SIGHT-BENCH: DIAGNOSING REFERENCE-FRAME CONSISTENCY AND GLOBAL SPATIAL REASONING IN MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce SIGHT-Bench, a diagnostic benchmark for spatial reasoning in real-world elevated outdoor scenes. It draws on stationary camera installations whose positions remain fixed while viewing orientations may vary. SIGHT-Bench contains 1,000 four-choice questions from 400 cameras, organized into five dimensions and 12 subtasks spanning geometry, scene layout, viewpoint changes, world-frame motion, and topology. We evaluate proprietary, general-purpose open, and spatially specialized multimodal models against a human reference. Even the strongest evaluated models retain substantial overall gaps to human performance, and those gaps vary across spatial dimensions. Across model configurations, spatial consistency is more reliable than direct connectivity. Additional thinking yields marked gains in world-frame motion but much smaller gains elsewhere. Subtask analyses further show that improvements in one spatial relation need not extend to related judgments. These results motivate treating spatial reasoning as a profile of distinct capabilities rather than a single aggregate score.

1 INTRODUCTION

Elevated outdoor cameras capture roads, buildings, and moving objects within the same scene. Spatial understanding in this setting requires more than recognition: models must compare spatial relations, align observations across views, interpret motion in a common reference frame, and reason about connectivity. These operations depend on different spatial representations and need not be equally reliable (Klatzky, 1998; Li et al., 2025a; Zhang et al., 2025; 2026a). We therefore treat outdoor spatial reasoning as a set of related but separable capabilities.

Existing benchmarks evaluate fine-grained spatial relations, multi-image and video understanding, and reasoning across viewpoints (Ma et al., 2024; Xu et al., 2025; Yang et al., 2025; 2024; Li et al., 2025a; Zhang et al., 2025; Wang et al., 2025; Yeh et al., 2025; Zhang et al., 2026a). Others focus on top-down imagery, UAV observations, heterogeneous air–ground views, and cross-view urban reasoning (Hou et al., 2025b; Zhang et al., 2026b; Li et al., 2026a; Xu et al., 2026a). These benchmarks cover a wide range of spatial settings, but aggregate accuracy alone does not show which spatial judgments a model handles well or poorly. Our goal is therefore diagnostic: to measure how performance changes across spatial judgments within a common outdoor observation setting.

We study this problem using real-world *elevated outdoor observations from stationary camera installations*. The camera position remains fixed at a site while its viewing direction may change. This allows viewpoint-dependent relations to be studied without introducing camera translation. The same setting also preserves scene structure across observations and supports reasoning about motion in a geographic frame. It therefore complements mobile aerial platforms and cross-platform views (Zhang et al., 2026b; Li et al., 2026a; Xu et al., 2026a), while allowing geometric, cross-view, motion, and topological judgments to be evaluated in the same environment.

We introduce **SIGHT-Bench**, which organizes elevated outdoor spatial reasoning into five dimensions: geometric comparison, scene configuration, viewpoint analysis, world-frame motion, and topological navigation (Figure 1). The dimensions correspond to different spatial operations rather than levels of difficulty. Each dimension is further divided into subtasks that isolate more specific spatial judgments.

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: **SIGHT-Bench: five dimensions of outdoor spatial reasoning.** Stationary elevated-camera observations support evaluations of geometry, scene configuration, viewpoint, world-frame motion, and topology across complementary spatial tasks.

Our evaluation shows that current multimodal models remain uneven in their spatial reasoning capabilities. Performance varies across dimensions and reasoning settings, suggesting that progress in outdoor spatial intelligence should be assessed through a multidimensional capability profile that captures different forms of spatial reasoning rather than through a single aggregate score alone.

2 Related Work

2.1 Spatial Intelligence Benchmarks

Spatial relations and multi-view understanding. Recent benchmarks have progressively broadened the evaluation of spatial intelligence in multimodal models, from localized spatial relations to more comprehensive reasoning across views. Early efforts such as 3DSRBench focus on three-dimensional spatial relations, while SpatialBench organizes spatial cognition into a broader capability-oriented framework (Ma et al., 2024; Xu et al., 2025). Beyond individual-view evaluation, MMSI-Bench and VSI-Bench extend evaluation to multiple images and video observations, requiring models to integrate spatial information across observations (Yang et al., 2025; 2024). A related line of work investigates viewpoint variation and perspective reasoning. ViewSpatial-Bench and All-Angles evaluate multi-perspective understanding, SpinBench studies rotation-based reasoning, MindCube examines spatial mental modeling from limited views, and MultiView-Bench explores world-centric integration across observations (Li et al., 2025a; Yeh et al., 2025; Zhang et al., 2025; Wang et al., 2025; Zhang et al., 2026a).

Outdoor and aerial spatial understanding. Recent benchmarks have further extended spatial reasoning evaluation to large-scale outdoor environments and embodied observation settings. TDBench studies top-down imagery, while SpatialSky-Bench and SpatialUAV investigate spatial understanding and navigation from UAV observations. AirGroundBench considers heterogeneous air–ground views, and CityCube evaluates cross-view reasoning in urban environments (Hou et al., 2025a; Zhang et al., 2026d;c; Li et al., 2026b; Xu et al., 2026b). Different from these settings, SIGHT-Bench studies spatial reasoning from stationary elevated cameras, where the observation location remains fixed while viewing directions vary. This setting enables controlled evaluation of geometry, scene configuration, viewpoint changes, world-frame motion, and topology within a shared outdoor environment.

Table 1: **Comparison with related spatial-intelligence benchmarks.**

Benchmark	Elevated / Aerial	Multi-view	Orient. Variation	Fixed Center	Frame Trans.	Global Integration	Topology / Route
VSI-Bench (Yang et al., 2024)	×	✓	✓	×	✓	✓	✓
MMSI-Bench (Yang et al., 2025)	×	✓	✓	×	×	×	×
All-Angles (Yeh et al., 2025)	×	✓	✓	×	×	×	×
SpinBench (Zhang et al., 2025)	×	✓	✓	×	×	×	×
MindCube (Wang et al., 2025)	×	✓	✓	×	×	✓	×
MultiView-Bench (Zhang et al., 2026a)	×	✓	✓	×	✓	✓	×
TDBench (Hou et al., 2025a)	✓	×	✓	×	×	×	×
SpatialSky-Bench (Zhang et al., 2026d)	✓	×	×	×	×	×	×
SpatialUAV (Zhang et al., 2026c)	✓	✓	✓	×	✓	×	×
AirGroundBench (Li et al., 2026b)	✓	✓	✓	×	✓	×	×
CityCube (Xu et al., 2026b)	✓	✓	✓	×	×	×	×
SIGHT-Bench	✓	✓	✓	✓	✓	✓	✓

✓ denotes an explicit setting or evaluation target; × otherwise. Multi-view includes sequential video views; Frame Trans. includes egocentric–allocentric transformations. Topology / Route denotes either connectivity or route tasks, rather than requiring both.

2.2 Spatial Reasoning Enhancement in Multimodal Models

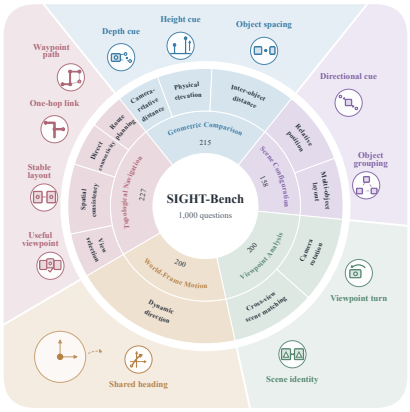
Recent efforts have improved spatial reasoning in multimodal models through spatial supervision, grounded representations, and specialized training strategies. One line of work introduces spatially grounded signals: SpatialVLM adopts spatially focused supervision, SpatialRGPT grounds model responses in spatial evidence, and SpatialLadder explores progressive training over spatial tasks (Chen et al., 2024; Cheng et al., 2024; Li et al., 2025b). Another line enhances spatial intelligence through large-scale spatial training and optimization. Spatial-SSRL introduces self-supervised spatial signals into reinforcement learning, while SenseNova-SI scales spatial training data across capabilities including metric reasoning, spatial relations, perspective taking, and mental reconstruction (Liu et al., 2025; Cai et al., 2025). While these approaches improve model capabilities through training, SIGHT-Bench provides a diagnostic framework for analyzing capability-specific spatial reasoning.

3 SIGHT-Bench: A Five-Dimensional Diagnostic Design

SIGHT-Bench uses real-world outdoor imagery from 400 cameras at elevated, stationary installations. Heights and installation configurations vary across sites. Within each site, viewing orientations may change while the installation position remains fixed. Questions use a single observation, multiple orientations, or temporally paired frames as required by the task. The setting reduces variation due to camera travel; it does not remove occlusion, appearance changes, object motion, or other real-world ambiguity. The distinction between viewpoint-dependent and world-referenced relations is related to established distinctions between egocentric and allocentric spatial representations and to recent multi-view spatial benchmarks (Klatzky, 1998; Yang et al., 2024; Zhang et al., 2026a).

The benchmark separates five dimensions by the primary spatial operation required to answer a question (Figure 2). **GC** concerns geometric comparisons, whereas **SC** concerns relations and arrangements among objects. **VA** asks how observations relate across viewing orientations; **WF** instead asks how observed object motion relates to a geographic direction. **TN** concerns evidence for connectivity, relational consistency, and routes. Following the diagnostic perspective of prior visual-reasoning benchmarks, the dimensions are intended to expose capability-specific strengths and failures rather than collapse them into a single aggregate measure (Johnson et al., 2017; Hudson & Manning, 2019). The five dimensions are defined by the primary spatial operation required by each question, and the subtasks further separate distinct relation types within those dimensions.

Naming and interpretation. View Selection is the functional name for the TN questions originally labeled Adjacent View Matching. The geometric subtasks measure relative judgments rather than calibrated metric depth. Representative items for all dimensions and subtasks, including visual options and correct answers, are provided in Appendix A.1.



Subtask	N	Share (%)
GC Geometric Comparison		
Camera-relative distance	51	5.1%
Physical elevation	64	6.4%
Inter-object distance	100	10.0%
SC Scene Configuration		
Relative position	86	8.6%
Multi-object layout	72	7.2%
VA Viewpoint Analysis		
Camera rotation	100	10.0%
Cross-view scene matching	100	10.0%
WF World-Frame Motion		
Dynamic direction	200	20.0%
TN Topological Navigation		
View selection	52	5.2%
Spatial consistency	82	8.2%
Direct connectivity	51	5.1%
Route planning	42	4.2%

Figure 2: **Five complementary diagnostic dimensions of SIGHT-Bench.** The benchmark contains 1,000 questions across five dimensions and 12 subtasks. The wheel summarizes the five dimensions and their question counts, while the accompanying panel lists the constituent subtasks. Sector sizes are proportional to the number of questions in each dimension and do not indicate a difficulty hierarchy. View Selection corresponds to Adjacent View Matching in the original task taxonomy.

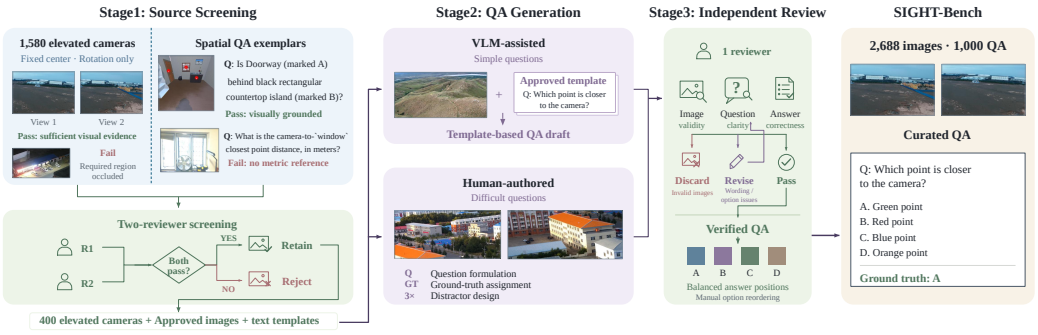


Figure 3: **Construction workflow.** Camera and observation screening, task-oriented question construction, and independent verification. We screened 1,580 candidate elevated cameras and retained 400 (25.3%). The retained cameras provided 2,688 images for question construction.

3.1 Question Construction and Verification

Task-oriented construction. We select observations from the elevated stationary-camera setting and construct questions around the spatial operations summarized in Figure 2. Straightforward questions were drafted with VLM assistance. Annotators drafted or revised questions involving multiple images or more complex spatial reasoning. Each question has exactly one correct option and three distractors. Distractors represent plausible competing directions, configurations, correspondences, or connectivity and route interpretations. Controlled question construction and capability-specific evaluation have also been used in prior visual-reasoning and multimodal benchmarks (Hudson & Manning, 2019; Li et al., 2024). Reference images and image-based options are explicitly associated with their roles, allowing both textual and visual choices within the same format. Figure 3 summarizes the complete construction and verification workflow.

Independent verification. An annotator who did not construct the item checks the correctness and uniqueness of its answer, clarity of its wording, and sufficiency of the provided visual evidence. Ambiguous questions, unintended textual cues, insufficient evidence, or multiple defensible answers trigger revision or removal. This emphasis on visually grounded questions and manual quality control follows longstanding concerns in visual question answering about language priors and insufficient dependence on visual evidence (Goyal et al., 2017; Li et al., 2024). Reviewers also

check that distractors are plausible but incorrect. We inspect correct-option positions and maintain an approximately balanced distribution across the four choices. All 1,000 final questions underwent this independent review across the five dimensions and 12 subtasks.

Diagnostic interpretation. Dimension-level accuracy measures coverage across spatial capabilities, while subtask-level accuracy reveals differences within a dimension. Paired answer changes distinguish corrections from newly introduced errors. The subtasks contain different instances and answer choices, so we compare their scores as capability profiles and use paired analyses when responses to the same questions are available for direct comparison.

4 EVALUATION

4.1 EXPERIMENTAL SETUP

SIGHT-Bench contains 1,000 four-choice questions across five dimensions: Geometric Comparison (GC; 215), Scene Configuration (SC; 158), Viewpoint Analysis (VA; 200), World-Frame Motion (WF; 200), and Topological Navigation (TN; 227). We evaluate general-purpose open models, spatially specialized checkpoints, and proprietary API models with VLMEvalKit (Duan et al., 2024). Each model receives the finalized question, its four choices, and the associated images, including image-based choices where applicable. Proprietary models are queried through official APIs.

We report the accuracy and answer-extraction results recorded in the evaluation outputs, counting responses without a valid final option as incorrect. We use *thinking* and *no-reasoning* for the two recorded inference configurations. For mixture-of-experts (MoE) models, we report total and active parameter counts when available, since only a subset of experts is activated for each input (Fedus et al., 2022). Parameter count describes model architecture here, not inference compute.

In addition to model results, we include uniform random guessing, a Qwen3-VL-8B text-only control, and a human reference. We use the text-only setting to check how much performance remains when the benchmark images are removed, following prior concerns that VQA systems can sometimes exploit linguistic regularities without relying sufficiently on visual evidence (Goyal et al., 2017). Five volunteers independently answered all 1,000 questions using the same observations and answer choices. None took part in benchmark construction or had professional or research experience related to the task domain. We averaged their scores within each category. For human and text-only baselines, Overall is reconstructed from available category summaries using the benchmark category sizes reported in Table 2 for all five dimensions.

4.2 FIVE-DIMENSIONAL PERFORMANCE

Table 2 shows substantial variation in the five dimensions, which is not apparent from the overall accuracy alone. The recorded GPT thinking configuration reaches 71.8% Overall and 96.5% on WF, but only 63.26% on GC and 45.57% on SC. For comparison, the human reference scores 98.1% on GC and 81.8% on SC. We therefore see a model that performs very strongly on world-frame direction while remaining much less reliable on geometric and scene-level judgments. The human reference is not uniformly higher on every dimension: its WF score is 95.0%.

Visual-input removal. Qwen3-VL-8B scores 49.5% with images and a reconstructed 29.30% in the text-only setting. Removing the images lowers performance on GC, SC, VA, and TN, whereas WF changes from 23.0% to 26.0%. The large Overall drop suggests that the benchmark depends substantially on visual input, although the effect is not identical across dimensions. This control is useful in light of known language-prior effects in visual question answering (Goyal et al., 2017).

Because removing the images also removes the content of image-based answer options, the text-only setting does not preserve the full candidate information available in the multimodal condition. Its Overall score is reconstructed from category-level summaries. Removing the images also removes the content of image-based answer options, so the text-only setting does not preserve the full candidate information available in the multimodal condition. Its Overall score is also approximate because we only have category-level summaries and cannot verify the original denominators.

Table 2: **Main results across the five diagnostic dimensions.** Accuracy (%). Within each category, the model with the highest Overall accuracy is highlighted.

Model / configuration	Overall	GC	SC	VA	WF	TN
Human and controls						
Human	**90.09**	98.10	81.80	91.50	95.00	82.70
Random	25.00	25.00	25.00	25.00	25.00	25.00
Text-only (Qwen3-VL-8B)	29.30	26.51	29.17	26.50	26.00	37.42
Proprietary MLLMs						
GPT-5.5 (thinking)	**71.80**	63.26	45.57	77.00	96.50	71.81
GPT-5.5 (no-reasoning)	55.90	56.74	38.61	75.00	34.00	69.60
Claude Opus 4.8	53.30	48.84	38.61	62.50	46.50	65.64
Open MLLMs						
Qwen3-VL-235B-A22B	**52.10**	56.74	41.14	69.00	24.00	65.20
InternVL3.5-241B-A28B	48.20	50.23	41.77	68.00	21.00	57.27
Qwen3-VL-30B-A3B	48.70	49.77	44.30	65.00	23.00	59.03
InternVL3.5-30B-A3B	44.10	41.40	37.34	67.00	18.50	53.74
Kimi-VL-A3B-Instruct	42.50	38.14	41.14	52.50	31.00	48.90
GLM-4.6V-Flash (no-thinking)	42.50	43.72	41.77	54.50	21.50	49.78
InternVL3-8B	41.50	35.81	42.41	58.00	20.50	50.22
InternVL3.5-8B	45.90	39.53	43.67	64.50	26.00	54.63
MiMo-VL-7B-RL	43.10	42.79	37.34	55.50	27.50	50.22
MiMo-VL-7B-SFT	43.70	42.33	41.14	57.50	25.00	51.10
Qwen3-VL-8B	49.50	49.77	46.20	69.50	23.00	57.27
Qwen3-VL-2B	32.90	36.28	41.14	24.50	18.50	44.05
InternVL3.5-2B	39.20	36.74	50.63	50.50	22.50	38.33
Gemma 4 E4B	36.30	34.88	51.90	36.00	26.50	35.68
Spatially enhanced models						
SenseNova-SI (Qwen3-VL-8B)	**52.30**	48.37	49.37	75.50	30.00	57.27
SenseNova-SI (InternVL3-8B)	47.80	46.98	51.27	73.50	19.00	48.90
Spatial-SSRL-7B	42.80	42.79	48.10	57.50	21.50	44.93

‡ Overall is reconstructed from rounded category summaries; Human Overall is approximately 90.09%. B/A indicate total/active MoE parameters where specified, not inference compute.

4.3 Human–Model Gaps Across Spatial Dimensions

Figure 4a compares the human reference with the highest-Overall proprietary, general-purpose open, and spatially specialized models. Their Overall accuracies are 71.8%, 52.1%, and 52.3%, respectively, compared with approximately 90.1% for humans. Thus, the strongest evaluated model still trails the human reference by 18.3 percentage points Overall, while the leading open and specialized models remain about 38 points behind. The specialized model narrowly exceeds the leading general-purpose open model Overall, yet both retain large gaps to the human reference across the spatial judgments evaluated in this outdoor benchmark.

The gap is not uniform across dimensions. GPT-5.5 with thinking reaches 96.5% on WF, slightly above the human reference of 95.0%, yet trails humans by 34.8 points on GC and 36.2 points on SC. The open and specialized models also show distinct dimension profiles despite their similar Overall scores. These comparisons show that a high aggregate score, or parity on one spatial dimension, does not imply broadly reliable outdoor spatial reasoning.

4.4 Spatial Consistency and Direct Connectivity

Figure 4b shows a common topological pattern across seven model configurations: accuracy on Spatial Consistency exceeds accuracy on Direct Connectivity in every plotted comparison. Qwen3-

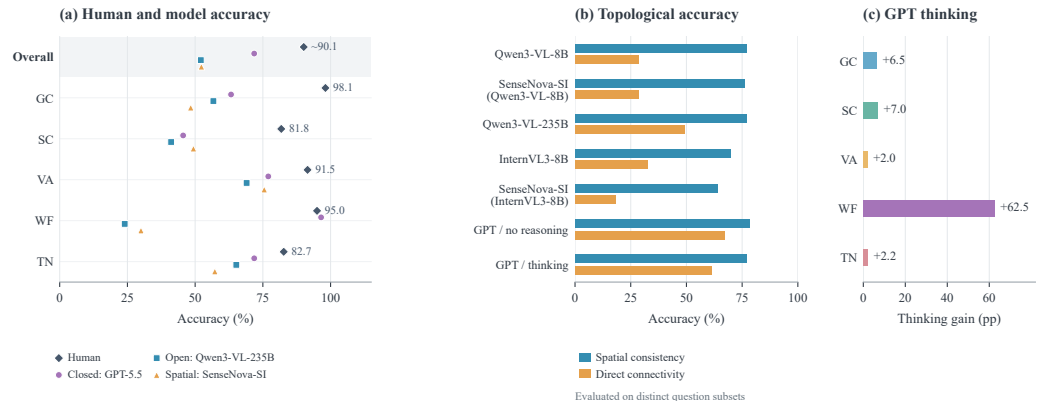


Figure 4: **Spatial performance remains uneven across model groups and reasoning settings.** (a) Human accuracy and the highest-Overall proprietary, general-purpose open, and spatially specialized models. Human Overall is reconstructed from category summaries. (b) Spatial Consistency and Direct Connectivity accuracy for seven configurations on distinct question subsets. (c) Percentage-point accuracy gains from recorded GPT thinking.

VL-8B, for example, scores 77.1% on consistency but only 28.6% on connectivity. The GPT thinking configuration narrows this difference, yet its corresponding scores remain 77.1% and 61.2%. The pattern spans model families and inference settings, indicating that coherent scene judgments need not be matched by equally reliable decisions about road connections.

These subtasks contain different questions and should not be interpreted as successive stages of a single reasoning process. Their score differences identify a recurring evaluation gap rather than establish where any one model's inference fails. We analyze the related route and connectivity subtasks more closely in Section 5.2, where the fine-grained results show how a similar Overall score can conceal different topological strengths and weaknesses.

4.5 Thinking: Gains Concentrated in WF

We compare two recorded inference configurations of the same GPT model on the same SIGHT-Bench questions. Both receive the finalized benchmark items, and their responses are paired by question to isolate the observed configuration contrast. The paired design also allows us to distinguish newly corrected answers from newly introduced errors.

The thinking configuration raises Overall accuracy from 55.9% to 71.8%, a gain of 15.9 points (Figure 4c). Most of that improvement comes from WF. Of the 159 net additional correct answers, 125 are WF questions, accounting for 78.6% of the total net gain. WF itself increases by 62.5 points, whereas GC, SC, VA, and TN increase by approximately 6.5, 7.0, 2.0, and 2.2 points, respectively, when computed from paired item counts.

Taken together, Figure 4 shows substantial human–model gaps, a recurring difference between topological judgments, and gains from thinking that concentrate in one dimension. These patterns motivate the subtask analysis that follows, where we examine which geometric and topological judgments account for the broader dimension-level results.

5 Diagnostic Analysis

The five-dimensional evaluation shows that improvements are uneven across spatial capabilities. We therefore look more closely at GC and TN, where the aggregate scores hide substantial variation among subtasks, and examine how GPT predictions change between the thinking and no-reasoning settings. This type of fine-grained decomposition has been used in diagnostic benchmarks to reveal failure modes that are difficult to see from aggregate accuracy alone (Johnson et al., 2017; Hudson

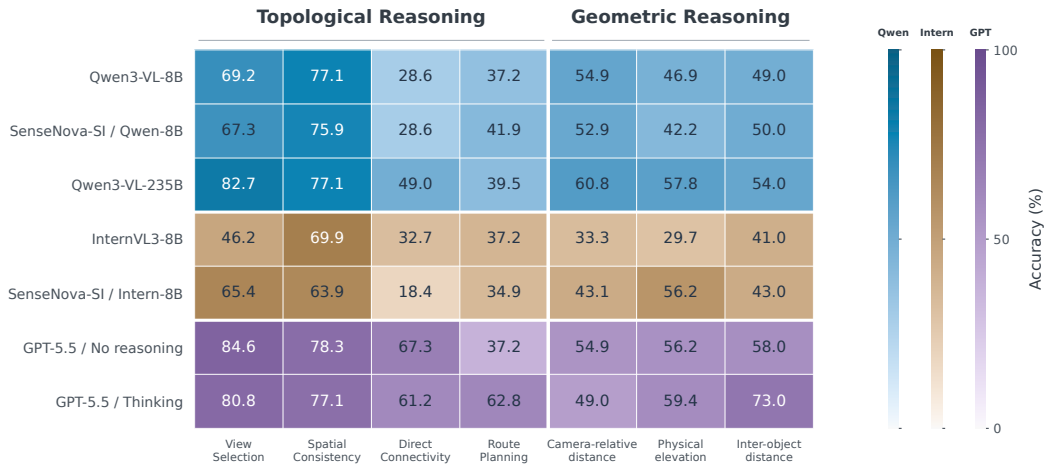


Figure 5: **Subtask-level performance in GC and TN.** Cells show accuracy (%). The three family color scales share a 0–100 range. View Selection refers to the evidence-selection questions originally named Adjacent View Matching. The subsets contain different instances and should not be interpreted as successive stages of a single reasoning pipeline.

& Manning, 2019; Ribeiro et al., 2020). We use these subtask and paired comparisons to identify where aggregate gains are concentrated and where errors persist.

5.1 GC: Reference-Dependent Geometric Gains

GC contains three types of geometric judgments: Camera-Relative Distance (51 questions), Physical Elevation (64), and Inter-Object Distance (100). Separating them allows us to ask whether improvements in geometric reasoning transfer across different reference relations, rather than treating geometry as a single capability (Johnson et al., 2017; Hudson & Manning, 2019).

Figure 5 shows that the answer depends strongly on the relation being judged. Qwen3-VL-8B scores 54.9%, 46.9%, and 49.0% on the three subsets, compared with 60.8%, 57.8%, and 54.0% for the 235B checkpoint. The larger model improves on all three, although sizeable errors remain. SenseNova-SI/Qwen reaches 52.9%, 42.2%, and 50.0%, despite its higher Overall score.

The GPT comparison is more uneven. Thinking raises Inter-Object Distance from 58.0% to 73.0% and Physical Elevation from 56.2% to 59.4%, while Camera-Relative Distance changes by only three answers, from 28/51 to 25/51 correct (54.9% to 49.0%).

5.2 TN: Evidence Assessment and Topological Decisions

TN separates View Selection (52 questions), Spatial Consistency (82), Direct Connectivity (41), and Route Planning (42). These tasks differ in how much they require the model to move beyond selecting or comparing visible evidence toward reasoning about connectivity and paths. Related route-oriented benchmarks and visually grounded navigation tasks similarly require models to connect local observations with larger spatial structure (Yang et al., 2024; Anderson et al., 2018). In elevated outdoor scenes, several structures may be visible at once, but visual co-occurrence alone does not determine whether they are connected.

The subtask results reveal a large gap between these judgments. Qwen3-VL-8B obtains 77.1% on Spatial Consistency, but only 28.6% on Direct Connectivity and 37.2% on Route Planning. For Qwen3-VL-235B, the corresponding scores are 77.1%, 49.0%, and 39.5%. Thus, scaling produces a substantial gain in direct connectivity while leaving route planning nearly unchanged. SenseNova-SI/Qwen shows another pattern: it scores 75.9% on Spatial Consistency and 41.9% on Route Planning, while Direct Connectivity remains at 28.6%, the same as its base model.

Table 3: **Paired GPT answer changes.** Accuracy is no-reasoning → thinking. Corrections are wrong-to-correct transitions; regressions are correct-to-wrong transitions under recorded scoring.

Subset	N	Accuracy (%)	Corrections	Regressions
Spatial Consistency	82	$78.3 \rightarrow 77.1$	3	4
Direct Connectivity	51	$67.3 \rightarrow 61.2$	1	4
Route Planning	42	$37.2 \rightarrow 62.8$	14	3

These results show a large gap between consistency judgments and connectivity or route decisions across several model configurations. Because the subtasks use different questions, the contrast identifies a recurring capability difference across judgments rather than a sequence of errors within a single inference process for any particular model.

5.3 Paired Diagnosis: Thinking Selectively Improves Routes

The small aggregate TN gain under GPT thinking masks very different changes across its subtasks. To separate genuine corrections from newly introduced errors, we compare predictions on the same questions under the two recorded configurations (Ribeiro et al., 2020). Table 3 summarizes the resulting corrections and regressions for each subset.

Route Planning improves from 16/43 to 27/43 correct answers, with 14 corrections and three regressions. Direct Connectivity shows the opposite balance, with one correction and four regressions, while Spatial Consistency changes by only one net answer. The decreases are small. The main result is instead the asymmetry of the gains: the improvement in route planning is not matched by similar improvements in the other topological subtasks.

Paired outcomes expose changes that aggregate scores miss, including corrections that coexist with newly introduced errors. They also show why a gain in route planning should not be read as a general improvement in topological reasoning: direct-connectivity answers can move in the opposite direction. Separating planning from visual grounding would require a further intervention that supplies the model with verified scene structure.

Qualitative failures. Representative errors show that a prediction can preserve part of a spatial relation while reversing another. Models may, for example, retain the correct left–right ordering while confusing front and rear, or select a visually plausible route that violates local connectivity. These examples complement the quantitative results above and are provided in Appendix A.2.

6 Conclusion

SIGHT-Bench evaluates spatial reasoning in elevated outdoor scenes with 1,000 questions across five dimensions and 12 subtasks. Our evaluation shows that current multimodal models still exhibit substantial gaps from human performance across proprietary, general-purpose open, and spatially specialized models, with the remaining challenges varying considerably across spatial dimensions. Across the evaluated configurations, spatial consistency is generally more reliable than direct connectivity, suggesting that models may capture certain local or relational structures without consistently recovering broader topological organization. Additional thinking produces a large improvement in world-frame motion reasoning, while other dimensions benefit much less and continue to exhibit distinct patterns of difficulty.

The finer-grained analyses reinforce this uneven capability profile. Models with similar Overall scores can exhibit different strengths and weaknesses across spatial dimensions, and improvements in one relation type, such as inter-object distance or route planning, do not necessarily transfer to related judgments such as camera-relative distance or direct connectivity. By reporting dimension- and subtask-level results alongside aggregate accuracy, SIGHT-Bench provides a more diagnostic view of spatial intelligence in multimodal models. We hope this benchmark encourages future progress toward models that achieve not only higher overall accuracy, but also more consistent and generalizable spatial reasoning across diverse real-world settings.

AI USE STATEMENT

Generative AI tools assisted with language revision, code development, and manuscript formatting. The authors remain responsible for the resulting claims, analyses, references, and presentation in the submitted manuscript and its accompanying research materials.

ETHICS STATEMENT

This work studies spatial reasoning capabilities of multimodal models through a diagnostic benchmark based on real-world elevated outdoor camera observations. We do not collect new human-subject data or conduct user studies. The benchmark is constructed from existing visual observations and focuses on evaluating model behavior under controlled research settings.

Our benchmark may inherit limitations from the underlying data sources, including visual biases, geographic coverage limitations, annotation errors, and ambiguities caused by real-world environments. Although the benchmark is designed to evaluate spatial reasoning rather than enable deployment, models evaluated on SIGHT-Bench may still exhibit unreliable behavior under distribution shifts or unfamiliar environments. Therefore, the results should be interpreted as research diagnostics rather than guarantees of real-world spatial understanding or safe deployment.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the SIGHT-Bench construction process, including camera selection, question generation, verification procedures, task definitions, evaluation protocols, and baseline configurations.

Upon publication, we will release the benchmark data, evaluation code, question-answer annotations, and supporting resources through a public repository. For any underlying visual sources or third-party resources subject to licensing restrictions, we will provide the corresponding data descriptions, usage instructions, and access procedures where applicable.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025. URL https://arxiv.org/abs/2511.13719.

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. *arXiv preprint arXiv:2401.12168*, 2024. URL https://arxiv.org/abs/2401.12168.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*, 2024. URL https://arxiv.org/abs/2406.01584.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Amit Agarwal, Zhe Chen, Mo Li, Yubo Ma, Hailong Sun, Xiangyu Zhao, Junbo Cui, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024. URL https://arxiv.org/abs/2407.11691.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Kaiyuan Hou, Minghui Zhao, Lilin Xu, Yuang Fan, and Xiaofan Jiang. TDBench: Benchmarking vision-language models in understanding top-down images. *arXiv preprint arXiv:2504.03748*, 2025a. URL https://arxiv.org/abs/2504.03748.

Kaiyuan Hou, Minghui Zhao, Lilin Xu, Yuang Fan, and Xiaofan Jiang. Tdbench: Benchmarking vision-language models in understanding top-down images. *arXiv preprint arXiv:2504.03748*, 2025b.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Roberta L. Klatzky. Allocentric and Egocentric Spatial Representations: Definitions, Distinctions, and Interconnections. In Christian Freksa, Christopher Habel, and Karl F. Wender (eds.), *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, volume 1404 of *Lecture Notes in Computer Science*, pp. 1–17. Springer, 1998. doi: 10.1007/3-540-69342-4_1.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025a. URL https://arxiv.org/abs/2505.21500.

Haotian Li, Yida Wang, Leyuan Wang, Jinshan Lai, Keyang Wang, Zonghao Guo, Qiang Ma, Liuyu Xiang, Jianwei Hu, and Zhaofeng He. Airgroundbench: Probing spatial intelligence in multimodal large models under heterogeneous multi-view embodied collaboration. *arXiv preprint arXiv:2606.28049*, 2026a.

Haotian Li, Yida Wang, Leyuan Wang, Jinshan Lai, Keyang Wang, Zonghao Guo, Qiang Ma, Liuyu Xiang, Jianwei Hu, and Zhaofeng He. AirGroundBench: Probing spatial intelligence in multimodal large models under heterogeneous multi-view embodied collaboration. *arXiv preprint arXiv:2606.28049*, 2026b. URL https://arxiv.org/abs/2606.28049.

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. *arXiv preprint arXiv:2510.08531*, 2025b. URL https://arxiv.org/abs/2510.08531.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025. URL https://arxiv.org/abs/2510.27606.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M. de Melo, Alan Yuille, and Jieneng Chen. 3DSRBench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024. URL https://arxiv.org/abs/2412.07825.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Li Fei-Fei, Jiajun Wu, and Manling Li. Mindcube: Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025. URL https://arxiv.org/abs/2506.21458.

Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin, and Yong Li. Citycube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9190–9215, 2026a.

Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin, and Yong Li. CityCube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9190–9215, 2026b. doi: 10.18653/v1/2026.acl-long.416.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025. URL https://arxiv.org/abs/2511.21471.

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. URL https://arxiv.org/abs/2412.14171.

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. *arXiv preprint arXiv:2505.23764*, 2025. URL https://arxiv.org/abs/2505.23764.

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

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

Haoyu Zhang, Meng Liu, Qianlong Xiang, Kun Wang, Yaowei Wang, and Liqiang Nie. Spatialuav: Benchmarking spatial intelligence for low-altitude uav perception, collaboration, and motion. *arXiv preprint arXiv:2606.27876*, 2026b.

Haoyu Zhang, Meng Liu, Qianlong Xiang, Kun Wang, Yaowei Wang, and Liqiang Nie. SpatialUAV: Benchmarking spatial intelligence for low-altitude uav perception, collaboration, and motion. *arXiv preprint arXiv:2606.27876*, 2026c. URL https://arxiv.org/abs/2606.27876.

Lingfeng Zhang, Yuchen Zhang, Hongsheng Li, Haoxiang Fu, Yingbo Tang, Hangjun Ye, Long Chen, Xiaojun Liang, Xiaoshuai Hao, and Wenbo Ding. Is your vlm sky-ready? a comprehensive spatial intelligence benchmark for uav navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25904–25913, 2026d.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. *arXiv preprint arXiv:2509.25390*, 2025. URL https://arxiv.org/abs/2509.25390.

A APPENDIX

A.1 TASK EXAMPLES

Figure 6 provides representative questions across the five diagnostic dimensions and their constituent subtasks.



Figure 6: **Representative task examples across the five diagnostic dimensions.** Rows correspond to Geometric Comparison (GC), Scene Configuration (SC), Viewpoint Analysis (VA), World-Frame Motion (WF), and Topological Navigation (TN). The examples illustrate the question and option formats used across the constituent subtasks of SIGHT-Bench.

A.2 ADDITIONAL QUALITATIVE FAILURE CASES

Figure 7 presents representative incorrect predictions from Scene Configuration, Viewpoint Analysis, and Topological Navigation. These examples are illustrative rather than a frequency estimate or an exhaustive taxonomy of errors.

In the Scene Configuration example, the error preserves the lateral rightward relation but reverses the front–rear component. The Viewpoint Analysis example shows a related partial error: the horizontal pan direction is preserved while the tilt direction is reversed. The Topological Navigation example instead selects an alternative route that is visually plausible but differs in the required local transition through the scene. Together, these examples show that incorrect predictions can arise even when part of the relevant spatial relation is preserved, and that plausible scene-level interpretations do not guarantee the correct relational or topological decision.

A.3 ADDITIONAL MODEL COMPARISONS

Figures 8–10 complement the main three-panel comparison. Nominal parameter groups do not ensure equal active parameters, training, or inference compute. Comparisons of specialization use Qwen3-VL-8B and InternVL3-8B as the respective SenseNova bases.

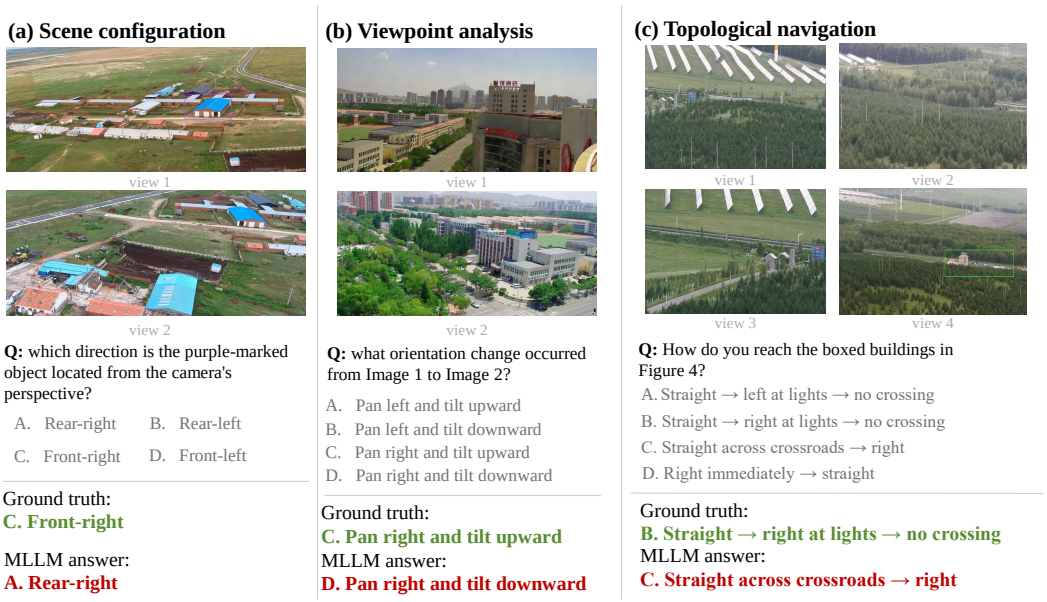


Figure 7: **Illustrative failure cases across spatial dimensions.** (a) In Scene Configuration, the model confuses a front-right relation with rear-right from the camera's perspective. (b) In Viewpoint Analysis, the predicted pan direction is correct, but the vertical orientation change is reversed. (c) In Topological Navigation, the model selects a plausible route that conflicts with the ground-truth connectivity sequence. Green text marks the ground-truth answer and red text marks the illustrative incorrect prediction. These examples describe observable errors rather than internal reasoning mechanisms.

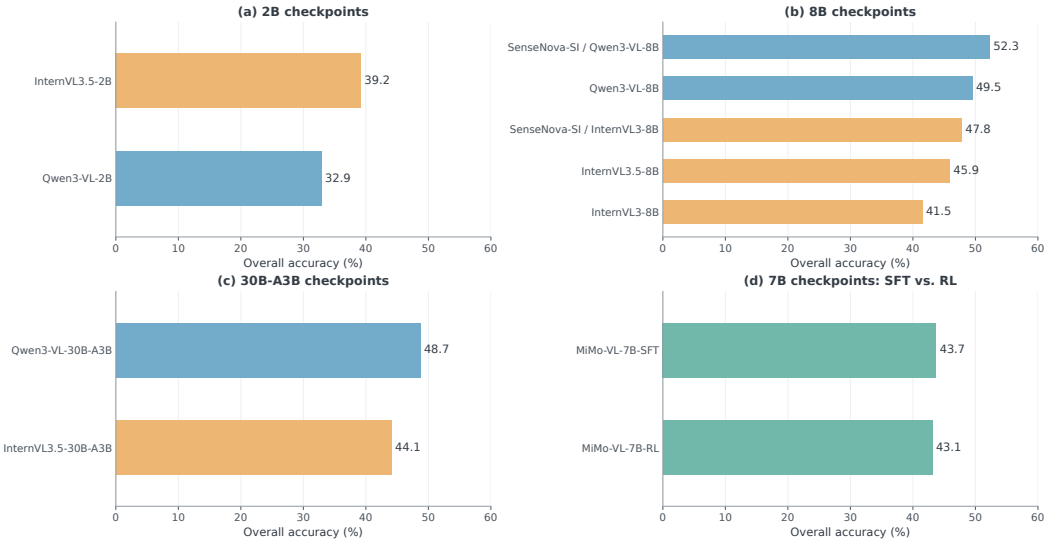


Figure 8: **Comparisons within nominal size groups.** Equal size labels are not a controlled comparison of training or inference compute.

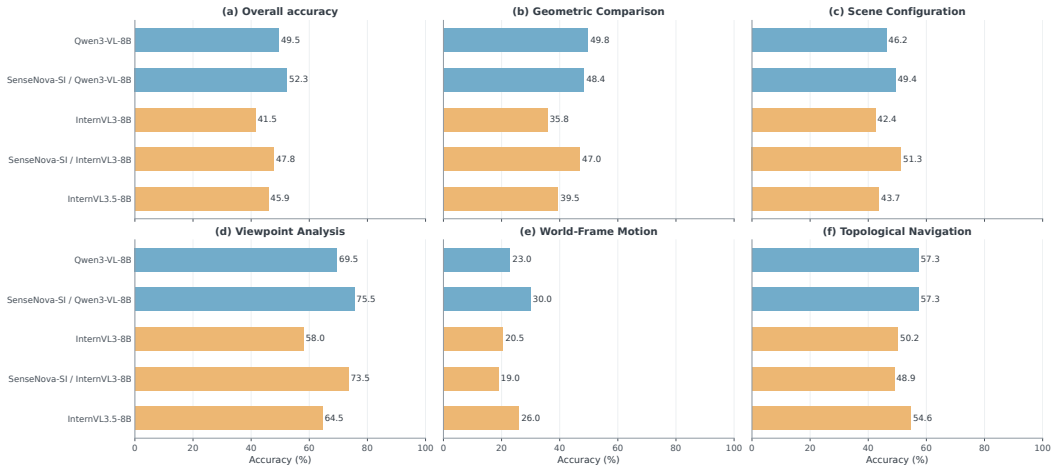


Figure 9: **8B capability profiles.** Overall improvements can conceal selective benefits and regressions across dimensions.

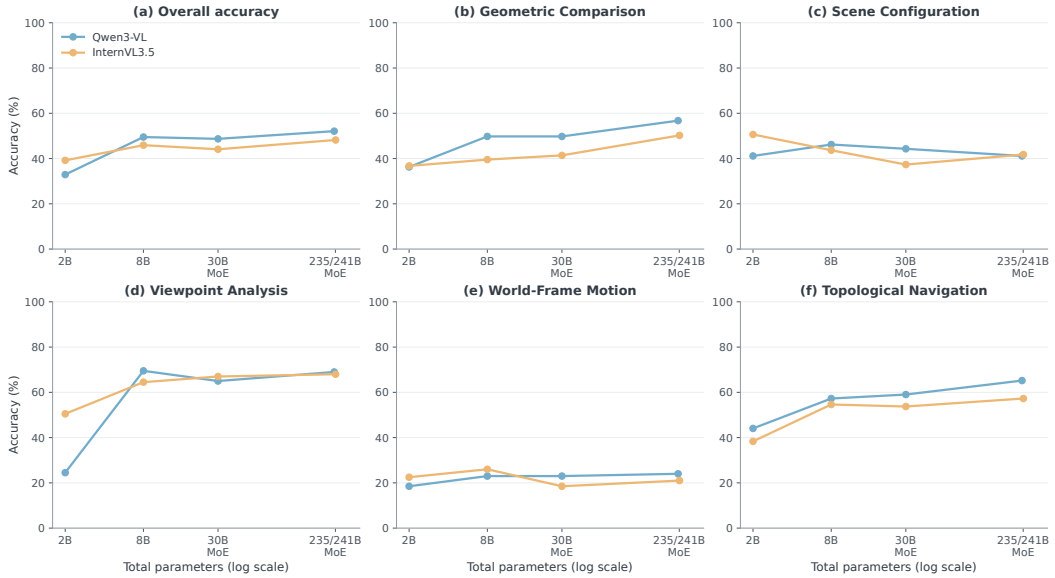


Figure 10: **Size comparisons by dimension.** Dense and MoE checkpoints are shown together; these curves are not fitted scaling laws.

A.4 Expanded Task Examples

Figures 11–34 show two selected questions for each of the twelve subtasks. Each figure includes the answer options and the ground-truth answer.

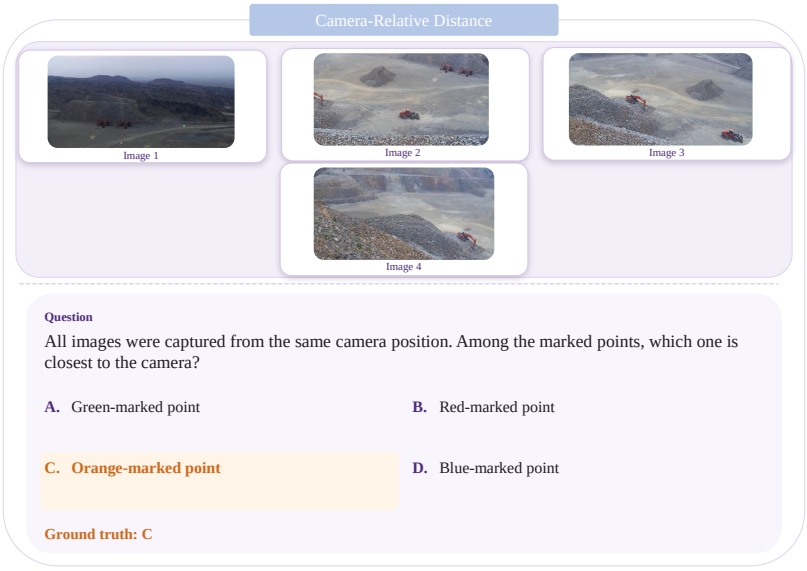


Figure 11: **GC: Camera-relative distance.** Selected example QA 01.

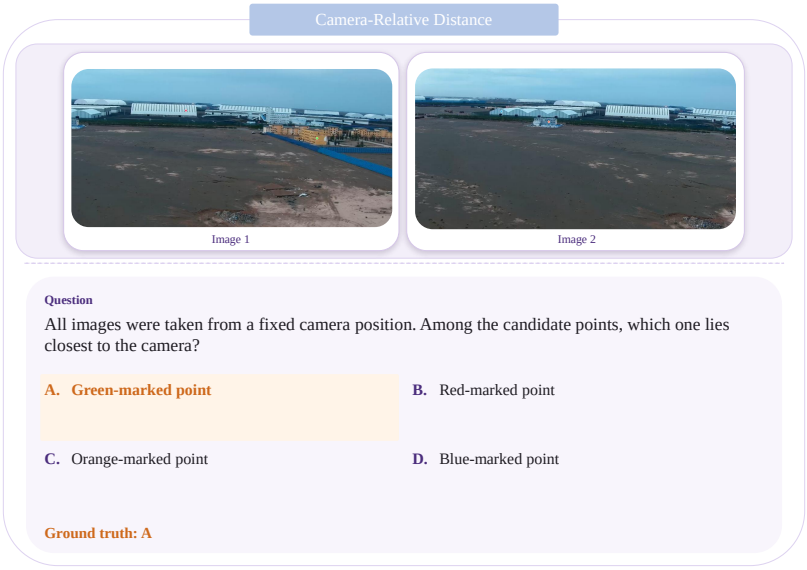


Figure 12: **GC: Camera-relative distance.** Selected example QA 02.

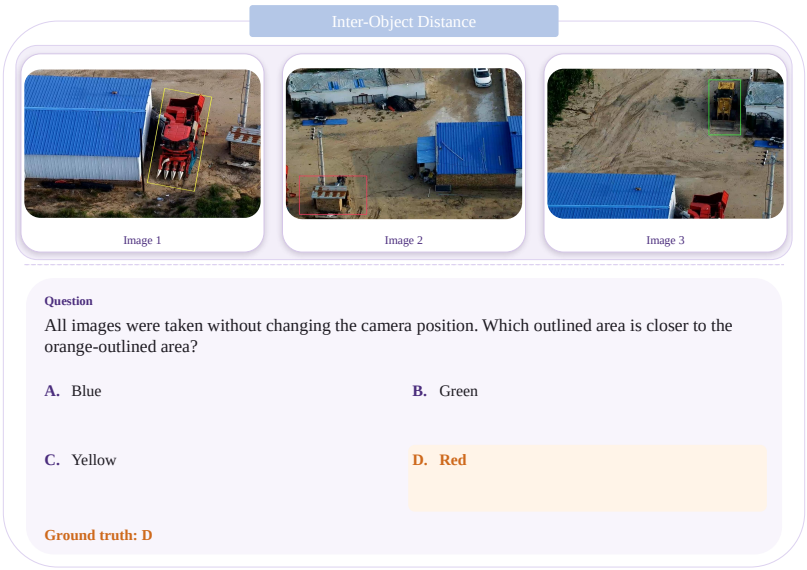


Figure 13: **GC: Inter-object distance.** Selected example QA 03.

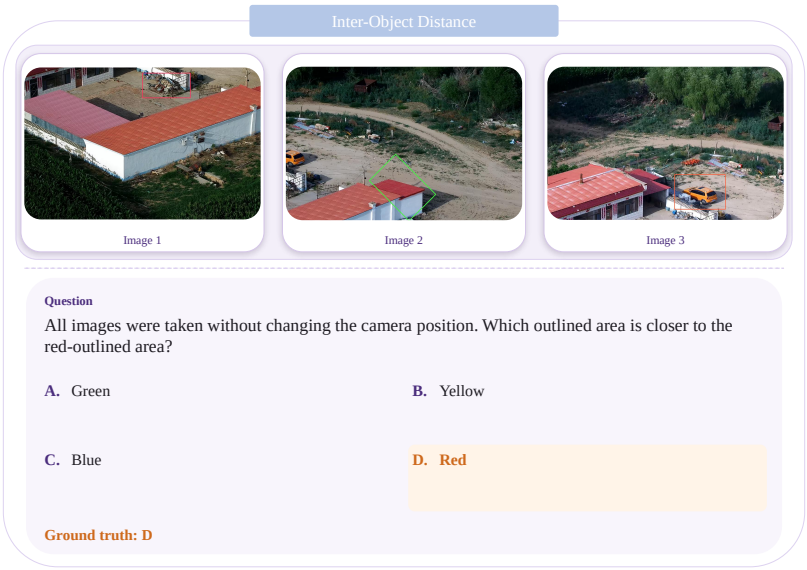


Figure 14: **GC: Inter-object distance.** Selected example QA 04.

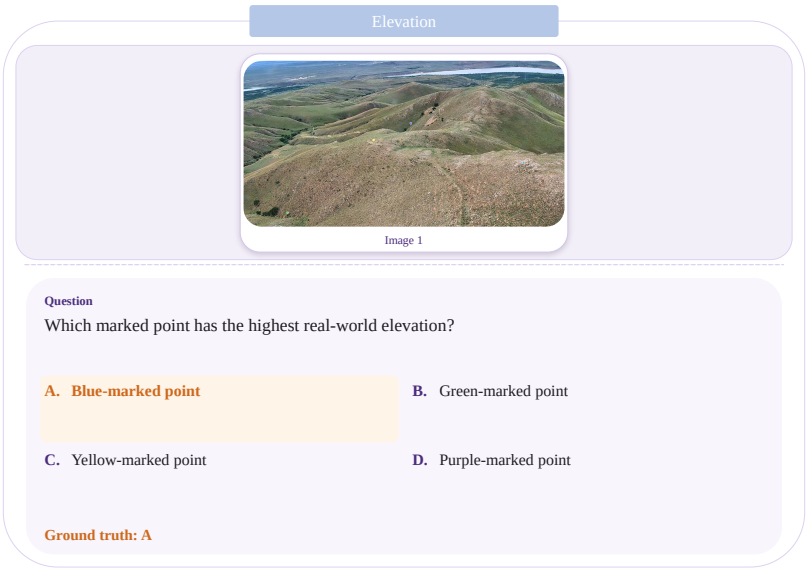


Figure 15: **GC: Elevation.** Selected example QA 05.

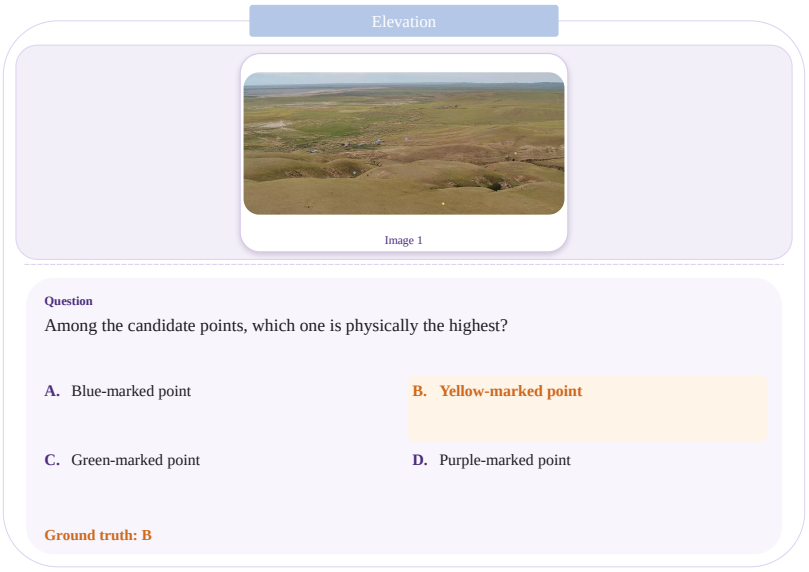


Figure 16: **GC: Elevation.** Selected example QA 06.

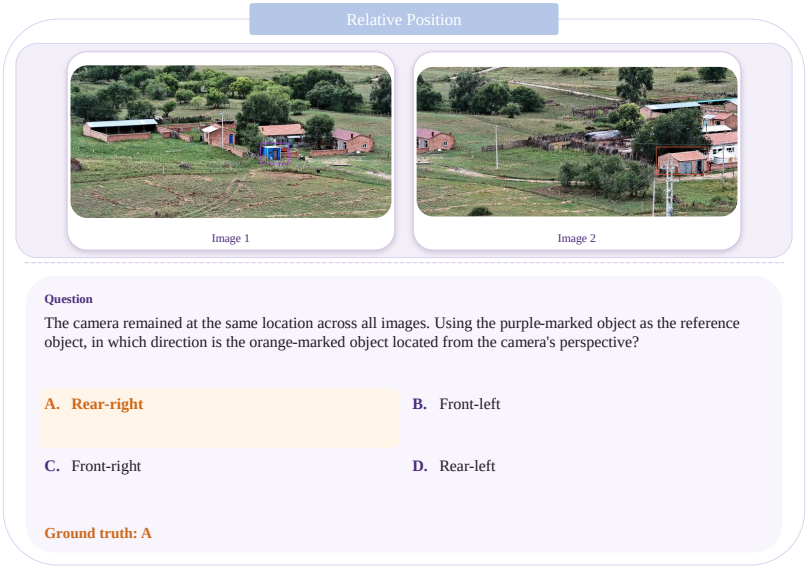


Figure 17: **SC: Relative position.** Selected example QA 07.

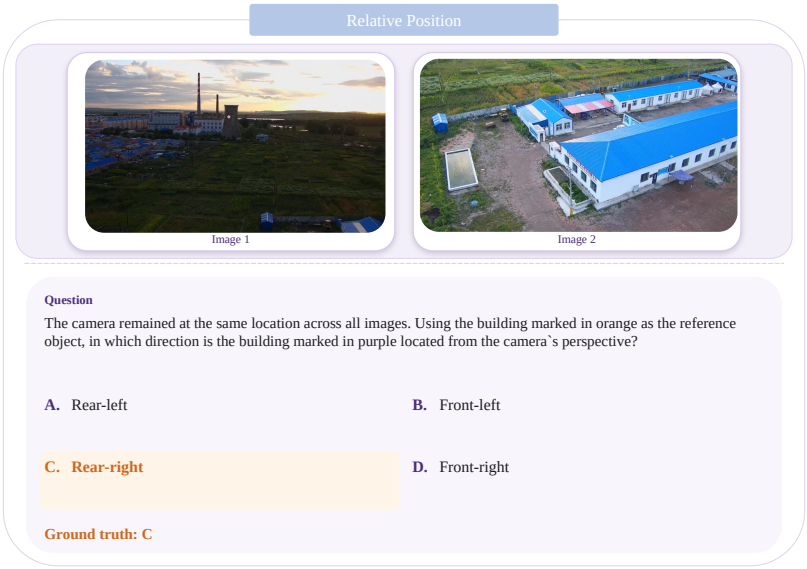


Figure 18: **SC: Relative position.** Selected example QA 08.

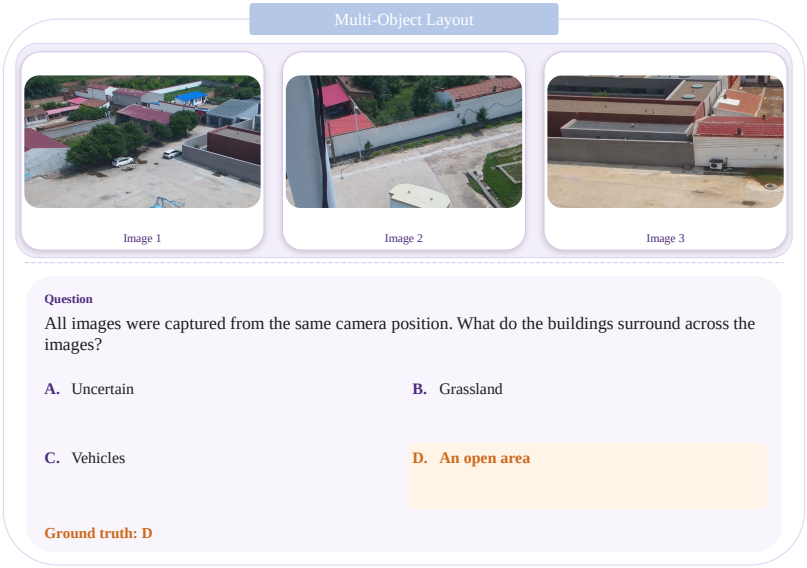


Figure 19: **SC: Multi-object layout.** Selected example QA 09.

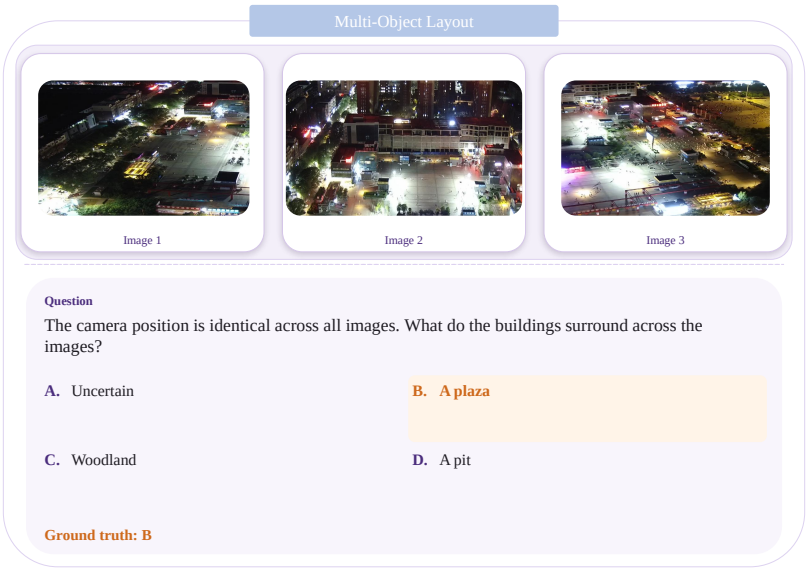


Figure 20: **SC: Multi-object layout.** Selected example QA 10.

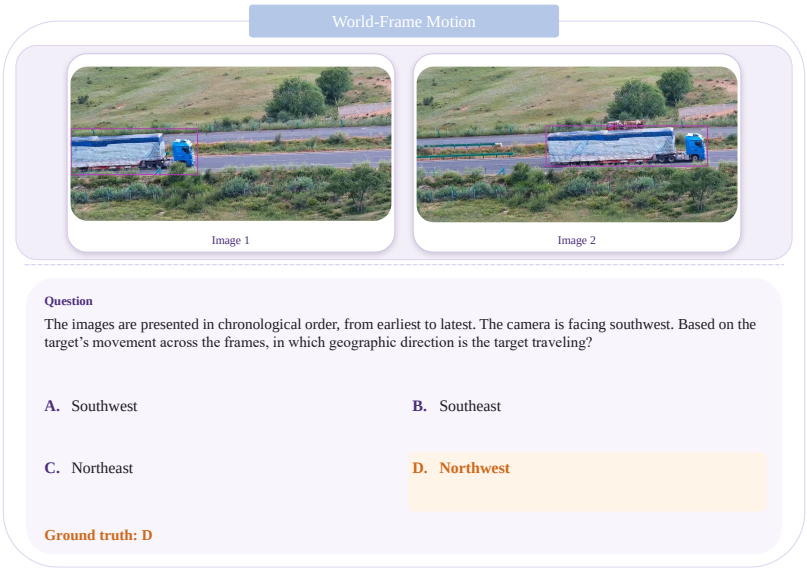


Figure 21: **WF: World-frame motion.** Selected example QA 11.

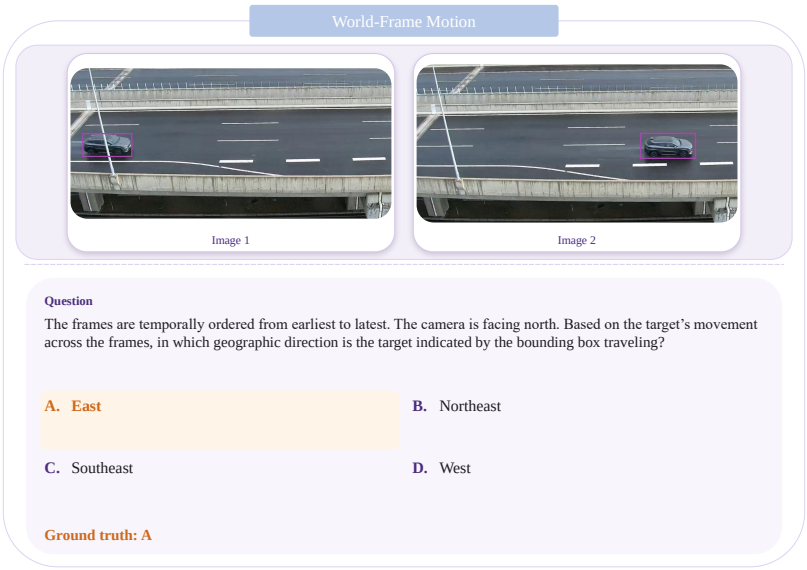


Figure 22: **WF: World-frame motion.** Selected example QA 12.

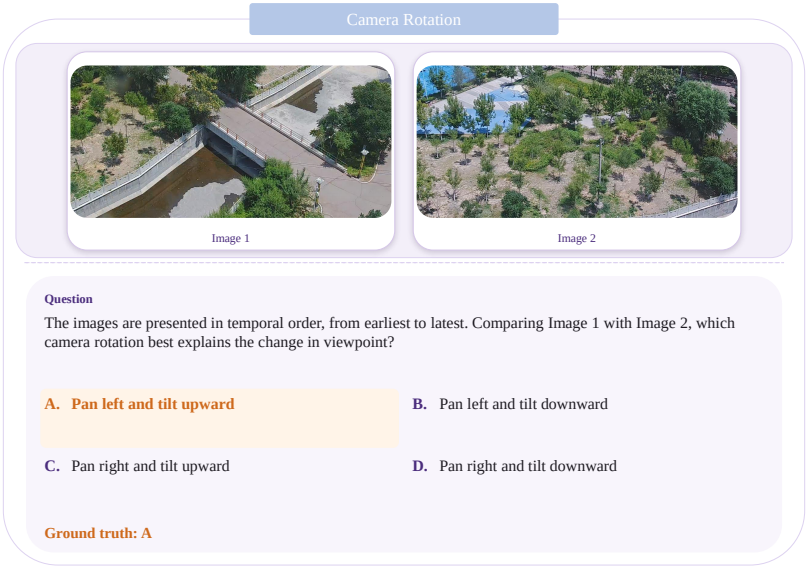


Figure 23: **VA: Camera rotation.** Selected example QA 13.

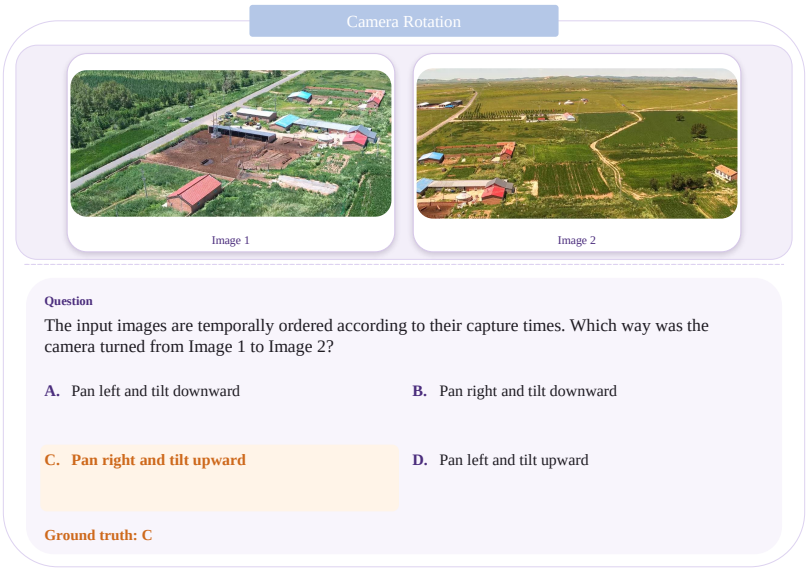


Figure 24: **VA: Camera rotation.** Selected example QA 14.

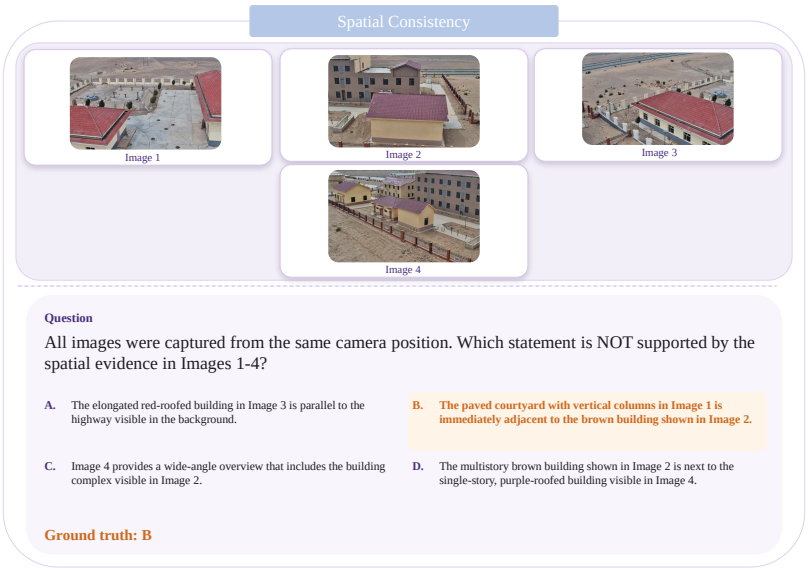


Figure 25: **TN: Spatial consistency.** Selected example QA 15.

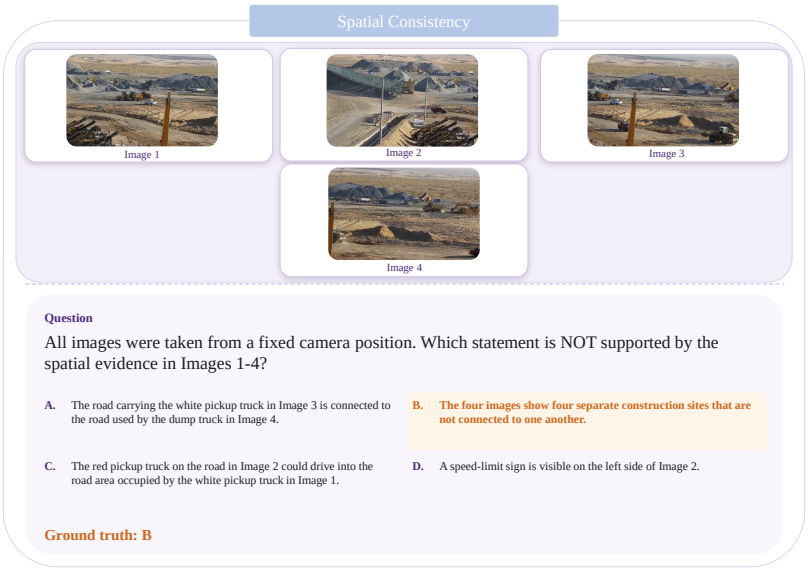


Figure 26: **TN: Spatial consistency.** Selected example QA 16.

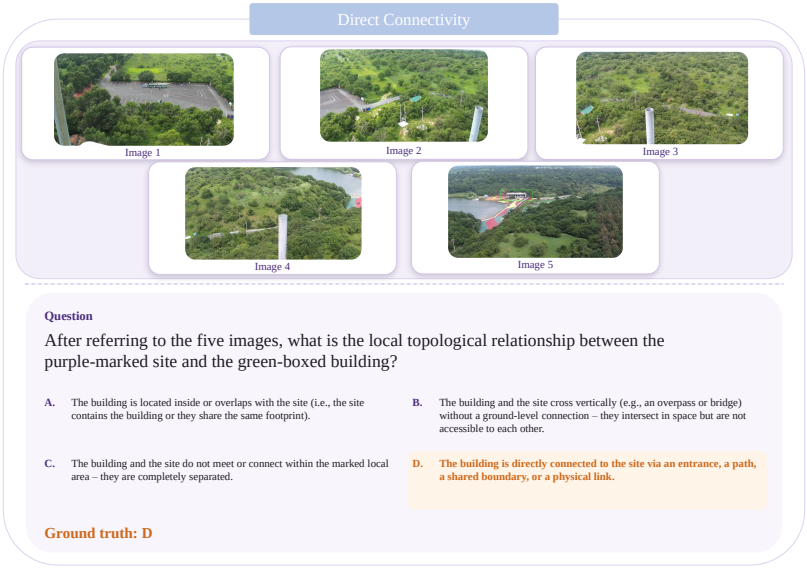


Figure 27: **TN: Direct connectivity.** Selected example QA 17.

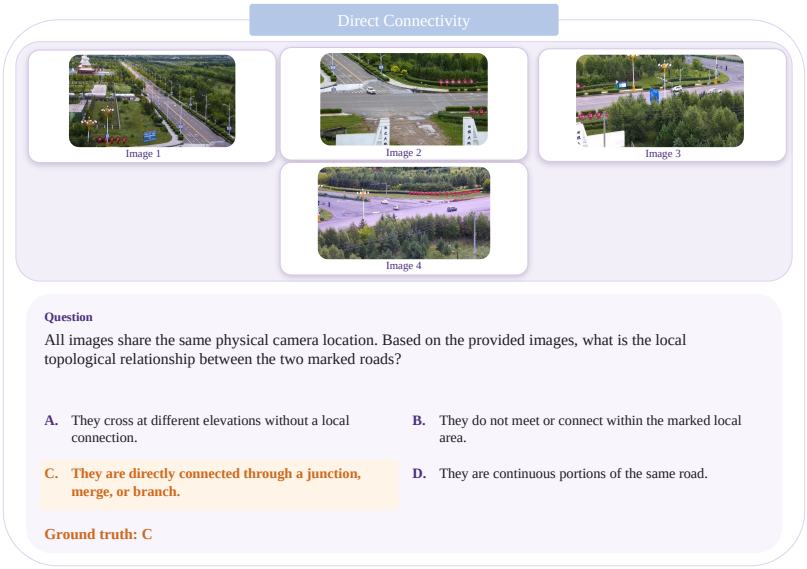


Figure 28: **TN: Direct connectivity.** Selected example QA 18.

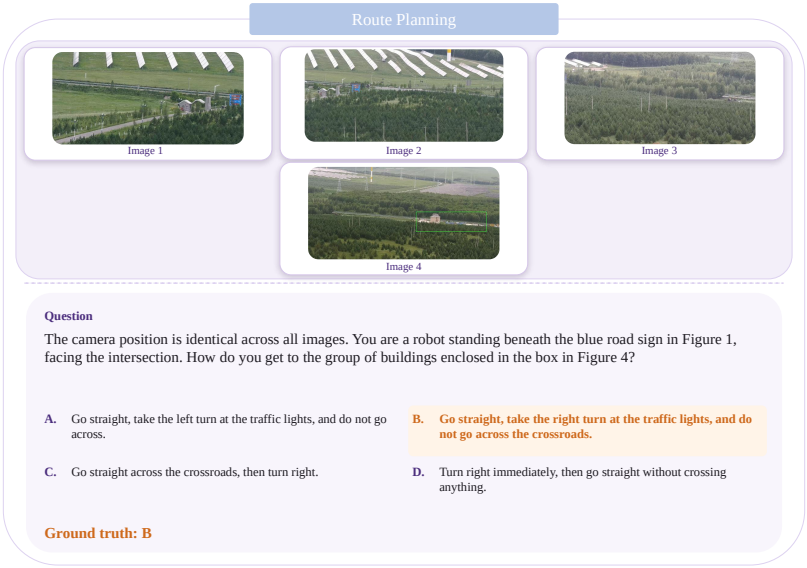


Figure 29: **TN: Route planning.** Selected example QA 19.

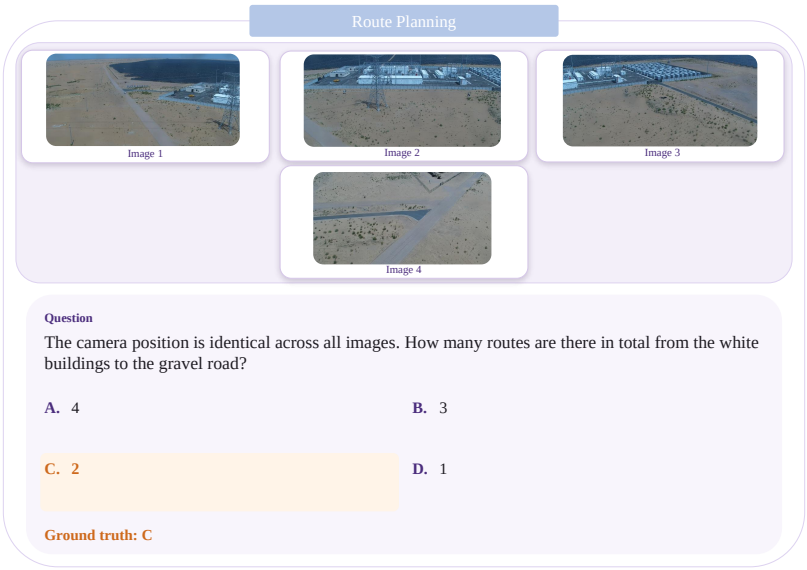


Figure 30: **TN: Route planning.** Selected example QA 20.

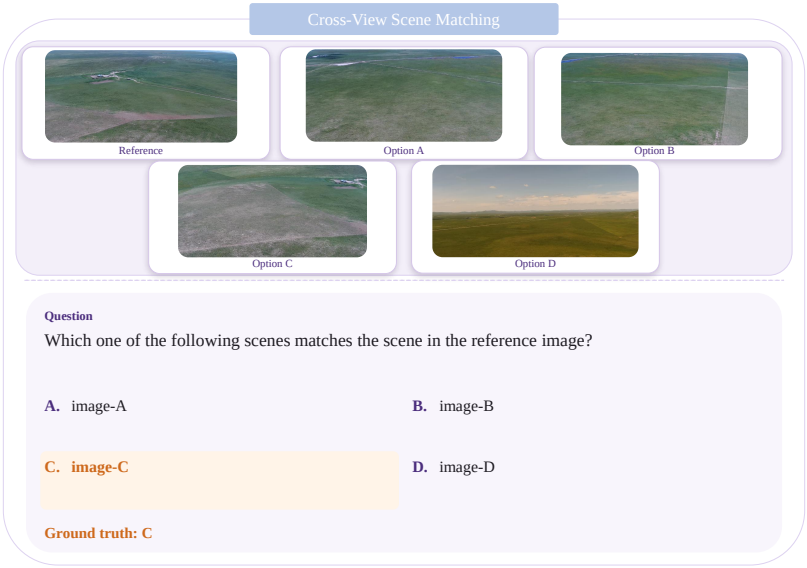


Figure 31: **VA: Cross-view scene matching.** Selected example QA 21.

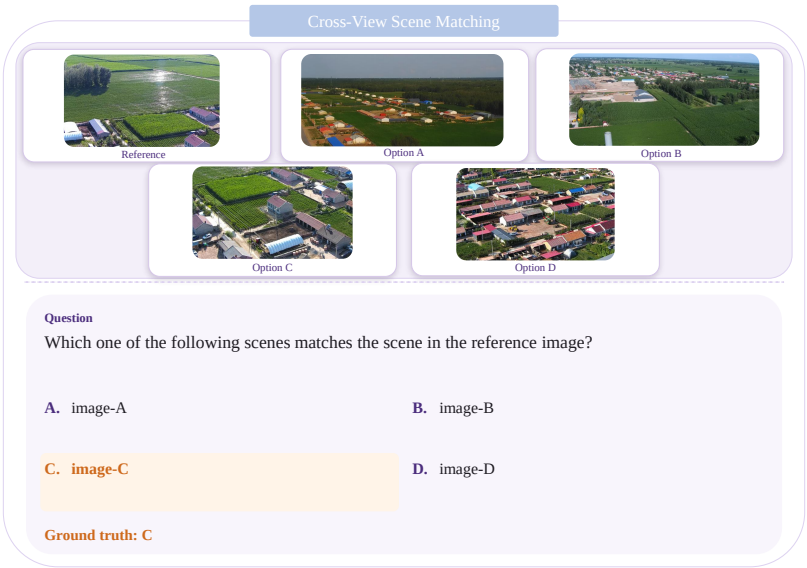


Figure 32: **VA: Cross-view scene matching.** Selected example QA 22.

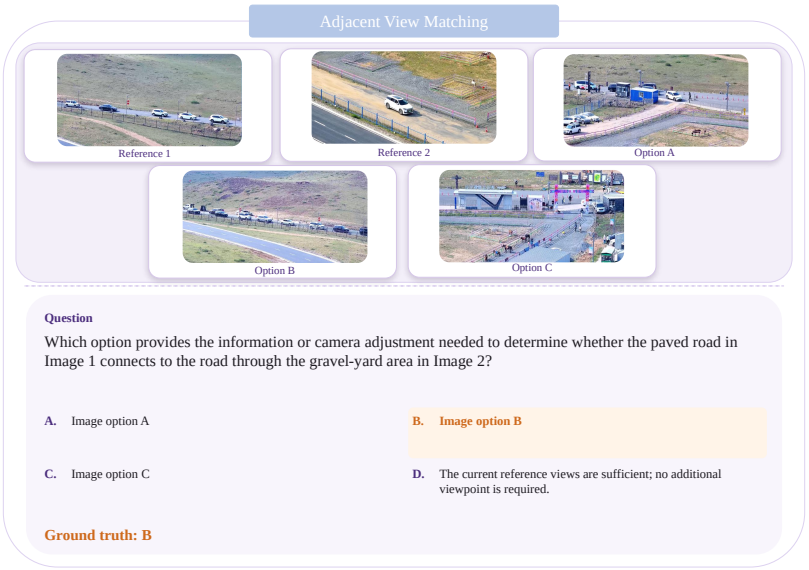


Figure 33: **TN: Adjacent view matching.** Selected example QA 23.

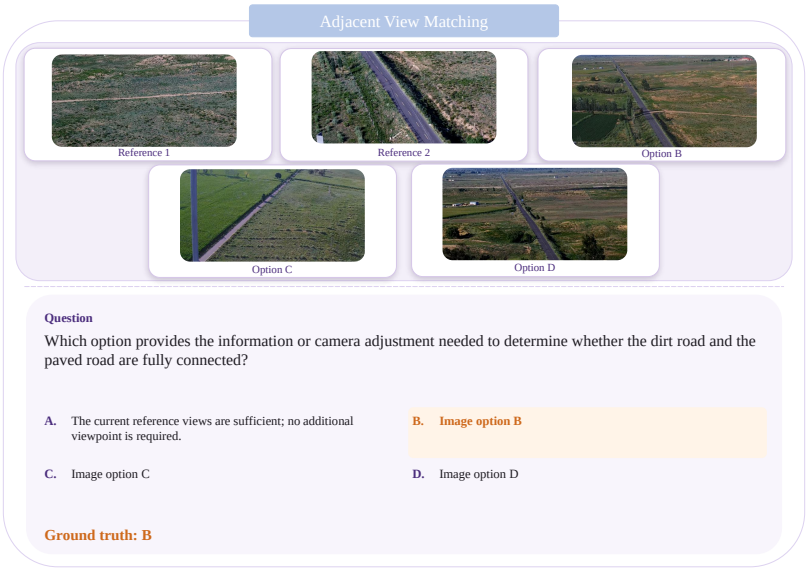


Figure 34: **TN: Adjacent view matching.** Selected example QA 24.