50 question presentations, with 10 presentations assigned to each of five

input modes. The questionnaire contains 46 distinct underlying questions, with four questions repeated across modes (presentation pairs 5/25, 12/23, 16/29, and 17/22). Responses are scored by exact agreement with the recorded answer key. Across 600 participant–question responses, 208 are correct, giving a participant-level mean accuracy of 34.67% with a sample standard deviation of 5.48 percentage points (range: 28–48%). Because the questionnaire mixes 42 four-option and eight two-option items, uniform random guessing has an expected accuracy of 29.00%. The difference of 5.67 percentage points is descriptive; no significance test against this chance expectation is reported. The 600 responses are repeated observations from 12 participants, not 600 independent participants.

Table S7: Descriptive human results on the diagnostic questionnaire. Accuracy is participant-level mean $\pm$ sample standard deviation (%); each mode has 10 presentations per participant. Chance denotes the expected accuracy of uniform guessing among the available options. Different modes use different question sets, so their differences do not isolate input-mode effects.

Input condition	**Accuracy (%)**	**Chance (%)**
Src	29.17 ± 19.29	30.00
Opt. BBox	42.50 ± 13.57	30.00
Opt. BBox + Name	34.17 ± 5.15	30.00
Main BBox + Name	47.50 ± 15.45	30.00
Main Object	20.00 ± 12.06	25.00
Overall	34.67 ± 5.48	29.00

Reading the table. The entries are participant-level descriptive means from the diagnostic questionnaire. Each input mode uses ten presentations per participant, and the chance column reflects the option structure of that mode. Because question content differs across modes, these values are not a matched causal ablation.

Macro benchmark input-mode distribution. To summarize the five model input modes, we select one representative configuration from each of five groups: GPT, Gemini, Qwen, InternVL, and Other. Table S8 reports the mean and range across these representatives. This comparison describes the input-mode pattern and is not a causal ablation.

Table S8: Mean and range of accuracy across the five benchmark input modes. Statistics are computed from five representative family configurations in Table 2.

Input mode	**Mean accuracy (%)**	**Range (%)**
Src	35.38	32.21–37.35
Opt. BBox	32.98	30.58–35.20
Main BBox	35.23	31.56–37.29
Opt. Coord	39.60	37.77–43.70
Main Coord	35.24	33.02–37.85

Reading the table. Each row averages the corresponding input-mode score across five representative family configurations copied from Table 2; the range records the minimum and maximum among those representatives. This compact summary describes the overall pattern without creating a new model ranking.

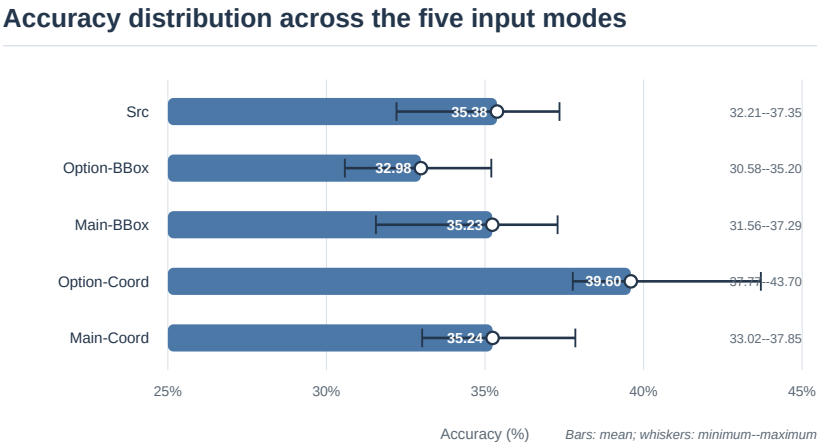


Figure S2: Mean accuracy across the five input modes. Bars show the mean across five representative model-family configurations; whiskers show the minimum–maximum range.

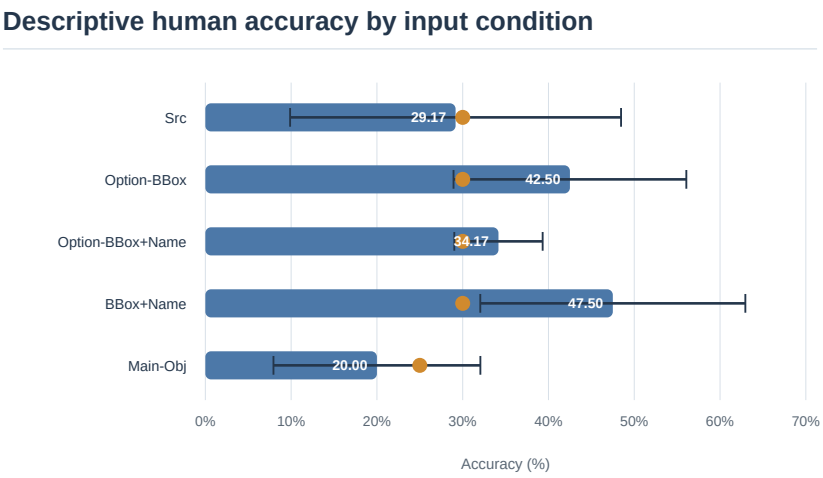


Figure S3: Descriptive human accuracy by questionnaire condition. Bars show participant-level means, whiskers show the corresponding sample standard deviation across participants, and dots show the expected accuracy under uniform guessing. The conditions use different question sets and are therefore not an input-only ablation.

Macro model context, not a human–model comparison. Table S9 gives model-score context for the input-mode summary. Each row corresponds to a configuration in Table 2; the table is neither a model ranking nor a human ceiling. The full benchmark has 3,577 entries per mode, the human questionnaire has 10 presentations per mode per participant, and the post-training split has one 857-question held-out test set.

Reading the table. These rows provide model-score context for the human-facing input conditions; they are not human baselines or matched human–model comparisons.

Failure taxonomy and withheld claims. Archived aggregate error summaries do not reliably separate request failures, empty or malformed outputs, parser failures, and incorrect predictions. We therefore retain unsuccessful outputs in the reported denominators and do not interpret an aggregate error rate as a reasoning-error rate. Model-level reasoning rankings and causal explanations of failure require the original request logs, serialized inputs, and per-question predictions, which are not available for every run.

Table S9: Representative model accuracy by input mode. Values are taken from Table 2 and are not human-study scores.

Configuration	Src	Opt. BBox	Main BBox	Opt. Coord	Main Coord
Gemma-4-31B	34.58	33.07	35.92	37.77	34.44
MiniCPM-V-4.5	32.37	30.78	33.77	36.43	31.00
InternVL3.5-14B	32.21	32.35	31.56	32.51	33.02
Qwen3-VL-30B	31.14	29.89	32.12	34.64	33.46
Kimi-VL-A3B	32.63	30.56	31.56	32.88	30.25
Llama-4-Scout-17B-16E	27.34	27.57	26.78	38.80	27.48
GLM-4.1V-9B	28.35	27.82	29.27	31.31	28.26
GPT-4.1-mini	35.56	31.70	35.53	39.53	34.61
Gemini-2.5-Pro	34.58	31.98	36.32	35.50	33.86

Interpretation and limitations. Table S7 reports the observed accuracy for each mode-specific question set. The combined labeled-box condition has the highest observed accuracy and the main-object-only condition the lowest. However, the question sets differ in content and template composition. In the template-balanced audit, the unannotated condition contains eight DDF and two VPR presentations; both candidate-box conditions contain ten VPR presentations each; the combined labeled-box condition contains ten DDF presentations; and the main-object-only condition contains ten CCF presentations. This limited coverage does not represent all benchmark reasoning templates. Mode, template, and question content are consequently confounded, so the observed differences cannot be attributed solely to the visual input mode.

We report the study as a small descriptive human baseline, rather than a human performance ceiling or a matched comparison with the full model evaluation. The modest overall accuracy motivates further examination of question clarity and answer grounding under the human-facing presentation protocol. The audit checked answer indices, presentation identifiers, choices, scores, template labels, repeated questions, and questionnaire text; it did not independently validate geometric ground truth, every rendered image annotation, or every manual transcription of individual responses. Aggregate summaries without independent per-model predictions are not used as matched model baselines.

S5 Common-Question Comparisons

We align the base and GRPO results by model, input mode, and question index. The base side contains 857 indexed questions; missing or unparsable GRPO responses can reduce the valid denominator. We therefore report accuracy differences as descriptive comparisons rather than paired causal training effects.

Table S10: Accuracy differences between the base and GRPO models on the common question index set. Each cell reports GRPO minus base accuracy in percentage points. The base set contains 857 indexed questions; the valid GRPO denominator can be lower when responses are missing or unparsable.

Model	**Variant**	**Src**	**Main BBox**	**Opt. BBox**	**Main Coord**	**Opt. Coord**
InternVL3.5-2B	Q8	+8.81	+12.18	+13.62	+45.39	+38.74
InternVL3.5-2B	BF16	+8.81	+12.18	+13.97	+45.51	+39.09
InternVL3.5-8B	Q8	+7.99	+6.08	+7.11	+43.06	+34.31
InternVL3.5-8B	F16	+7.99	+6.20	+7.23	+42.82	+33.72
InternVL3.5-14B	Default	+6.24	+6.66	+9.35	+30.92	+28.70
Qwen2-VL-2B	Default	+7.03	+7.29	+15.76	+49.24	+48.19
Qwen2.5-VL-7B-Instruct	Default	+2.51	+7.27	+11.34	–	+28.12
Qwen3-VL-4B-Instruct	Default	+5.00	+10.10	+16.74	+33.37	+36.64
Qwen3-VL-8B-Instruct	Default	+31.62	+14.56	+20.82	+32.69	+36.95

Reading the table. The comparison is aligned by model, input mode, and indexed question set. It does not recover paired per-question outcomes when valid-response denominators differ.

Descriptive GRPO delta on common question indexes

	Src	Main-BBox	Option-BBox	Main-Coord	Option-Coord
InternVL3.5-2B (Q8)	+8.81	+12.18	+13.62	+45.39	+38.74
InternVL3.5-2B (BF16)	+8.81	+12.18	+13.97	+45.51	+39.09
InternVL3.5-8B (Q8)	+7.99	+6.08	+7.11	+43.06	+34.31
InternVL3.5-8B (F16)	+7.99	+6.20	+7.23	+42.82	+33.72
InternVL3.5-14B (Default)	+6.24	+6.66	+9.35	+30.92	+28.70
Qwen2-VL-2B (Default)	+7.03	+7.29	+15.76	+49.24	+48.19
Qwen2.5-VL-7B-Instruct (Default)	+2.51	+7.27	+11.34	--	+28.12
Qwen3-VL-4B-Instruct (Default)	+5.00	+10.10	+16.74	+33.37	+36.64
Qwen3-VL-8B-Instruct (Default)	+31.62	+14.56	+20.82	+32.69	+36.95

Cell values: GRPO accuracy minus baseline accuracy (percentage points)

Figure S4: GRPO–base accuracy differences on the common question index set. Each cell reports the difference in percentage points; valid-response denominators may differ across input modes.

S6 Application Illustration

Figure S5 provides a motivating embodied-laboratory use case for the panoramic perception and spatial-grounding capability evaluated by EgoSR-Bench. The illustration is conceptual and does not represent a separately evaluated deployment.



Figure S5: Motivating application in an embodied autonomous chemistry laboratory. The benchmark evaluates the panoramic perception and spatial grounding layer that can support subsequent laboratory navigation and manipulation.