SPATIALROUTE-BENCH: BENCHMARKING LONG-HORIZON SPATIAL INTELLIGENCE OF VISION-LANGUAGE MODELS IN OUTDOOR VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have made rapid progress in spatial understanding, yet existing benchmarks largely focus on local or scene-level spatial relations, leaving long-horizon spatial understanding in continuous outdoor videos underexplored. Such videos require models to integrate spatial evidence scattered across long time spans and large physical distances, as the observer's viewpoint and direction of travel change continuously along the route. To bridge this gap, we introduce **SpatialRoute-Bench**, a geometry-grounded benchmark for evaluating how VLMs perceive, reason over, and use spatial information along extended outdoor trajectories. SpatialRoute-Bench comprises 7,000 multiple-choice QA pairs across seven tasks, organized into a three-level hierarchy in which each level demands more extensive spatial integration than the one below: *perception*, *reasoning*, and *decision*. To enable scalable benchmark construction, we develop a fully automated pipeline that couples VLM-based semantic grounding with 3D reconstruction: semantic annotations determine which landmarks can be referred to, whereas reconstructed camera poses, trajectories, and scene geometry provide the spatial evidence from which answers are derived. Task-specific generators then convert this evidence into questions, retaining only those that pass geometric and semantic validity checks. Evaluations of representative VLMs reveal three critical limitations. First, there remains substantial room for improvement: even the strongest model reaches only 28.25% overall accuracy against a 25% random baseline. Second, camera-motion perception remains unreliable: even for a single motion event, average accuracy is only 26.2%. Third, spatial relations degrade over longer horizons: when the same relation must be recovered between landmarks observed minutes apart, average accuracy drops from 29.2% to 22.1%, below the random baseline. SpatialRoute-Bench exposes a clear gap between general-purpose video understanding and grounded spatial understanding of long outdoor traversals, and provides a diagnostic testbed for closing it.

1 INTRODUCTION

Recent vision-language models (VLMs) have shown increasingly strong spatial reasoning, from relative position and depth to distance, orientation, and other geometry-aware relations (Fu et al., 2024; Chen et al., 2024; Yang et al., 2025b). Long outdoor videos, however, pose a distinct challenge: the relevant spatial evidence is scattered across long time spans and large physical distances, and must be integrated as the observer's viewpoint and direction of travel change continuously. Existing spatial benchmarks capture this setting only partially. Many are built on indoor or spatially confined environments, or evaluate relatively short observation sequences (Yang et al., 2025b; Szymańska et al., 2024; Lang et al., 2026), while recent long-horizon and streaming benchmarks typically target a specific capability, such as spatial memory, continual state tracking, global mapping, or world-state tracking (Wang et al., 2026; Li et al., 2026; Zhang et al., 2026; Gu et al., 2026). As a result, it remains unclear how spatial evidence is perceived, integrated over time, and ultimately used for decisions along a continuous traversal.

Moreover, geometry-grounded spatial QA typically relies on pre-existing RGB-D, LiDAR, simulator states, annotated 3D scenes, or curated reconstructions, often together with substantial human

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.

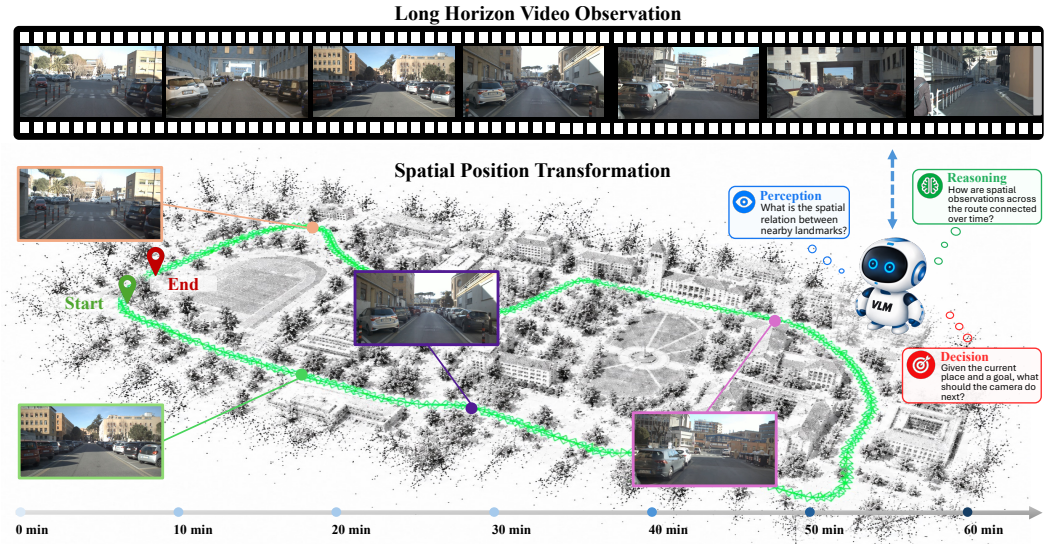


Figure 1: Motivation of SpatialRoute-Bench. Long-horizon outdoor videos scatter spatial evidence across extended trajectories and changing viewpoints. SpatialRoute-Bench evaluates whether VLMs can perceive local spatial states, integrate spatial evidence over time, and use it for destination-directed decisions.

annotation or verification (Yang et al., 2025b; Cheng et al., 2024; Szymańska et al., 2024; Tian et al., 2025; Yang et al., 2025a). These dependencies hinder scaling geometry-grounded evaluation to ordinary long outdoor videos, which come with neither 3D supervision nor manual geometric annotations.

To bridge these gaps, we introduce **SpatialRoute-Bench** (Figure 1), a geometry-grounded benchmark for evaluating how VLMs perceive, reason over, and use spatial information along extended outdoor trajectories. Starting from an ordinary long video, SpatialRoute-Bench recovers camera poses, trajectories, and scene geometry with the streaming 3D reconstruction model LingBot-Map (Chen et al., 2026), and extracts recognizable landmarks with a Qwen-based VLM (Alibaba Cloud, 2026). The two sources play complementary roles: semantic annotations determine what entities can be referred to in natural-language questions, whereas reconstructed geometry determines where they are and supplies the spatial evidence from which answers are derived. Task-specific generators then convert this evidence into multiple-choice questions, retaining only those that pass geometric and semantic validity checks.

SpatialRoute-Bench comprises 7,000 multiple-choice QA pairs across seven tasks. Inspired by prior work on spatial intelligence and visual navigation (Li et al., 2026; Zhang et al., 2026; Gupta et al., 2017; Zhang et al., 2018), we organize these tasks into a three-level hierarchy in which each level demands more extensive spatial integration than the one below. *Perception* reads out local spatial states: the relative direction between landmarks observed within a short interval and the camera's heading change when passing a landmark. *Reasoning* relates and composes these states across longer route segments through long-range relative direction, route-distance comparison, and sequential turns. *Decision* uses the accumulated evidence to choose the turn or waypoint that leads toward a specified target landmark, which we refer to as the destination.

Using this hierarchy as a diagnostic lens, we evaluate ten representative VLMs on SpatialRoute-Bench and identify three critical limitations. First, there remains substantial room for improvement: even the strongest model reaches only 28.25% overall accuracy against a 25% random baseline, and no model exceeds 41% on any task. Second, camera-motion understanding remains weak: average accuracy is only 26.2% for recognizing a single camera-motion event and 25.4% for composing two consecutive turns. Third, spatial relations degrade over longer horizons: when the same four-way relation must be recovered between landmarks observed at least three minutes apart, average accuracy drops from 29.2% to 22.1%, below chance, and nine of the ten models degrade. Moreover, frontier proprietary models show no clear advantage over open-weight ones: both groups fall within the same narrow band of 23.82–28.25%. Models with similar overall accuracy also exhibit markedly

different task-wise profiles, so aggregate scores alone obscure important differences in spatial capability. Together, these findings expose a clear gap between general-purpose video understanding and grounded spatial understanding of long outdoor traversals.

Our contributions are as follows:

- We introduce SpatialRoute-Bench, a geometry-grounded benchmark for long-horizon spatial understanding in outdoor videos, comprising 7,000 multiple-choice QA pairs across seven tasks organized into a three-level hierarchy of perception, reasoning, and decision.
- We develop a fully automated pipeline that constructs geometry-grounded QA directly from long outdoor videos: VLM annotations determine which landmarks can be referred to, while reconstructed 3D geometry supplies the spatial evidence from which answers are derived, requiring neither pre-existing 3D annotations nor manual spatial labels.
- We benchmark representative VLMs and reveal three critical limitations: substantial room for improvement, near-chance perception of camera motion, and a sharp degradation of spatial relations over longer temporal horizons. These findings expose a clear gap between general-purpose video understanding and grounded spatial understanding, and establish SpatialRoute-Bench as a diagnostic testbed for closing it.

2 Related Work

2.1 Spatial Reasoning in Vision-Language Models

Spatial reasoning in VLMs ranges from qualitative relations to metric depth and distance estimates (Chen et al., 2024; Stogiannidis et al., 2025). Evaluation has progressed from single images to multi-view and video settings. BLINK recasts classical vision tasks into multimodal questions (Fu et al., 2024), while SpatialVLM targets qualitative and quantitative 3D reasoning (Chen et al., 2024). MMSI-Bench evaluates reasoning across multiple images (Yang et al.), and VSI-Bench focuses on egocentric indoor videos (Yang et al., 2025b). Despite this, models lag behind humans in precise geometric and compositional reasoning (Chen et al., 2024; Yang et al.; Stogiannidis et al., 2025). Moreover, many benchmarks entangle spatial reasoning with object recognition, obscuring where models fail (Stogiannidis et al., 2025).

2.2 Long-Horizon Spatial Understanding in Video

Long-horizon spatial understanding requires reasoning beyond current views, as relevant landmarks may be observed minutes earlier from different viewpoints (Gupta et al., 2017; Parisotto & Salakhutdinov, 2017; Zhang et al., 2018; Cartillier et al., 2021; Mirowski et al., 2018; Chaplot et al., 2020). Among recent benchmarks, VSI-SUPER and LongSpace-Bench test spatial recall over long videos (Yang et al., 2026; Lang et al., 2026), while S3-Bench, UCS-Bench, and OVO-S-Bench evaluate spatial reasoning on streaming inputs (Wei et al., 2026; Wang et al., 2026; Li et al., 2026). GST-Bench examines global spatial consistency (Zhang et al., 2026), and HourVideo includes navigation questions over hour-long videos (Chandrasegaran et al., 2024). Concurrently, VSI-Super-Wild evaluates world-state tracking in diverse indoor and outdoor videos (Gu et al., 2026), and KilometerVision builds questions from hour-long walking tours, although each question covers at most ten minutes (Mahendran et al., 2026). These benchmarks typically focus on one capability or short video segments. SpatialRoute-Bench instead evaluates a perception–reasoning–decision hierarchy over outdoor traversals, deriving answers from geometry recovered from RGB video.

2.3 3D-Grounded Spatial Evaluation

Image appearance alone is an unreliable basis for spatial labels, as perspective, occlusion, and viewpoint changes make depth and relative position ambiguous (Eigen et al., 2014; Zhou et al., 2017). Many benchmarks ground their questions in explicit 3D information. Some derive 3D cues from images: SpatialVLM extracts spatial supervision from geometric cues in real photos (Chen et al., 2024), and SpatialRGPT combines depth with 3D scene graphs (Cheng et al., 2024). Others build questions from 3D data. Space3D-Bench and VSI-Bench use reconstructed or scanned indoor scenes (Szymańska et al., 2024; Yang et al., 2025b), NuScenes-SpatialQA uses 3D annotations

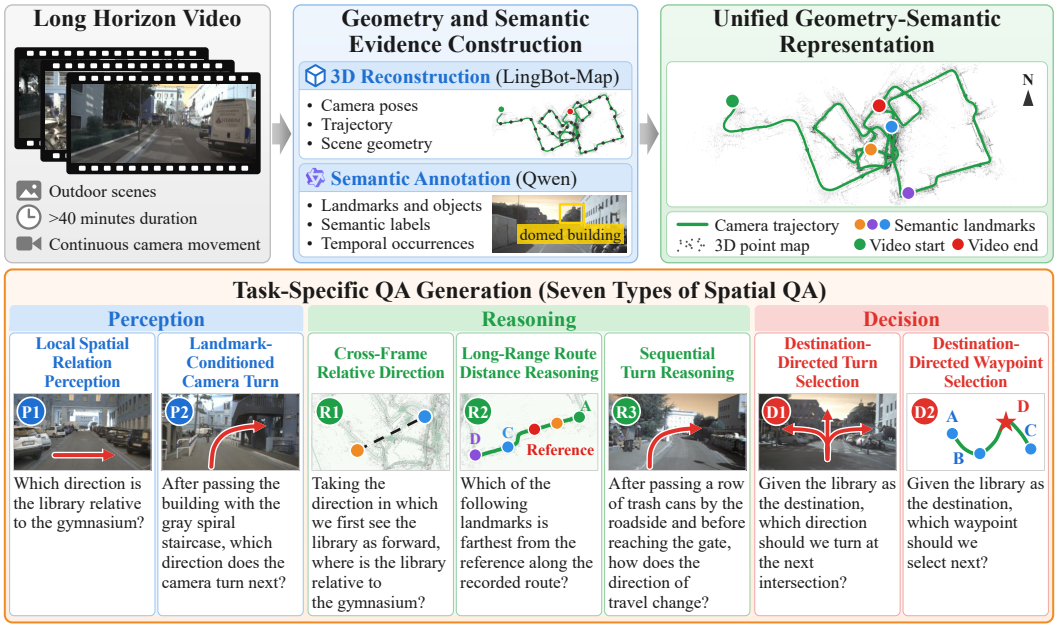


Figure 2: Overview of the SpatialRoute-Bench pipeline. Given a long outdoor video, we reconstruct camera motion and scene geometry, ground semantic landmarks with a VLM, and automatically generate seven geometry-grounded QA tasks spanning perception, reasoning, and decision.

from driving data (Tian et al., 2025), and 3D-GRAND generates spatial language from structured 3D scenes (Yang et al., 2025a). These works show that 3D geometry makes spatial labels objective and scalable. However, they either operate on single images or rely on 3D data from scans, RGB-D sensors, or LiDAR. Building cross-time spatial questions directly from long outdoor videos, with geometry recovered from video, remains less explored.

3 Method

As shown in Figure 2, SpatialRoute-Bench builds spatial questions from long outdoor videos by combining two evidence streams. A geometric stream recovers camera motion and scene structure, and a semantic stream identifies static landmarks for reference. The two streams are aligned on the video timeline so semantic observations can be grounded in the reconstructed scene. Throughout the pipeline, semantics determines which landmarks can be referred to, whereas all spatial labels and answers are derived exclusively from reconstructed geometry and recorded camera motion.

3.1 Geometric and Semantic Evidence Construction

Given a video $\mathcal{V}$, we first build a shared evidence representation

$$\mathcal{E}(\mathcal{V}) = (\mathcal{E}_{\text{geo}}, \mathcal{E}_{\text{sem}}, \mathcal{M}), \tag{1}$$

where $\mathcal{E}_{\text{geo}}$ holds the reconstructed geometry and camera motion, $\mathcal{E}_{\text{sem}}$ holds frame-level landmark observations, and $\mathcal{M}$ links observations of the same persistent landmark. All task generators operate on this shared representation, whose components we describe below.

Geometric evidence. We apply LingBot-Map (Chen et al., 2026) to a temporally regular frame stream to recover camera poses, depth maps, intrinsics, and confidence metadata, while keeping original video timestamps. Valid camera centers form the recorded trajectory. Because the reconstruction is defined only up to a global scale, all geometric quantities are expressed in relative scene units rather than metric distances. We split the trajectory into continuous segments at invalid poses, large temporal gaps, abrupt spatial jumps. A validated world-up direction then defines the horizontal plane on which observer-relative relations, route distances, and turn directions are computed.

Table 1: Comparison with representative video spatial understanding benchmarks. Duration is the average video length in minutes. *Geometry-Grounded*: answers are derived from 3D geometry. *Motion Composition*: questions compose multiple camera-motion events, such as consecutive turns. *Destination Decision*: questions ask which turn or waypoint leads toward a specified destination. ✓: supported; △: partially supported; ✗: not supported; N/R: not reported.

Dataset	Video Setting	Duration	#QA	Geometry-Grounded	Motion Composition	Destination Decision
VSI-Bench	Real / Indoor	1.2	5,130	✓	✗	✓
VSI-SUPER	Constructed / Indoor	N/R	700	✗	✗	✗
LongSpace-Bench	Real / Indoor	21.4	4,073	✗	✗	✗
OVO-S-Bench	Real / Indoor & Outdoor	N/R	1,680	△	△	✗
GST-Bench	Synthetic / Indoor	N/R	2,762	✓	△	✗
VSI-Super-Wild	Real / Indoor & Outdoor	38.62	6,980	△	✗	✗
SpatialRoute-Bench (Ours)	**Real / Outdoor**	**54.68**	**7,000**	✓	✓	✓

Semantic grounding. On a sparser frame stream, a Qwen-based VLM (Alibaba Cloud, 2026) identifies visible static landmarks and outputs referring expressions, bounding boxes, and semantic quality attributes. The VLM is never asked to predict spatial relations, camera motion, route structure, or answers. Each observation is matched to the nearest valid reconstruction frame by its original-video timestamp and lifted to 3D using that frame's depth, intrinsics, and pose. Observations with unreliable temporal alignment, geometry, or depth are discarded.

Landmark association. For tasks that require persistent landmark identity, we conservatively associate observations that are both semantically compatible and geometrically consistent into landmark tracks stored in $\mathcal{M}$. These tracks only indicate which observations likely show the same physical landmark and never determine benchmark answers.

3.2 Task Taxonomy and Formulation

Building on $\mathcal{E}(\mathcal{V})$, we define seven tasks, each with its own geometric and semantic validity checks, grouped by how spatial evidence is used. Perception tasks read out a local spatial state, reasoning tasks relate or compose evidence along an extended route, and decision tasks select a maneuver or waypoint toward a destination. Two perception tasks have matched reasoning counterparts: P1 and R1 share the same relation space but differ in temporal horizon, while P2 and R3 contrast a single motion event with an ordered pair of turns.

3.2.1 Perception

Local Spatial Relation Perception (P1). P1 asks for the relative position of two landmarks observed within a short time window. Given grounded landmark positions $\mathbf{x}_A$ and $\mathbf{x}_B$, we project their displacement onto the horizontal right and forward axes of the camera at the moment A is observed:

$$r_{AB} = \Delta\mathbf{x}_{AB}^{\top}\mathbf{e}_A^{\text{right}}, \qquad f_{AB} = \Delta\mathbf{x}_{AB}^{\top}\mathbf{e}_A^{\text{forward}}, \tag{2}$$

where $\Delta\mathbf{x}_{AB} = \mathbf{x}_B - \mathbf{x}_A$. Following VSI-Bench (Yang et al., 2025b), the signs of (r_{AB}, f_{AB}) define one of four observer-relative quadrants: front-left, front-right, behind-left, or behind-right. Ambiguous or unreliable landmark pairs are discarded.

Landmark-Conditioned Camera Turn (P2). P2 asks how the camera moves near a recognizable landmark. We detect canonical left and right turns from changes in the horizontal trajectory tangent, take low-curvature intervals as straight events, and then associate a nearby landmark with each event. The answer is one of $\{\text{left}, \text{right}, \text{straight}\}$, with reverse used only as a distractor.

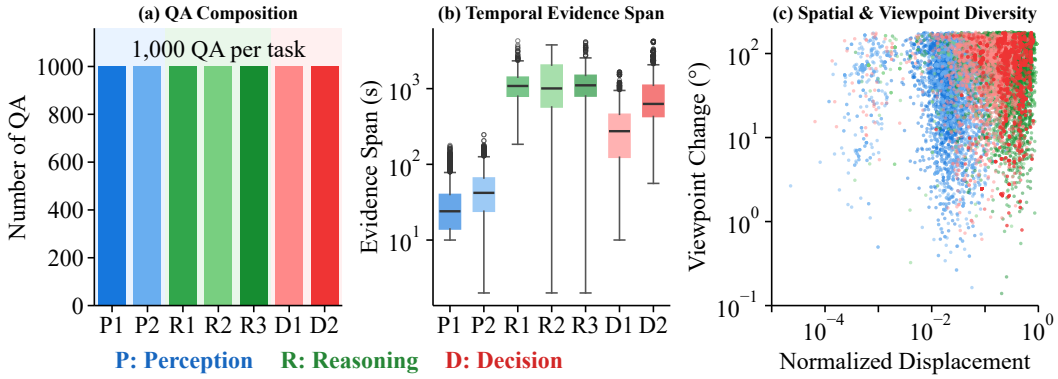


Figure 3: Statistics of SpatialRoute-Bench. (a) Number of QA pairs across the seven tasks. (b) Distribution of temporal evidence spans. R2 exhibits a broad distribution without extreme outliers. (c) Distribution of scene-normalized spatial displacement and relative viewpoint change.

3.2.2 Reasoning

Cross-Frame Relative Direction (R1). R1 extends P1 to long horizons. It uses the same four-way relation, evaluated in the camera frame of the earlier observation, but the two landmarks are observed at least 180 seconds apart. With the relation space held fixed, P1 and R1 differ primarily in the temporal separation between the two observations.

Long-Range Route Distance Reasoning (R2). R2 asks the model to compare landmarks by the distance traveled along the recorded route rather than by straight-line proximity. Within each continuous trajectory segment, we parameterize the horizontal path by cumulative arc length s and define the route distance between an anchor landmark a and a candidate c as

$$d_{\text{route}}(a, c) = |s_c - s_a|. \tag{3}$$

Given four distinct candidates distributed along the route around the anchor, the correct answer is the one with the largest route distance. Questions that cross trajectory discontinuities or have ambiguous distance margins are discarded.

Sequential Turn Reasoning (R3). R3 extends P2 from a single motion event to an ordered pair of turns. We take two adjacent canonical events from the complete trajectory timeline and keep only clear left/right pairs, yielding one of $\{\text{LL}, \text{LR}, \text{RL}, \text{RR}\}$. A landmark before the first turn and a distinct landmark after the second delimit the queried portion of the route. Because adjacency is defined before QA filtering, no intervening canonical event can be skipped.

3.2.3 Decision

Decision tasks then put this route evidence to use: given a destination, the model must choose where to go next.

Destination-Directed Turn Selection (D1). D1 asks which local maneuver leads toward a destination. Regions that the trajectory revisits define decision nodes, and the distinct recorded departures from a node form alternative outgoing corridors, each associated with a left, right, or straight maneuver. Given a current landmark near the node and a later destination landmark, we keep only cases in which the destination lies downstream of exactly one corridor and the alternative corridors remain geometrically separated from it. The answer is the maneuver of this unique destination-connected corridor, rather than the direct bearing to the destination or an unconstrained shortest-path plan.

Destination-Directed Waypoint Selection (D2). D2 poses an analogous choice over waypoints instead of maneuvers. At a recurring region with multiple recorded branches, each branch is represented by a later branch-specific destination and by its first eligible landmark, which serves as the waypoint. Given the current region and a destination landmark, the model must select the waypoint on the unique branch that leads toward that destination. As in D1, the answer follows the recorded branch–destination correspondence rather than unconstrained shortest-path planning.

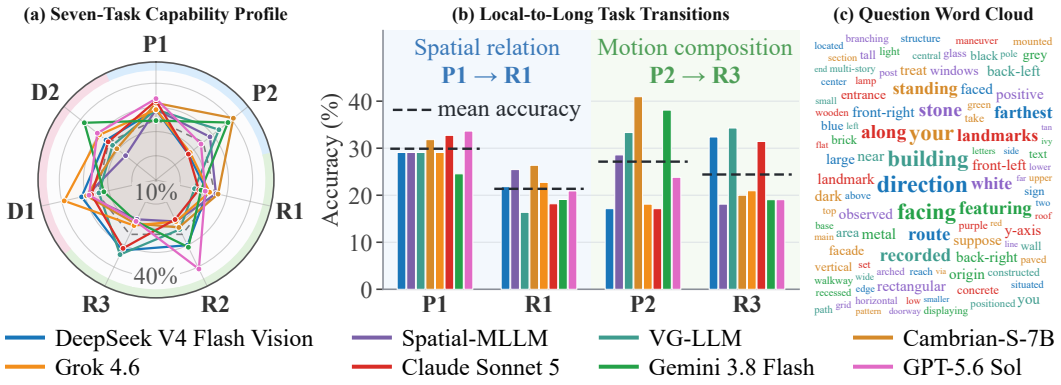


Figure 4: Diagnostic views of SpatialRoute-Bench. (a) Seven-task accuracy profiles across evaluated VLMs. (b) Accuracy transitions from local perception to long-horizon reasoning for the matched task pairs P1→R1 and P2→R3; dashed lines indicate mean task accuracy. (c) Word clouds of the question vocabulary. For visual clarity, panels (a) and (b) show eight selected models.

3.3 Benchmark Composition and Statistics

Table 1 compares SpatialRoute-Bench with representative video spatial benchmarks. Whereas existing benchmarks mainly target egocentric observation, spatial memory, streaming reasoning, or global spatial awareness, SpatialRoute-Bench focuses on continuous outdoor traversal, where spatial evidence is spread over extended trajectories and changing viewpoints. Its seven tasks jointly cover local perception, cross-time reasoning, and destination-directed decisions. To construct the benchmark, we first reconstructed and semantically annotated a pool of several hundred long outdoor traversal videos. Three annotators then manually inspected the reconstruction quality, derived spatial relations, and semantic labels of each video, and selected 22 hour-long videos with the most reliable geometry, spatial relations, and landmark annotations. All QA pairs are then generated automatically from these videos.

SpatialRoute-Bench contains 7,000 multiple-choice QA pairs, balanced at 1,000 per task. The QA pairs are derived from 22 hour-long outdoor traversal videos, as each video requires dense geometric and semantic annotation over its full trajectory and supports hundreds of long-horizon questions. As shown in Figure 3, the tasks occupy distinct temporal and geometric regimes. For example, P1 and R1 share the same relation space but differ substantially in temporal span. Beyond temporal separation, the QA pairs also span a wide range of spatial displacements and viewpoint changes. This variation supports the diagnostic analyses in Section 4.

4 Experiments

4.1 Experimental Setup

We evaluate ten representative VLMs spanning open-weight/open-lineage and proprietary systems (Table 2). The former group includes DeepSeek V4 Flash Vision (DeepSeek, 2026), Spatial-MLLM (Wu et al., 2026), VG-LLM (Zheng et al., 2026), and Cambrian-S-7B (Yang et al., 2026), with the latter three specifically developed for spatial or geometry-aware visual understanding. The proprietary group comprises Grok 4.6 (xAI, 2026), Gemini 3.7 Flash (Google DeepMind, 2026a), Claude Sonnet 5 (Anthropic, 2026), Gemini 3.8 Flash (Google DeepMind, 2026b), GPT-5.6 Terra (OpenAI, 2026), and GPT-5.6 Sol (OpenAI, 2026). All models are evaluated zero-shot under the same protocol.

To control evaluation cost while preserving coverage across videos and task types, we randomly sample up to five QA instances from each task in every video. The same sampled questions are used across all models to ensure a consistent comparison. The question vocabulary is dominated by spatial relations, landmark attributes, camera motion, route structure, and destination-related concepts, as summarized by the word clouds in Figure 4(c).

Table 2: Main results on SpatialRoute-Bench. P1–P2 denote perception tasks, R1–R3 denote reasoning tasks, and D1–D2 denote decision tasks. Best and second-best scores are highlighted in **bold** and underlined, respectively. All tasks are evaluated by multiple-choice accuracy (%); Avg. denotes the overall accuracy across all evaluated questions.

Model	Perception		Reasoning			Decision		Avg.
	P1	P2	R1	R2	R3	D1	D2	
Open-weight and Open-lineage Models								
DeepSeek V4 Flash Vision	29.09	17.14	21.82	30.00	32.38	31.58	26.44	26.87
Spatial-MLLM	29.09	28.57	25.45	19.09	18.10	29.47	16.09	23.82
VG-LLM	29.09	33.33	16.36	22.73	**34.29**	24.21	22.99	26.18
Cambrian-S-7B	31.82	**40.95**	**26.36**	21.82	20.00	29.47	20.69	27.42
Proprietary Models								
Grok 4.6	29.09	18.10	22.73	19.09	20.95	**38.95**	29.89	25.21
Gemini 3.7 Flash	19.09	21.90	**26.36**	31.82	**34.29**	24.21	26.44	26.32
Claude Sonnet 5	32.73	17.14	18.18	18.18	31.43	27.37	25.29	24.24
Gemini 3.8 Flash	24.55	38.10	19.09	30.91	19.05	22.11	**37.93**	27.15
GPT-5.6 Terra	**33.64**	22.86	23.64	29.09	24.76	27.37	20.69	26.18
GPT-5.6 Sol	**33.64**	23.81	20.91	**40.91**	19.05	28.42	31.03	**28.25**

4.2 Main Results

Table 2 shows SpatialRoute-Bench remains far from solved. Every model stays close to the 25% random baseline: the strongest model, GPT-5.6 Sol, reaches only 28.25% overall accuracy, followed by Cambrian-S-7B at 27.42%, and all ten models fall within a narrow range of 23.82–28.25%. Averaged over models, task accuracy ranges from 22.09% on R1, the only task whose average falls below chance, to 29.18% on P1, and even the highest single-task score, 40.95% by Cambrian-S-7B on P2, remains far from reliable performance. Model family does not clearly separate the models: proprietary models (24.24–28.25%) and open-weight models (23.82–27.42%) occupy largely overlapping ranges, and the top two models, one from each group, differ by only 0.83 percentage points.

This narrow overall range, however, conceals large differences at the task level. The gap between the best and worst model reaches 23.81 points on P2 and 21.84 points on D2, compared with only a 4.43-point spread in overall accuracy. Seven of the ten models achieve the best score on at least one task, yet none leads on more than two, and strengths rarely carry over across tasks. Gemini 3.7 Flash, for example, has the lowest P1 accuracy (19.09%) yet ties for the best score on both R1 and R3, while Grok 4.6 leads on D1 (38.95%) but ranks near the bottom on P2 (18.10%). The top two models illustrate the same contrast: despite differing by only 0.83 points overall, GPT-5.6 Sol outperforms Cambrian-S-7B by 19.09 points on R2 (40.91% vs. 21.82%), whereas Cambrian-S-7B leads by 17.14 points on P2 (40.95% vs. 23.81%). Similar overall scores can therefore arise from substantially different spatial capabilities, which motivates the finer-grained analysis in Section 4.3.

4.3 Discussion

Capability analysis from perception to decision. Accuracy does not decline steadily along the perception–reasoning–decision hierarchy (Figure 4): averaging task accuracies over models gives 27.69% on perception, 24.63% on reasoning, and 27.03% on decision, so failures must be traced task by task rather than attributed to the level of the hierarchy. Local perception is already fragile. Only four models exceed 30% on P1, and six of the ten models score below 25% on P2. Even the strongest P2 models, Cambrian-S-7B (40.95%) and Gemini 3.8 Flash (38.10%), recover local camera motion only partially: because reverse serves only as a distractor, a model that merely rules it out would reach about 33%, only 5–8 points below them. Longer horizons further degrade relational reasoning. With the same four-way relation, mean accuracy falls from 29.18% on P1 to 22.09% on R1 (-7.09 points), and nine of the ten models score lower. Because R1 pairs also span larger displacements and viewpoint changes, this drop reflects the long-horizon setting as a whole rather

Table 3: Diagnostic perturbations using **GPT-5.6 Sol**. Original Video corresponds to the GPT-5.6 Sol entry in Table 2. Shuffled Video randomly permutes the temporal order of input frames, Black Video removes all visual content while preserving the input length, and Shuffled Options randomly permutes answer positions. All values are multiple-choice accuracy (%); Avg. denotes micro-averaged accuracy across all evaluated questions.

Setting	P1	P2	R1	R2	R3	D1	D2	Avg.
Original Video	33.64	23.81	20.91	40.91	19.05	28.42	31.03	28.25
Shuffled Video	30.00	18.10	22.73	18.18	20.95	29.47	21.84	22.99
Black Video	23.64	18.10	20.91	20.00	18.10	22.11	20.69	20.50
Shuffled Options	35.45	20.95	21.82	28.18	22.86	31.58	29.89	27.15

than elapsed time alone. Unlike this consistent drop, motion composition is highly model-dependent. P2 and R3 have similar mean accuracies (26.19% vs. 25.43%), yet three models gain and three lose more than 10 points from P2 to R3, with the two strongest P2 models showing the largest drops, so single-event recognition does not predict sequential turn reasoning. Finally, destination-directed decisions remain unsolved, with average accuracies of 28.32% on D1 and 25.75% on D2. Since decision accuracy is not lower than reasoning accuracy, decision failures do not appear to be simply inherited from the reasoning level, although all three level averages lie within three points of chance.

Cross-model analysis. The overall ranking in Section 4.2 should be interpreted with caution: the full 4.43-point spread across the ten models is only modestly larger than the 95% confidence interval of a single model's overall accuracy (about ± 3.2 points). Within this narrow range, we observe no consistent aggregate advantage for either model family or spatially trained models. The three spatially trained models span a wide range: Cambrian-S-7B ranks second overall (27.42%), VG-LLM lies in the middle (26.18%), and Spatial-MLLM ranks last (23.82%). Although Cambrian-S-7B and VG-LLM achieve top scores on P2 and R3, respectively, comparable peaks are reached by general-purpose models: Gemini 3.8 Flash comes within three points of Cambrian-S-7B on P2, while Gemini 3.7 Flash ties VG-LLM on R3. These task-wise strengths cannot be attributed to spatial supervision alone without controlled base-model comparisons. Overall, models are better distinguished by their task-level capability profiles than by model family or training category.

Benchmark validity analysis. Since every model remains close to chance, we examine whether performance depends on visual and temporal evidence by perturbing the input of GPT-5.6 Sol (Table 3). Replacing the video with black frames lowers overall accuracy from 28.25% to 20.50%, with decreases on six of the seven tasks and below-chance performance on all seven. This suggests that question and option text alone are insufficient to sustain the original performance.

Shuffling the temporal frame order reduces overall accuracy to 22.99%, with the largest drop on R2 from 40.91% to 18.18% (-22.73 points, 25 fewer correct answers). Smaller drops occur on D2 and P2, while other tasks change modestly. Meanwhile, shuffling option choices yields 27.15% accuracy (a minor 1.10-point drop), showing little reliance on option bias. Together, these perturbations confirm that visual content and temporal order contribute meaningfully to performance despite low overall scores.

5 Conclusion

We introduced SpatialRoute-Bench, a geometry-grounded benchmark that evaluates long-horizon spatial understanding in outdoor videos through a three-level hierarchy of perception, reasoning, and decision. Its fully automated pipeline couples 3D reconstruction with VLM-based semantic grounding: semantics determines which landmarks can be referred to, whereas all answers are derived from geometry recovered from the video itself. Evaluating ten representative VLMs, we find that all of them remain close to chance, that camera-motion understanding remains unreliable, and that the same spatial relation becomes harder to recover when its landmarks are observed minutes apart. SpatialRoute-Bench is currently limited by reconstruction quality, relative rather than metric scale, and recorded rather than interactive trajectories. Nevertheless, it offers a scalable testbed for developing VLMs with persistent spatial memory and reliable long-horizon spatial reasoning.

ETHICS STATEMENT

SpatialRoute-Bench is constructed from publicly available outdoor videos. Our benchmark focuses on camera motion, pointmap, and static landmarks, and does not annotate or evaluate personal identities or sensitive attributes.

AI USE STATEMENT

In this work, we used LLM to improve the clarity, grammar, and presentation of the manuscript, and to assist in checking potential overlap in wording and research framing with related work.

REFERENCES

Alibaba Cloud. Alibaba cloud model studio: Visual understanding models, 2026. URL https://www.alibabacloud.com/help/en/model-studio/vision-model. Accessed: 2026-09-26.

Anthropic. Claude sonnet 5 system card, June 2026. URL https://www.anthropic.com/system-cards.

Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 964–972, 2021.

Keshigeyan Chandrasegaran et al. HourVideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026.

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. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

DeepSeek. Deepseek-v4-flash-vision-exp release: Multimodal api now live. https://api-docs.deepseek.com/news/news260821/, August 2026. Accessed: 2026-09-25.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

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

Google DeepMind. Gemini 3.7 flash: Model card, August 2026a. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Google DeepMind. Gemini 3.8 flash: Model card, September 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Tianjun Gu, Tianyu Xin, Kuan Zhang, Bowen Yang, Kok-Chung Chua, Peize Li, Xinran Zhang, Yupeng Chen, Qiyue Zhao, Qinlei Xie, et al. Towards spatial supersensing in the wild. *arXiv preprint arXiv:2607.13681*, 2026.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2616–2625, 2017.

Shiqiang Lang, Jing Liu, Haoyang He, Peiwen Sun, Yuanteng Chen, Tao Liu, Lan Yang, Longteng Guo, and Honggang Zhang. Longspace: Exploring long-horizon spatial memory from perception to recall in video. *arXiv preprint arXiv:2606.05677*, 2026.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. Ovo-s-bench: A hierarchical benchmark for streaming spatial intelligence in multimodal llms. *arXiv preprint arXiv:2606.03890*, 2026.

Aravindh Mahendran et al. KilometerVision: A new frontier for large-scale spatial intelligence in VLMs. In *European Conference on Computer Vision (ECCV)*, 2026.

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. *Advances in neural information processing systems*, 31, 2018.

OpenAI. Gpt-5.6 system card, July 2026. URL `https://deploymentsafety.openai.com/gpt-5-6`.

Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. *arXiv preprint arXiv:1702.08360*, 2017.

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

Emilia Szymańska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark. In *European Conference on Computer Vision*, pp. 68–85. Springer, 2024.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4626–4635. IEEE, 2025.

Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep it in mind: User centric continual spatial intelligence reasoning in egocentric video streams. *arXiv preprint arXiv:2606.15200*, 2026.

Yuxi Wei, Wei Huang, Qirui Chen, Lu Hou, and Xiaojuan Qi. See, remember, explore: A benchmark and baselines for streaming spatial reasoning. *arXiv preprint arXiv:2603.23864*, 2026.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 2026.

xAI. Introducing grok 4.6, August 2026. URL `https://x.ai/news/grok-4-6`.

Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29501–29512. IEEE, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025b.

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

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. corr, abs/2505.23764, 2025b. doi: 10.48550. *arXiv preprint ARXIV.2505.23764*.

Mengmi Zhang, Keng Teck Ma, Shih-Cheng Yen, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Egocentric spatial memory, 2018. URL https://arxiv.org/abs/1807.11929.

Qifeng Zhang, Kaixiang Huang, Heng Dong, Huang Fang, Junting Chen, Junjie Zhu, Yonghang Chen, Zhiyu Zhang, and Wei Li. Gst-bench: Can vlms develop global spatial awareness from video? *arXiv preprint arXiv:2608.05747*, 2026.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6612–6619. IEEE, 2017.

A Appendix

A.1 Qualitative QA Examples

Figures 5–11 present representative QA examples from the seven SpatialRoute-Bench tasks. Each example shows the sampled video input, the multiple-choice question and pseudo ground-truth answer, representative evidence frames, and the corresponding geometry reconstructed from the video. We also report model predictions and a brief analysis to illustrate typical successes and failures. Together, these examples provide a qualitative view of how SpatialRoute-Bench grounds questions in visual observations and reconstructed trajectory geometry across perception, reasoning, and decision tasks.

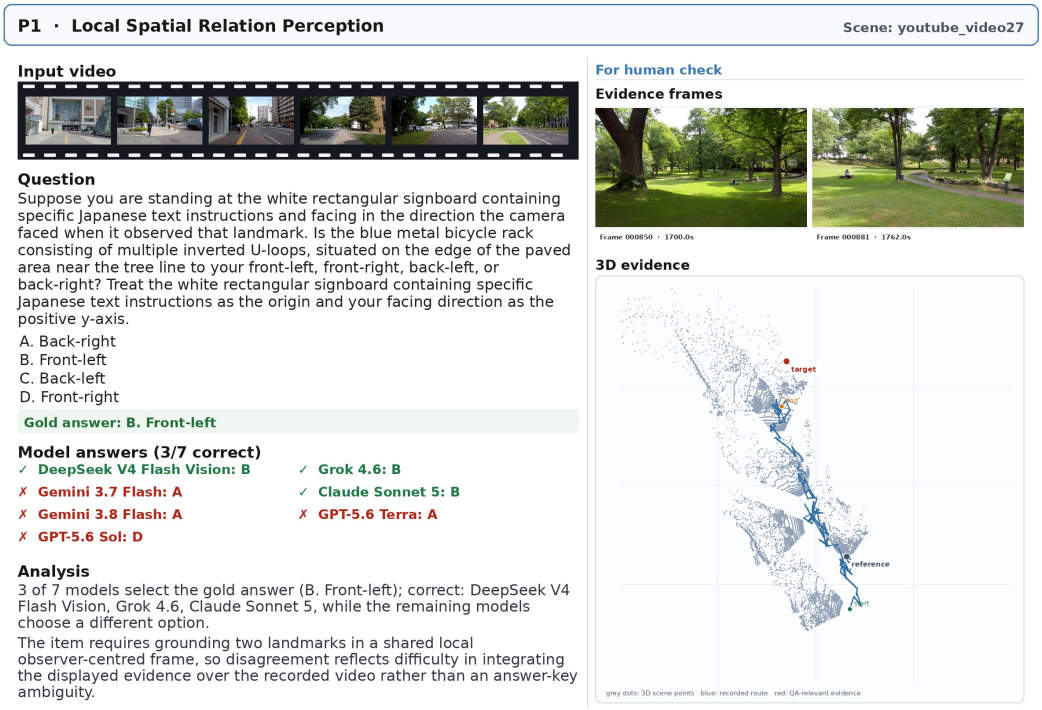


Figure 5: Representative P1 example for local spatial relation perception.

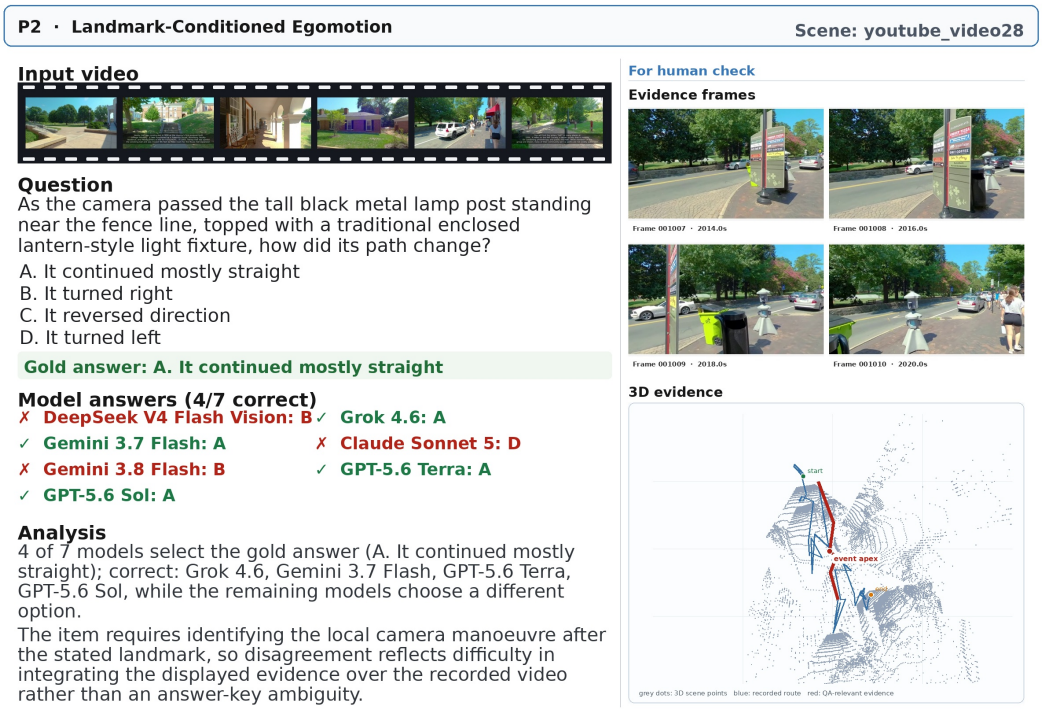


Figure 6: Representative P2 example for landmark-conditioned egomotion perception.

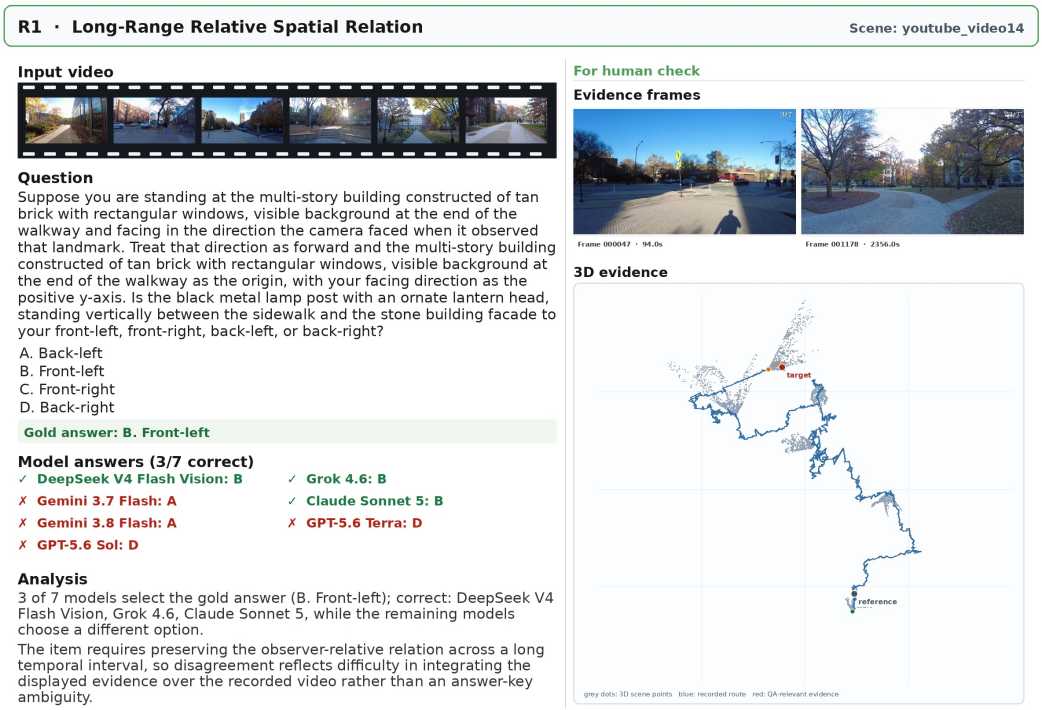


Figure 7: Representative R1 example for long-range relative spatial reasoning.

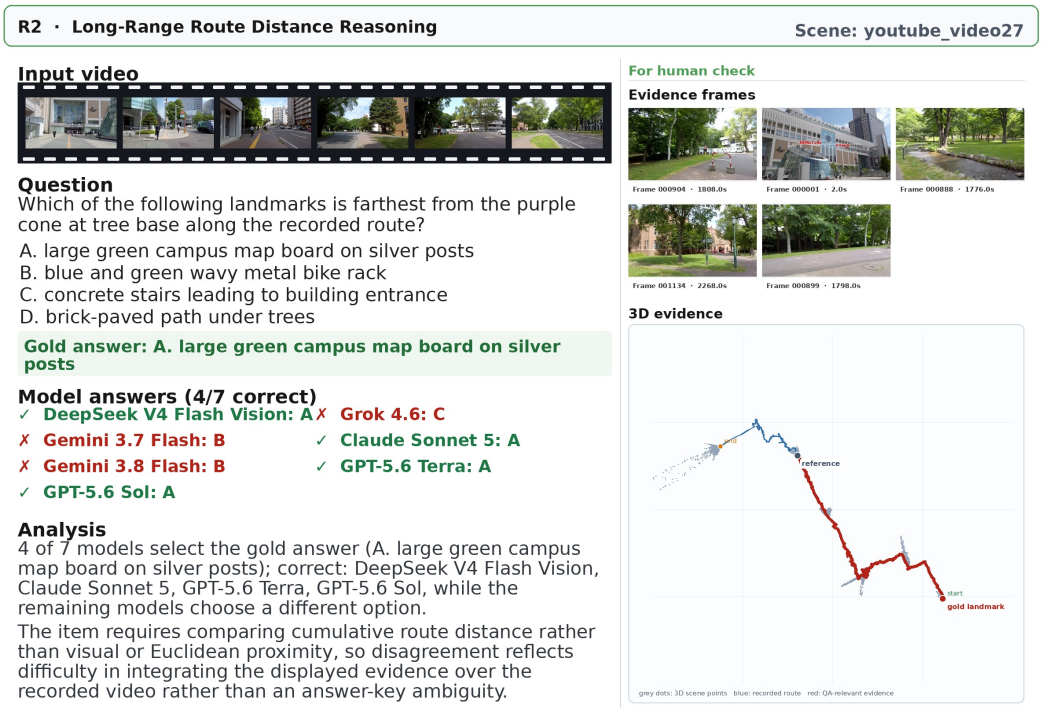


Figure 8: Representative R2 example for long-range route-distance reasoning.

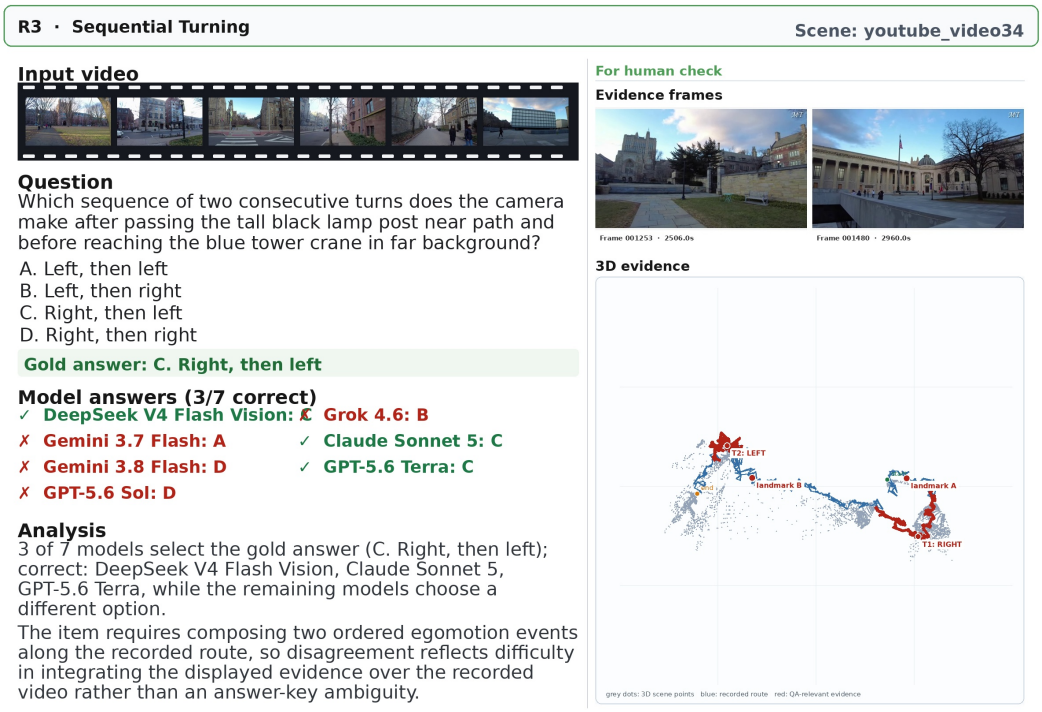


Figure 9: Representative R3 example for sequential turn reasoning.

D1 · Goal-Directed Turn Selection — Scene: youtube_video6

Input video



Question

At the branching area near the vertical corner edge of the grey stone structure where the rusticated facade meets the side wall, showing the depth of the masonry blocks, which maneuver should the camera take to reach the corner of a light blue building featuring distinct horizontal rusticated stone grooves?

A. Continue mostly straight
B. Reverse direction
C. Turn left
D. Turn right

Gold answer: C. Turn left

Model answers (3/7 correct)

- ✓ DeepSeek V4 Flash Vision: C
- ✗ Gemini 3.7 Flash: unparsed
- ✗ Gemini 3.8 Flash: B
- ✗ GPT-5.6 Sol: D
- ✓ Grok 4.6: C
- ✗ Claude Sonnet 5: A
- ✓ GPT-5.6 Terra: C

Analysis

3 of 7 models select the gold answer (C. Turn left); correct: DeepSeek V4 Flash Vision, Grok 4.6, GPT-5.6 Terra, while the remaining models choose a different option.

The item requires selecting the recorded initial manoeuvre conditioned on a downstream goal, so disagreement reflects difficulty in integrating the displayed evidence over the recorded video rather than an answer-key ambiguity.

For human check

Evidence frames



3D evidence

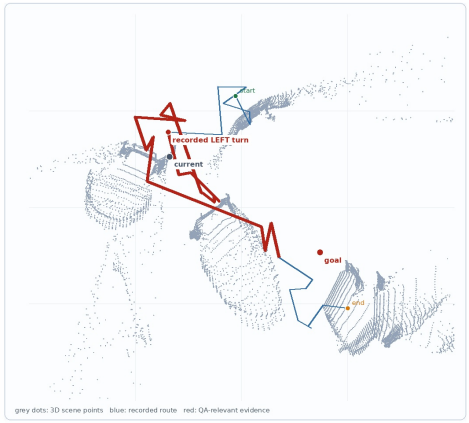


Figure 10: Representative D1 example for destination-directed turn selection.

D2 · Goal-Conditioned Waypoint Selection — Scene: youtube_video7

Input video



Question

At the recurring area near the central building featuring a prominent grey dome and a white portico with tall classical columns, situated at the far end of the green lawn, which listed waypoint should the camera take to follow the recorded route toward the grand classical building entrance featuring a triangular pediment supported by tall white columns, contrasting with the dark brick facade of the surrounding structure?

A. The waypoint near the detailed view of the purple brick building featuring a central white doorway, symmetric windows with white frames, and a dark slate roof with dormers.
B. The waypoint near the corner section of a grey stone building featuring a distinctive arched window on the upper level and rectangular windows below.
C. The waypoint near the multi-story building constructed of grey stone blocks featuring regularly spaced rectangular windows, visible through gaps in the foreground foliage.
D. The waypoint near the tall, black metal lamppost featuring a decorative fluted base and a single spherical white glass globe at the top, standing on the edge of the lawn.

Gold answer: A. The waypoint near the detailed view of the purple brick building featuring a central white doorway, symmetric windows with white frames, and a dark slate roof with dormers.

Model answers (4/7 correct)

- ✗ DeepSeek V4 Flash Vision: D
- ✗ Gemini 3.7 Flash: D
- ✓ Gemini 3.8 Flash: A
- ✗ GPT-5.6 Sol: D
- ✓ Grok 4.6: A
- ✓ Claude Sonnet 5: A
- ✓ GPT-5.6 Terra: A

Analysis

4 of 7 models select the gold answer (A. The waypoint near the detailed view of the purple brick building featuring a central white doorway, symmetric windows with white frames, and a dark slate roof with dormers.); correct: Grok 4.6, Claude Sonnet 5, Gemini 3.8 Flash, GPT-5.6 Terra, while the remaining models choose a different option.

The item requires linking the specified goal to the correct next waypoint on the recorded branch, so disagreement reflects difficulty in integrating the displayed evidence over the recorded video rather than an answer-key ambiguity.

For human check

Evidence frames



3D evidence

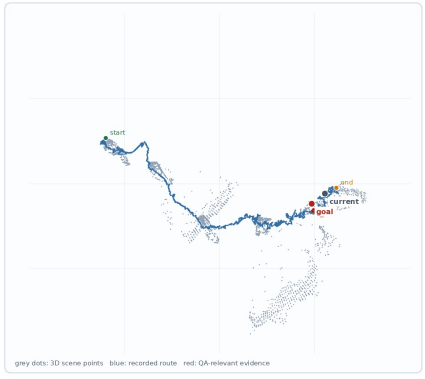


Figure 11: Representative D2 example for destination-directed waypoint selection.