Learning Situated Awareness in the Real World

Anonymous CVPR submission

Paper ID 48

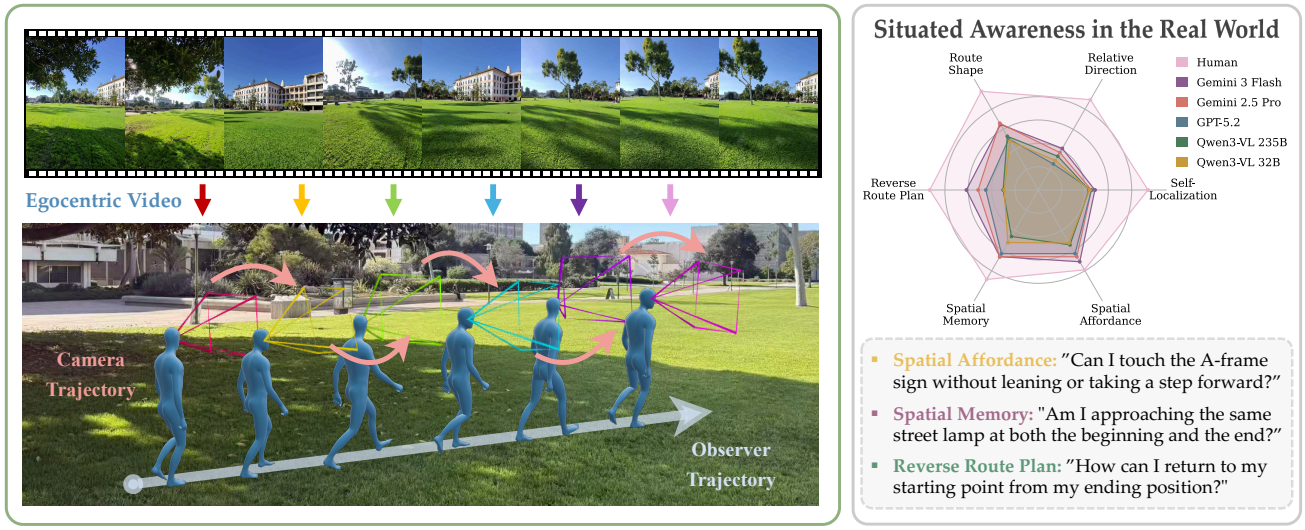

Figure 1. **(Left) Situated Awareness in the Real World.** A real-world example in which the observer walks along a straight trajectory while frequently rotating their head. The resulting egocentric video exhibits substantial camera orientation changes despite linear translational motion. **(Right) Reasoning Task Performance.** Radar plot compares human performance with representative MFMs across six situated awareness tasks in SAW-BENCH.

Abstract

A core aspect of human perception is situated awareness, the ability to relate ourselves to the surrounding physical environment and reason over possible actions in context. However, most existing benchmarks for multimodal foundation models (MFMs) emphasize ***environment-centric*** *spatial relations (relations among objects in a scene), while largely overlooking* ***observer-centric*** *relationships that require reasoning relative to agent's viewpoint, pose, and motion. To bridge this gap, we introduce* SAW-BENCH *(Situated Awareness in the Real World), a novel benchmark for evaluating egocentric situated awareness using real-world videos.* SAW-BENCH *comprises 786 self-recorded videos captured with Ray-Ban Meta (Gen 2) smart glasses spanning diverse indoor and outdoor environments, and over 2,071 human-annotated question-answer pairs. It probes a model's observer–centric understanding with six different awareness tasks. Our comprehensive evaluation reveals a human-model performance gap of 37.66%, even with the best-performing MFM, Gemini 3 Flash. Beyond this gap, our in-depth analysis uncovers several notable findings; for example, while models can exploit partial geometric cues in egocentric videos, they often fail to infer a coherent camera geometry, leading to systematic spatial reasoning errors. We position* SAW-BENCH *as a benchmark for situated spatial intelligence, moving beyond passive observation to understanding physically grounded, observer-centric dynamics. This work is accepted to ICML 2026.*

1. Introduction

Human perception of space is inherently *situated*. As people move through the world, they do not perceive scenes from a detached, global viewpoint; instead, they experience the environment relative to their own body and perspective.

At any given moment, a person maintains an implicit sense of their location, orientation, and the "intentional arc" of their movements [36]. Collectively, these **observer-centric** capabilities constitute **situated awareness** [13, 14, 39, 46], which operates continuously during everyday interaction and forms a foundational layer that supports more complex spatial intelligence.

Situated awareness is not only central to human perception, but also critical for autonomous systems operating in physical environments. In robotics, knowing *what* an object is is insufficient; an agent must also track precisely *where* it is relative to its own body to effectively plan grasping and navigation. Similarly, in augmented and virtual reality (AR/VR), systems must continuously synchronize virtual content with the users' physical perspective to maintain immersion. In both domains, failure in this translation decouples the system's understanding from physical reality. Cognitive science studies suggest that such spatial intelligence relies on path integration, where local, situated updates are accumulated to a larger observer-aware map [35]. Consequently, failures in situated awareness do not merely result in local errors; they create a cascading drift that significantly degrades the system's capability to understand space and time [50].

Yet, despite the importance of situate awareness, the current evaluation landscape has largely overlooked these observer-centric capabilities. While there is growing interest in spatial reasoning for MFMs, existing benchmarks, such as VSI-Bench [55] and MindCube [59], predominantly focus on observer-independent tasks. These benchmarks emphasize object-object interaction, discrete mental simulation, and metric distance estimation – skills that define spatial reasoning from a detached, static, and third person perspective. As a result, models are evaluated under the assumption that they are passive spectators rather than active embodied agents with their own view point, motion, and position. This leaves observer-centric situated awareness, the ability to understand space relative to the observer and how it evolves overtime, largely unexplored.

To bridge this gap, we introduce SAW-BENCH, **S**ituated **A**wareness in the Real **W**orld, a novel video understanding benchmark designed to assess MFMs' situated awareness capabilities. SAW-BENCH comprises 2,071 carefully curated, human-annotated multiple-choice question-answer pairs spanning six distinct awareness tasks (localization, relative direction, route shape, reverse route plan, spatial memory, and spatial affordance), applied to 786 egocentric videos encompassing both indoor and outdoor environments (§3.2). Videos in SAW-BENCH are all self-recorded using Ray-Ban Meta (Gen 2) glasses. The question-answer pairs are carefully designed to ensure that models need to reason about the observer itself and the environment, making situated awareness essential for solving our tasks (§3.1).

We conduct a comprehensive evaluation of 16 open-source and 8 proprietary MFMs. The best performing model, Gemini 3 Flash, yields a performance of 53.89% on SAW-BENCH, significantly below that of human-level performance (91.55%). Beyond this overall performance gap, we identify four systematic patterns in model behavior (§5): (1) models often conflate egocentric camera rotation with translational movement; (2) model accuracy degrades significant as trajectory complexity increases; (3) models exhibit a mismatch between view-level internal memory and persistent world-state memory; and (4) environment openness alone is an insufficient proxy for spatial reasoning difficulty.

2. Related Work

General video understanding benchmarks. Video understanding has become a central capability in MFMs, serving as a key testbed for evaluating their ability to perceive, reason, and integrate visual and linguistic information over time. Recent benchmarks emphasize more complex forms of reasoning such as long-form video understanding [6, 8, 48, 51, 62]; visual-temporal reasoning [31, 40, 53]; domain-specific reasoning [12, 21, 65]; and comprehensive video understanding [16, 29, 38]. Despite this growing complexity, these benchmarks remain largely allocentric, evaluating models as passive observers of scene-level events. Consequently, observer-centric spatial reasoning, understanding one's own position and relationship to the environment, remains largely unexamined, motivating the curation of SAW-BENCH.

3D spatial intelligence. Research in 3D spatial intelligence has predominantly focused on reasoning over explicit, reconstructed geometric representations such as point clouds and meshes [1, 22–24]. Early datasets like ReferIt3D [2] and ScanRefer [7] evaluate the capability to ground natural language into specific 3D coordinates, while more recent datasets like 3D-VisTA [68], 3D-GRAND [54], and EmbodiedScan [47] assess more complex grounding and reasoning over holistic 3D scenes. Among these datasets, SQA3D [34] is the most relevant to our work, as it explicitly introduces situated questions in a 3D scene (*e.g.* asking about the environment relative to a specific location given a 3D scene context). While prior work often relies on reconstructed 3D scenes, such data is costly to obtain and difficult to scale in real-world settings. In practice, self-recorded egocentric video is a lower-friction input to MFMs, as it requires neither 3D capture nor externally mounted cameras. For that reason, SAW-BENCH adopts egocentric video as its primary modality to evaluate situated awareness in realistic real-world settings.

Visual-spatial intelligence benchmarks. In the visual domain, existing benchmarks often focus on high-level semantic or commonsense reasoning, while overlooking fine-

Figure 2. **Overview of SAW-BENCH.** We illustrate six representative tasks (§ 3.1) evaluating different aspects of situated awareness: Self-Localization, Relative Direction, Route Shape, Reverse Route Plan, Spatial Memory, and Spatial Affordance. During data collection, human annotators follow pre-defined trajectories when recording egocentric videos (§ 3.2); these trajectories are visualized as purple dashed arrows. For all tasks, the model input consists solely of egocentric video without access to any bird's-eye or global scene representations; the visualizations shown here are provided for illustrative purposes only.

grained and precise spatial intelligence. Evaluations that do focus on spatial awareness largely frame spatial reasoning through observer-independent tasks, overlooking the inherently observer-centric nature of embodied tasks. Benchmarks such as VSI-Bench [55], VSI-Super [57], Spatial-Bench [52], and SpaCE-10 [20] emphasize object-object interaction, counting, and distance estimation, which treat models as a detached, external observer. Similarly, Mind-Cube [59], MomaGraph [26] and All-Angles Bench [58] assess spatial reasoning using discrete multi-view inference. While they evaluate models' ability to integrate spatial information across viewpoints, they typically assume the observer's state is given or static. As a result, models' capability for **situated awareness**, the continuous real time update from observer's own pose and perspective relative to the environment, largely unexplored. A more comprehensive comparison is provided in Table 1.

3. Situated Awareness Benchmark

3.1. Situated Awareness Tasks in SAW-BENCH

While cognitive science does not prescribe a canonical task taxonomy for situated awareness, prior work across navigation [5, 45], spatial updating [15, 37], spatial working memory [32, 42], and affordance [18, 19] has studied these abilities as separable components, each capturing a distinct aspect of how observers relates themselves to the environment. Motivated by this decomposition, we introduce six situated awareness tasks over 2,071 human-annotated question-answer pairs derived from 786 videos, each of which requires models to understand and reason over the relationship between itself and the environment:

- **Self-Localization** (9.92%): infer the observer's position within the environment from an egocentric viewpoint;
- **Relative Direction** (40.27%): reason about the ob-

Table 1. **Summary of Visual-Spatial Reasoning Benchmarks.** We compare existing works from: [1]Egocentric setting, [2]Self-collected data, [3]Input modality, [4]Self-localization tasks, [5]Motion Understanding, [6]Trajectory Reasoning, [7]Spatial Memory, and [8]Action Feasibility.

Benchmark	Ego.	Self-Col.	Mod.	Self-Loc.	Motion	Traj.	Mem.	Act. Feas.
Spatial-MM [41]	✓*	✗	Images	✗	✗	✗	✗	✗
ViewSpatial Bench [28]	✓*	✗	Images	✗	✗	✗	✗	✗
Open3D-VQA [63]	✗	✗	Images	✗	✗	✗	✗	✗
SpatialBench [52]	✓	✓	Videos	✗	✗	✗	✗	✗
MindCube [59]	✗	✓*	Images	✗	✗	✗	✗	✗
3SDRBench [33]	✗	✗	Images	✗	✗	✗	✗	✗
All-Angles Bench [58]	✓	✗	Images	✗	✗	✗	✓*	✗
VSI-Super Recall [57]	✓	✗	Videos	✗	✗	✗	✓*	✗
VSI-Super Count [57]	✓	✗	Videos	✗	✗	✗	✓*	✗
OmniSpatial [25]	✓*	✗	Images + Videos	✗	✓*	✗	✗	✓*
SpaCE-10 [20]	✓*	✗	Images + Videos	✗	✗	✓*	✓*	✗
MomaGraph [26]	✓	✗	Images	✗	✗	✓*	✗	✓*
VLM4D [66]	✓*	✗	Videos	✗	✓	✗	✗	✗
MMSI-Bench [56]	✓*	✗	Images	✗	✓	✗	✗	✗
VSI-Bench [55]	✓	✗	Videos	✗	✓	✗	✓*	✗
MMSI-Video-Bench [30]	✓*	✓*	Videos	✗	✓	✗	✓	✗
SAW-BENCH (Ours)	✓	✓	Videos	✓	✓	✓	✓	✓

Figure 3. **Benchmark Curation Pipeline.** We first pre-define 37 camera trajectories and annotate their metadata (details are provided in §D). Human video collectors then record egocentric videos by following these trajectories in selected scenes. Low-quality recordings are filtered and re-captured to ensure consistent video quality.

server's relative position across time by relating starting and ending viewpoints;

- **Route Shape** (26.36%): characterize the geometric shape of the observer's movement trajectory;
- **Reverse Route Plan** (11.06%): infer a sequence of movements that returns the observer to the starting location;
- **Spatial Memory** (4.83%): reason about changes in the environment by comparing spatial information across time;
- **Spatial Affordance** (7.82%): determine whether a specific action is feasible under physical constraints from the observer's viewpoint.

3.2. Video Collection

Egocentric video is a natural sensing modality for studying situated awareness, as it captures the environment from the observer's own viewpoint and preserves the observer-centric structure of spatial perception during real-world interaction. Unlike third-person footage, egocentric video directly encodes where objects appear relative to the camera wearer, how the field of view evolves with head and body movement, and how visibility of the environment changes over time.

To reflect this setting, all videos in SAW-BENCH are recorded from an egocentric perspective using Ray-Ban Meta (Gen 2) smart glass worn by human participants. Most videos are captured as single, continuous clips without interruption. For tasks involving **Spatial Memory**, we apply limited post-processing by concatenating two short clips recorded in the same physical scene: one before and one after a controlled modification of the environment. No other temporal reordering or editing is performed. Audio is excluded from all videos to ensure that all reasoning is grounded solely in visual information.

Our video collection process spans a diverse set of real-world environments, including 10 outdoor scenes (*e.g.* courtyards, parking lots, lawns, and plazas) and 5 indoor scenes (*e.g.* lecture halls, classrooms, recreation rooms, and household environments). Within each scene, we collect approximately 40-60 distinct videos to support tasks that benefit from dense coverage of a fixed environment, such as **Self-Localization** and **Route Shape**. For tasks that are more difficult to scale within a limited set of scenes, par-

ticularly **Spatial Memory** and **Spatial Affordance**, we additionally collect a set of videos across a broader range of environments outside these core scenes. This supplemental collection prioritizes diversity over dense coverage, enabling evaluation of memory and action feasibility across varied layouts and physical constraints without requiring exhaustive sampling of each scene.

During video collection, participants followed a lightweight recording protocol, consisting of high-level guidelines intended to ensure consistency across scenes while preserving natural behavior. For tasks involving **Self-Localization**, participants were instructed to record videos from a set of predefined reference locations (*e.g.* corners, side, or center) to ensure coverage of diverse viewpoints. Beyond these coverage requirements, the protocol did not prescribe specific paths, motions, or camera poses. Instead, participants were instructed to follow coarse trajectory shapes (*e.g.* zigzag or two consecutive turns), while retaining flexibility in how these shapes were executed within each environment. Recording protocols are provided in §D.1.

3.3. QA Annotation and Quality Check

Question-answer annotation. Question-answer (QA) pairs in SAW-BENCH are annotated based on the predefined recording protocol and the known trajectory of each video. All QA pairs are annotated by the same human participants who recorded the videos, who followed predefined reference locations and coarse trajectory shapes during data collection. As a result, annotation is restricted to these predefined configurations rather than open-ended interpretation of the video content. This design allows questions to target well-defined aspects of agent-centric situated awareness while minimizing ambiguity in ground-truth answers.

Quality control. We perform quality check at both the video and annotation levels. For video quality control, each recorded video is manually reviewed by human reviewers. Videos exhibiting issues such as rapid head motion, poor visibility of key objects, or other factors that could impair spatial reasoning were discarded and re-filmed following the same recording protocol. To ensure high-quality QA annotation, each QA pair in the protocol is independently annotated by two annotators. We report inter-annotator agreement score in §D.2. Disagreements were resolved through a final review process following the same annotation guidelines. This rigorous approach ensures consistency and accuracy across all annotated QA pairs.

4. Experiments

4.1. Evaluation Protocol

Models. We evaluate a diverse set of general-purpose MFMs in a zero-shot setting. For *open-source* MFMs, we test: Qwen2.5-VL [4], Qwen3-VL [3], InternVL 2 [9], InternVL 3 [67], LLaVA-NeXT-Video [64], and LLaVA-OneVision [27]. We also evaluate the following *proprietary* MFMs: GPT [43] and Gemini [11, 17]. We provide full model evaluation configuration in Table 4.

Baselines. We include five baselines: (1) *chance level (random)*, which is defined as the expected accuracy under uniform random answer selection; (2) *chance level (frequent)*, which is defined as the accuracy achieved by always selecting the most frequent answer; (3) *blind LLM*, which answer the multiple-choice questions without access to any visual information from the video. We use GPT-5.2 as our blind evaluation for this baseline. Details are provided in §C.1; (4) *socratic models* [61], which generate a single holistic caption for each video using a video captioner and use this caption as a language-based representation of the video for downstream question answering. In evaluation, the model is provided with the question and the caption only. We use GPT-5.2 for both video captioning and question answering. Details are provided in §C.2; (5) *human level*, which is measured based on the answers given by two graduate students, who have access to the full videos and no time constraints. Details are provided in §C.3.

Accuracy evaluation. We use accuracy as the primary metric to evaluate model performance on SAW-BENCH. Following recent benchmarks for foundation model evaluation [40, 49, 65], we first apply a regular-expression–based parser to extract the predicted answer from each model's raw response. If the parser fails, we additionally use GPT-4o-mini to extract the answer from the raw output. Prompt used for answer extraction is provided in §B.3.

4.2. Main Results

We provide quantitative results on SAW-BENCH for all models in Table 2.

Widespread difficulty in situated awareness. Our evaluation reveals that situated awareness remains a fundamental challenge for current MFMs. Even the top performing model, Gemini 3 Flash, achieves only 53.89% accuracy overall, leaving a 37.66% performance gap compared to human accuracy of 91.55%. Notably, humans perform remarkably well on **Self-Localization** and **Route Shape** tasks, suggesting that these categories are naturally intuitive to observers. Interestingly, the performance gap between human and top-performing MFMs is much smaller on **Spatial Memory** and **Spatial Affordance**, highlighting that current models may be relatively strong on tasks that rely more heavily on spatial memorization and depth cues.

Proprietary *vs.* open-source model. In general, proprietary MFMs outperform open-source MFMs, with the largest performance gap appearing on **Reverse Route Plan**, a task that requires sustained reasoning over egocentric trajectories and explicit tracking of intermediate movements.

Table 2. **Evaluation Results on SAW-BENCH.** Unless otherwise specified, all models process videos at 2 fps (frames per second). Frame level sensitivity analyses are provided in §F. **Bold** and underlined numbers indicate the best and second-best performance in each category, respectively. Model configurations are provided in Table 4. †: Human baseline details are provided in §C.3. *: Models do not support fps-based sampling and process a fixed total of 32 frames per video. ‡: 8 frames per video sampling due to compute limitations.

Model	All	Self-Localization	Relative Direction	Route Shape	Reverse Route Plan	Spatial Memory	Spatial Affordance
Baselines							
Human Level †	**91.55**	**94.00**	**89.39**	**97.62**	**93.01**	**88.50**	**79.01**
Chance Level (Random)	27.49	34.00	25.90	21.43	27.51	28.00	56.17
Chance Level (Frequent)	29.55	38.00	25.90	27.11	27.51	27.00	50.62
Blind LLM (GPT-5.2)	31.34	38.00	23.02	36.63	24.02	38.00	54.32
Socratic Model (GPT-5.2)	31.34	40.50	20.62	41.58	24.02	32.00	50.62
Proprietary Multimodal Foundation Models							
Gemini 3 Flash	**53.89**	48.50	**41.13**	64.84	**61.57**	**66.00**	**70.99**
Gemini 2.5 Pro	50.80	45.50	37.05	**66.12**	51.53	**66.00**	66.05
Gemini 3 Pro	45.97	**50.00**	38.61	52.01	36.24	63.00	61.73
GPT-5.2	41.04	45.50	25.78	50.55	44.98	63.00	62.96
Gemini 2.5 Flash	39.79	44.00	25.30	57.33	37.99	49.00	46.91
GPT-5 Mini	33.80	43.50	27.46	36.08	22.27	56.00	49.38
Open-Source Multimodal Foundation Models							
Qwen3-VL 235B-A22B	**41.40**	43.50	**33.41**	**53.11**	**30.13**	46.00	54.32
Qwen3-VL 32B	38.58	44.00	29.14	48.35	29.26	52.00	52.47
Qwen3-VL 30B-A3B	36.55	39.00	29.62	43.04	27.07	**54.00**	50.00
Qwen2.5-VL 32B	36.46	**53.00**	28.06	41.03	24.89	45.00	54.94
Qwen2.5-VL 72B	36.17	51.50	26.74	41.76	25.33	45.00	**56.79**
Qwen3-VL 8B	36.12	40.00	27.82	46.70	23.58	48.00	48.77
LLaVA OneVision 72B*	33.70	39.00	22.30	46.15	24.45	41.00	52.47
InternVL3 8B*	33.70	43.50	26.86	36.45	27.95	46.00	48.15
LLaVA-Video 72B*	32.98	32.50	23.86	43.04	24.45	41.00	53.70
InternVL3 14B*	32.69	49.00	17.27	45.05	24.02	54.00	49.38
Qwen2.5-VL 7B	31.48	38.50	19.06	43.59	26.20	38.00	49.38
LLaVA-NeXT-Video 32B*	31.24	41.00	24.46	35.35	22.27	34.00	51.23
LLaVA-Video 7B*	30.81	41.00	25.06	32.78	24.45	32.00	49.38
InternVL2 40B‡	30.13	45.00	17.75	38.28	24.89	32.00	54.32
InternVL2 8B*	29.84	43.00	14.99	41.94	24.89	40.00	50.00
LLaVA OneVision 7B*	29.45	34.50	20.26	34.80	25.33	44.00	49.38
InternVL3 38B‡	27.71	35.50	23.50	37.55	24.45	46.00	51.23

We provide a qualitative analysis on model responses in Figure 4 **Right**. In these cases, open-source model tends to rely on shortcuts cues from some "key frames," typically the first and the last frame of the video; whereas proprietary model more consistently maintain a coherent observer-centric representation across the full extent of the video.

Blind LLM *vs.* socratic model. The blind LLM achieves an overall accuracy of 31.34%, which is only marginally above chance level baselines, indicating that effective performance on SAW-BENCH requires access to visual information. Compared to Blind LLM, Socratic model does not yield significant performance gains, achieving the same overall accuracy of 31.34%. Although Socratic model has indirect access to visual content through video captioning, reducing egocentric video into a static language-based representation discards critical observer-centric cues such as viewpoint changes, orientation, and temporal structure. Notably, the Socratic model exhibits slightly better performance on Route Shape than the Blind LLM, suggesting that captions can convey coarse trajectory-level information. However, this limited improvement does not extend to other tasks. As a result, high-level semantic summaries alone are insufficient to situated awareness.

5. Analysis

To better understand when and why models fail at situated awareness, we analyze representative error patterns that reflect core components of observer-centric reasoning.

Camera rotation as a source of trajectory errors. We identify a systematic failure mode in Route Shape occurring when changes in camera rotation are decoupled from the observer's translational movement. To isolate this effect, we compare three controlled scenarios: (1) a straight path with stable head orientation (Figure 5 **Left**); (2) the same straight path with frequent head rotations (Figure 5 **Middle**); and (3) a true zigzag trajectory (Figure 5 **Right**).

Despite identical translational motion in cases (1) and (2), even top-performing models frequently misclassify case (2) as a zigzag trajectory: Gemini 3 Flash does so in 60.0% of instances, while Qwen3-VL 235B misclassifies 53.3% of cases. As illustrated in Figure 4 **Right**, models justify these predictions by erroneously attributing camera orientation shifts to physical body displacement. This failure highlights a fundamental limitation in current MFMs:

Question: *From my viewpoint at the end of the video, how can I go back to my starting point? Turn around, go straight. Turn left, go straight, then turn left and continue straight.*

Gemini 3 Flash: To determine the path back to the starting point, I first need to analyze the path taken from the start to the end of the video...The camera pans **left**...The camera **moves forward**...The camera **turns right** (clockwise rotation)...The camera **moves forward** along the lawn...The camera turns **right** (clockwise rotation) again...The camera **moves forward** towards...

To go back to the starting point from the end (facing the blue fence): ... **Rotate 180 degrees** to face ... **Move back** across the lawn towards .... must be reversed **with a left turn** ... **Move along** the buildings ... must be reversed **with a left turn...Move back towards**.

Qwen3-VL 235B: To determine the path back to the starting point, I analyze the video's movement. The **final frames** show a field, while the **starting point (first frame)** is a building with grass. **The video likely moved from the start by turning right and proceeding, then turning left to reach the field.** To reverse this path: turn around (opposite direction), go straight (back along the last segment), turn right (opposite of the previous left turn), go straight, then turn left (opposite of the initial right turn).

Question: *What's the shape of my moving trajectory?*
A. Zigzag B. Straight line C. Square D. L-shape

Gemini 3 Flash: The video starts with the camera facing a courtyard. The person moves towards a building on the **left**, then turns **right** to walk along the side of the building, and finally turns **left** to enter a walkway between buildings. This sequence of movements--**moving forward, turning right, and then turning left--forms a zigzag pattern.**

Gemini 2.5 Pro: The video begins with me standing in a courtyard, looking around. I first pan my head **left** and **right**. Then, I start **walking forward** across the paved area of the courtyard. After walking straight for a bit, I make a **distinct turn to my left** and proceed down a concrete pathway between the buildings. This movement pattern--walking straight and then making a roughly 90-degree turn--forms an L-shaped trajectory.

Qwen3-VL 235B: The video frames show the camera (person's perspective) moving in alternating directions--**left**, **right**, **left**, **right**--creating a **back-and-forth pattern**. This **alternating movement corresponds to a zigzag trajectory**, as opposed to a straight line, square, or L-shape, which involve fewer or no directional changes.

Figure 4. **Error Case Analysis.** (**Left**) **Reverse Route Plan**: Gemini 3 Flash successfully reconstructs the return path by systematically inverting the actions from the forward pass. In contrast, Qwen3-VL 235B attempts to exploit a shortcut between the first and last frames, thereby neglecting the transitive dynamics and spatial transformations occurring throughout the frame sequence. (**Right**) **Route Shape**: While both Gemini 3 Flash and Qwen3-VL 235B effectively identify camera rotations, they falsely integrate these rotational pans into the observer's physical movement trajectory, leading to incorrect shape understanding.

Figure 5. **Camera Rotation and Observer's Trajectory.** Visualization of three controlled scenarios used to isolate the impact of head rotation on **Route Shape**. **(Left)** a straight path with steady head orientation; **(Middle)** the same straight path with frequent left-and-right head rotations; and **(Right)** a true zigzag trajectory.

the inability to maintain a robust observer-centric coordinate system that distinguishes egocentric rotational pans from global positional updates.

> *Current MFMs often conflate **egocentric camera rotation** with **translational movement**.*

Trajectory complexity and error accumulation. Spatial updating is an inherently accumulative process, where errors in estimating egocentric motion compound as the observer moves through an environment [35, 44, 50]. In human navigation, this integration is highly sensitive to "noise" introduced by changes in orientation [10].

Table 3. **Accuracy (%) on Relative Direction Tasks Stratified by the Number of Turns**. Performance for most models degrades significantly as geometric complexity increases.

Models	Straight	Single Turn	Two Turns
Human	**100.00**	**96.67** (-3.33%)	**90.00** (-10.00%)
Gemini 3 Flash	73.33	70.69 (-3.60%)	40.61 (-44.63%)
Gemini 3 Pro	63.33	56.90 (-10.16%)	36.46 (-42.44%)
Gemini 2.5 Pro	73.33	55.17 (-24.76%)	33.41 (-54.44%)
GPT-5.2	30.00	39.66 (+32.20%)	22.49 (-25.03%)
Qwen3-VL 235B	90.00	8.62 (-90.42%)	27.85 (-69.06%)
Qwen3-VL 32B	80.00	12.07 (-84.91%)	21.83 (-72.71%)

To investigate whether MFMs exhibit a similar sensitivity to trajectory complexity, we stratify results on the **Relative Direction** task by geometric complexity: (1) **Straight** (pure translation), (2) **Single Turn** (one rotational update), and (3) **Two Turns** (multiple rotational updates). As shown in Table 3, increasing geometric complexity is associated with a substantial accuracy degradation for most models, particularly when trajectories involve multiple orientation changes. When quantified using relative performance drop with respective to straight trajectories, MFMs often exhibit significant degradation under multi-turn conditions, while human performance remains largely stable. This human-model gap suggests that current MFMs struggle to reliable integrate successive egocentric orientation changes over time, resulting in compounding errors as trajectories move away from simple translational motion.

Question: *Which object changes between earlier and later in the video?*
A. Phone *B. Fire hydrant* *C. A-frame sign* *D. Traffic cone*

Gemini 3 Flash: I observed a yellow fire hydrant throughout the video, which remained unchanged. I did not see a phone or a traffic cone. I then focused on the **A-frame sign**. In the earlier frames (e.g., 00:10.500), I looked at the lamp post on the left side of the path and confirmed that **no sign was present**. ... I observed the same lamp post and saw that an A-frame sign had been placed next to it. This clear appearance of the **sign** indicates it is the object that changed between the earlier and later parts of the video.

Gemini 2.5 Pro ...I see a yellow fire hydrant multiple times, starting around 00:00... In the frame at 00:21, a small **orange traffic cone** orange traffic cone ('D') appears on the grass next to the fallen A-frame sign. **This cone was not visible in earlier frames** (e.g., at 00:15) ... **This The traffic cone changes from not being present to being present**. ...

GPT-5.2 In the earlier frames there is **no A-frame sign visible**, while in later frames **an A-frame sign appears** near the walkway by a tree/lamp post, indicating that this object changes between earlier and later in the video.

Qwen3-VL 30B ... A-frame sign **is absent in earlier frames but appears later** but appears later (e.g., frame 21.2), so it changes ... Thus, C is correct.

Figure 6. **Model Responses in Spatial Memory.** Across multiple models, non-visibility is incorrectly treated as non-existence: objects that exit the camera's field of view are inferred to have disappeared or changed, revealing a gap between *what is seen* and *what exists*.

> *Model accuracy **degrades** significantly as trajectory complexity increases.*

Failure to maintain persistent object memory. A recurring failure mode across Spatial Memory tasks arises from models' difficulty in maintaining object persistence across egocentric motion. Although models often provide accurate descriptions of what is visible in individual frames or short temporal windows, they fail to reason about objects that leave the camera's field of view. As shown in Figure 6, models tend to infer that objects are absent in earlier frames simply because they are not visible, incorrectly treating first observation as object appearance rather than recognizing that the object may have existed outside the field of view. These errors suggest that current MFMs rely primarily on view-dependent evidence, rather than maintaining a persistent world-state representation over time.

> ***Persistent tracking** of objects across frames remains an open challenge across models.*

Effect of openness on situated awareness.

Figure 7 summarizes model performance across indoor and outdoor environments. Contrary to the intuition that larger and more dynamic outdoor environments may increase spatial reasoning difficulty, no consistent performance degradation is observed in outdoor scenes. Across the four selected models, outdoor performance is often comparable to, and in several cases higher than, indoor performance. On average, the indoor–outdoor performance gap remains small.

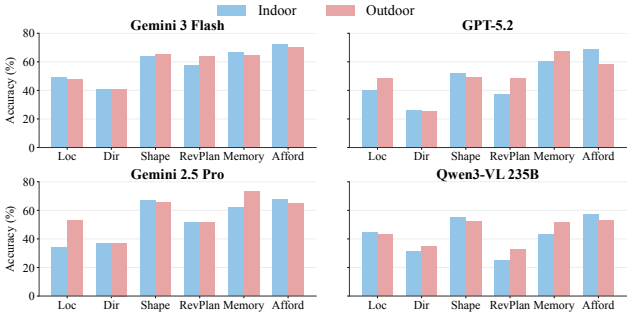

Figure 7. **Indoor *vs.* Outdoor Performance**. Comparison of zero-shot accuracy across six situated awareness tasks for Gemini 3 Flash, Gemini 2.5 Pro, GPT-5.2, and Qwen3-VL 235B.

These results suggest that environment scale alone does not determine spatial reasoning difficulty. While outdoor scenes typically span larger spatial extents, they often contain fewer objects and exhibit less structural clutter than indoor environments, which may reduce relational ambiguity. As a result, spatial reasoning difficulty is not monotonically correlated with scene size or openness. Instead, indoor environments can pose equally, if not more, complex spatial challenges due to higher object density and more intricate layout structures.

> *Environment openness alone is an **insufficient proxy** for spatial reasoning difficulty.*

6. Conclusion

Situated awareness underlies how humans continuously perceive, navigate, and act in the physical world, yet it remains insufficiently captured by existing multimodal evaluation frameworks. In this work, we introduce SAW-BENCH to explicitly evaluate observer-centric situated spatial understanding in MFMs using egocentric videos. Through a systematic evaluation of 24 models, we uncover fundamental gaps in current MFMs' ability to reason about observer-centric tasks. Our analysis identifies key factors underlying these limitations, offering insights for advancing MFMs toward more robust situated spatial intelligence. We hope this work sheds light on the development of AI systems that move beyond passive observation toward physically grounded, observer-centric, and interactive world understanding.

References

[1] Eslam Abdelrahman, Mohamed Ayman, Mahmoud Ahmed, Habib Slim, and Mohamed Elhoseiny. Cot3DRef: Chain-

of-Thoughts Data-efficient 3D Visual Grounding. In *ICLR*, 2024. 2

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *ECCV*, 2020. 2

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. 5

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[5] Neil Burgess. Spatial Memory: How Egocentric and Allocentric Combine. *Trends in Cognitive Sciences*, 10(12): 551–557, 2006. 3

[6] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-Hour Video-Language Understanding. In *NeurIPS*, 2024. 2, 15, 16

[7] Dave Zhenyu Chen, Angel X Chang, and Matthias Niessner. ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In *ECCV*, 2020. 2

[8] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understanding. In *ICLR*, 2025. 2

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*, 2024. 5

[10] Lucia A Cherep, Alex F Lim, Jonathan W Kelly, Devi Acharya, Alfredo Velasco, Emanuel Bustamante, Alec G Ostrander, and Stephen B Gilbert. Spatial Cognitive Implications of Teleporting through Virtual Environments. *Journal of Experimental Psychology: Applied*, 26(3):480, 2020. 7

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5

[12] Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. SciVideoBench: Benchmarking Scientific Video Reasoning in Large Multimodal Models. *arXiv preprint arXiv:2510.08559*, 2025. 2

[13] Mica R. Endsley. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 37:32 – 64, 1995. 2

[14] John M. Flach. Situation Awareness: Proceed with Caution. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 37:149 – 157, 1995. 2

[15] Nancy Franklin, Barbara Tversky, and Vicky Coon. Switching Points of View in Spatial Mental Models. *Memory & Cognition*, 20(5):507–518, 1992. 3

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *CVPR*, 2025. 2

[17] Gemini Team. A New Era of Intelligence with Gemini 3. 2025. Accessed: 2026-01-16. 5

[18] Eleanor J Gibson and Richard D Walk. The" Visual Cliff". *Scientific American*, 202(4):64–71, 1960. 3

[19] JJ Gibson. *The Ecological Approach to Visual Perception*. Psychology Press, 2014. 3

[20] Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Zhaokai Wang, Songyuan Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. SpaCE-10: A Comprehensive Benchmark for Multimodal Large Language Models in Compositional Spatial Intelligence. *arXiv preprint arXiv:2506.07966*, 2025. 3, 4

[21] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, and Xin Eric Wang. MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos. In *ICLR*, 2025. 2

[22] Joy Hsu, Jiayuan Mao, and Jiajun Wu. NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations. In *CVPR*, 2023. 2

[23] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view Transformer for 3D Visual Grounding. In *CVPR*, 2022.

[24] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds. In *ECCV*, 2022. 2

[25] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards Comprehensive Spatial Reasoning Benchmark for Vision Language Models. *arXiv preprint arXiv:2506.03135*, 2025. 4

[26] Yuanchen Ju, Yongyuan Liang, Yen-Jen Wang, Nandiraju Gireesh, Yuanliang Ju, Seungjae Lee, Qiao Gu, Elvis Hsieh, Furong Huang, and Koushil Sreenath. Moma-Graph: State-Aware Unified Scene Graphs with Vision-Language Model for Embodied Task Planning. *arXiv preprint arXiv:2512.16909*, 2025. 3, 4

[27] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. In *TMLR*, 2025. 5

[28] 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*, 2025. 4

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *CVPR*, 2024. 2

[30] Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, Wenbo Hu, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Video-Bench: A Holistic Benchmark for Video-Based Spatial Intelligence. *arXiv preprint arXiv:2512.10863*, 2025. 4

[31] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do Video LLMs Really Understand Videos? In *ACL*, 2024. 2

[32] Steven J Luck and Edward K Vogel. The Capacity of Visual Working Memory for Features and Conjunctions. *Nature*, 390(6657):279–281, 1997. 3

[33] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M de Melo, and Alan Yuille. 3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark. In *ICCV*, 2025. 4

[34] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated Question Answering in 3D Scenes. In *ICLR*, 2023. 2

[35] Bruce L McNaughton, Francesco P Battaglia, Ole Jensen, Edvard I Moser, and May-Britt Moser. Path Integration and the Neural Basis of the'Cognitive Map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006. 2, 7

[36] Maurice Merleau-Ponty, Donald Landes, Taylor Carman, and Claude Lefort. *Phenomenology of Perception*. Routledge, 2013. 2

[37] Pierre-Emmanuel Michon and Michel Denis. When and Why are Visual Landmarks Used in Giving Directions? In *International Conference on Spatial Information Theory*, pages 292–305. Springer, 2001. 3

[38] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-Bench: A Comprehensive Benchmark and Toolkit for Evaluating Video-based Large Language Models. In *Computational Visual Media*, 2025. 2

[39] Nadine B. Sarter and David D. Woods. Situation Awareness: A Critical But Ill-Defined Phenomenon. *The International Journal of Aviation Psychology*, 1:45–57, 1991. 2

[40] Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models. In *ICLR*, 2025. 2, 5, 12, 13, 14

[41] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An Empirical Analysis on Spatial Reasoning Capabilities of Large Multimodal Models. In *EMNLP*, 2024. 4

[42] Daniel J Simons and Daniel T Levin. Change Blindness. *Trends in cognitive sciences*, 1(7):261–267, 1997. 3

[43] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*, 2026. 5, 16

[44] Matthias Stangl, Ingmar Kanitscheider, Martin Riemer, Ila Fiete, and Thomas Wolbers. Sources of Path Integration Error in Young and Aging Humans. *Nature Communications*, 11(1):2626, 2020. 7

[45] Barbara Tversky. Cognitive Maps, Cognitive Collages, and Spatial Mental Models. In *European Conference on Spatial Information Theory*, pages 14–24. Springer, 1993. 3

[46] Barbara Tversky. Spatial Cognition: Embodied and Situated. 2009. 2

[47] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI. In *CVPR*, 2024. 2

[48] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An Extreme Long Video Understanding Benchmark. In *ICCV*, 2025. 2

[49] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. In *NeurIPS*, 2024. 5, 12

[50] Thomas Wolbers and Mary Hegarty. What Determines Our Navigational Abilities? *Trends in Cognitive Sciences*, 14(3): 138–146, 2010. 2, 7

[51] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *NeurIPS*, 2025. 2

[52] 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. 3, 4

[53] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the Arrow of Time in Large Multimodal Models. In *NeurIPS*, 2025. 2

[54] 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 *CVPR*, 2025. 2

[55] 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 *CVPR*, 2025. 2, 3, 4, 12

[56] 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*, abs/2505.23764, 2025. 4

[57] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards Spatial Supersensing in Video. *arXiv preprint arXiv:2511.04670*, 2025. 3, 4

[58] 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. In *AAAI*, 2026. 3, 4

[59] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial Mental Modeling from Limited Views. In *NeurIPS*, 2025. 2, 3, 4

[60] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. In *ACL*, 2025. 12

[61] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. In *ICLR*, 2023. 5, 16

[62] Hongjie Zhang, Lu Dong, Yi Liu, Yifei Huang, Yali Wang, Limin Wang, and Yu Qiao. LvBench: A Benchmark for Long-form Video Understanding with Versatile Multi-modal Question Answering. In *IJCV*, 2025. 2

[63] Weichen Zhang, Zile Zhou, Xin Zeng, Xuchen Liu, Jianjie Fang, Chen Gao, Yong Li, Jinqiang Cui, Xinlei Chen, and Xiao-Ping Zhang. Open3D-VQA: A Benchmark for Comprehensive Spatial Reasoning with Multimodal Large Language Model in Open Space. *arXiv preprint arXiv:2503.11094*, 2025. 4

[64] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data. In *TMLR*, 2025. 5

[65] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: Measuring Expert-Level Multi-Discipline Video Understanding. In *CVPR*, 2025. 2, 5, 12

[66] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: Towards Spatiotemporal Awareness in Vision Language Models. In *ICCV*, 2025. 4

[67] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025. 5

[68] Zhu Ziyu, Ma Xiaojian, Chen Yixin, Deng Zhidong, Huang Siyuan, and Li Qing. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. In *ICCV*, 2023. 2

A. Experiment Setup

A.1. Model Configuration

A.2. Implementation Details for Model Inference

All MFMs are evaluated in a zero-shot setting across all tasks, consistent with prior work [40, 49, 55, 60, 65]. Whenever possible, we use the official code provided by each model for video preprocessing. Proprietary models and Qwen-series models are evaluated via their official API services[1]. In the evaluation prompt (§ B.2), models are instructed to return responses in a JSON-like format that includes both the selected multiple-choice option and the corresponding reasoning trace, enabling structured analysis and automatic parsing. When standard regular-expression–based parsing fails, we employ GPT-4o-mini to extract the multiple-choice answers; the extraction prompt is provided in § B.3.

For thinking-enabled models, the `thinking_level` parameter for Gemini 3 Flash, Gemini 2.5 Pro, and Gemini 2.5 Flash is set to the default `dynamic` mode[2], allowing the model to adapt its reasoning budget based on task complexity. For Gemini 3 Pro, `thinking_level` is set to `low`, as the `dynamic` setting consistently produces reasoning traces that exceed the output context window, leading to incomplete or `null` responses. For GPT-5.2, the thinking level is set to `medium`. For the Qwen3-VL series, the `thinking` variants are used instead of the `instruct` variants, with `thinking_level` set to `medium`.

[1] OpenAI, Gemini, and Qwen-series
[2] Gemini thinking mode

Table 4. **Model Configurations for Evaluation.** Unset values indicate that their default values are being used. Configurations are based on official model repositories where available. **Temp.**: temperature. ‡: GPT-5 mini only accepts a temperature value of 1.0. *: GPT-5.2 does not support adjusting the temperature or top-p parameters. †: The `thinking_level` parameter for Gemini 3 Pro was set to `low`. When configured to `dynamic`, Gemini 3 Pro consistently generated reasoning traces that exceeded the output context window, resulting in incomplete or `Null` responses.

Model	API Checkpoint / HF Checkpoint	Max New Tokens	Temp.	Top-P	Top-K	Sampling Rate (fps)	Thinking Level
Proprietary Multimodal Foundation Models							
Gemini 3 Pro†	gemini-3-pro-preview	16384	0.0	1.0	1.0	2	low†
Gemini 3 Flash	gemini-3-flash-preview	16384	0.0	1.0	1.0	2	dynamic
Gemini 2.5 Pro	gemini-2.5-pro	16384	0.0	1.0	1.0	2	dynamic
Gemini 2.5 Flash	gemini-2.5-flash	16384	0.0	1.0	1.0	2	dynamic
GPT-5.2*	gpt-5.2-2025-12-11	16384				2	medium
GPT-5 mini‡	gpt-5-mini-2025-08-07	16384	1.0	1.0		2	
Open-Source Multimodal Foundation Models							
Qwen3-VL 235B	qwen3-vl-235b-a22b-thinking	16384	0.0	1.0		2	medium
Qwen3-VL 32B	qwen3-vl-32b-thinking	16384	0.0	1.0		2	medium
Qwen3-VL 30B	qwen3-vl-30b-a3b-thinking	16384	0.0	1.0		2	medium
Qwen3-VL 8B	qwen3-vl-8b-thinking	16384	0.0	1.0		2	medium
Qwen2.5-VL 72B	qwen2.5-vl-72b-instruct	8192	0.0	1.0		2	
Qwen2.5-VL 32B	qwen2.5-vl-32b-instruct	8192	0.0	1.0		2	
Qwen2.5-VL 7B	qwen2.5-vl-32b-instruct	8192	0.0	1.0		2	
LLaVA-NeXT-Video 32B	LLaVA-NeXT-Video-32B-Qwen	32768	0.0	1.0		2	
LLaVA-Video 7B	LLaVA-Video-7B-Qwen2	32768	0.0	1.0		2	
LLaVA-Video 72B	LLaVA-Video-72B-Qwen2	32768	0.0	1.0		2	
LLaVA OneVision 72B	llava-onevision-qwen2-72b-ov	32768	0.0	1.0		2	
LLaVA OneVision 7B	llava-onevision-qwen2-7b-ov	32768	0.0	1.0		2	
InternVL3 38B	InternVL3-38B	32768	0.0	1.0		2	
InternVL3 14B	InternVL3-14B	32768	0.0	1.0		2	
InternVL3 8B	InternVL3-8B	32768	0.0	1.0		2	
InternVL2 40B	InternVL2-40B	8192	0.0	1.0		2	
InternVL2 8B	InternVL2-8B	32768	0.0	1.0		2	

B. Prompts

B.1. System Prompt

System Prompt

```
You are the person wearing the AR glasses.
 All videos are recorded from your
first-person perspective. Treat the
camera's movement as your own head and
 body movement. Only reason about what
 is visible or inferable from the
egocentric video. Do not assume any
external knowledge beyond what appears
 in the video.
```

B.2. Evaluation Prompt

We follow the evaluation prompt used in TOMATO [40].

Evaluation Prompt

```
You will be provided with a sequence of
    frames uniformly sampled from a video,
     the frames are provided in
    chronological order of the video.
    Analyze these frames and provide the
    answer to the question about the video
     content. Answer the multiple-choice
    question about the video content.

You must use these frames to answer the
    multiple-choice question; do not rely
    on any external knowledge or
    commonsense.

<question>
{question}
</question>

<options>
{index2ans}
</options>

PLEASE ANSWER THE QUESTION WITH ONLY THE
    OPTIONS PROVIDED. When answering,
    please follow the template provided:

"options": <your choice>
"thinking_trace": <your thinking trace>
```

B.3. Answer Extraction Prompt

We follow the answer extraction prompt used in TOMATO [40].

Answer Extraction Prompt

```
You are given a response, a list of
    multiple-choice options, and a
    index2answer mapping. You are required
     to extract the letter option from GPT
    .

<response>
{response}
</response>

<all_choices>
{all_choices}
</all_choices>

<index2answer>
{index2ans}
</index2answer>

Only output the single parsed letter from
    the response. No other texts are
    needed.

If you think no options can match the
    index2answer dictionary, randomly
    select one letter.

Your extracted letter is:
```

C. Baselines

We summarize the baseline configurations in Table 2. Additional details for the Blind LLM baseline are provided in § C.1, the Socratic model in § C.2, and the human evaluation in § C.3.

C.1. Blind LLM

Blind LLM refers a language-only model that does not receive any visual input, such as images or videos, during the inference stage. Instead, the model is provided solely with the textual component of the task, including the question, instructions, and answer options. This setting isolates the contribution of linguistic priors and textual reasoning, serving as a diagnostic baseline for identifying language-only shortcuts and estimating models' performance without perceptual information. By comparison with vision-enabled settings, this baseline quantifies the extent to which visual information contributes to task performance. We follow the same blind LLM prompt used in HourVideo [6].

Blind LLM Prompt

```
You are tasked with assisting in answering
     a few difficult questions about short
     egocentric videos. The goal is to
    establish a baseline for how many
    multiple-choice questions (MCQs) can
    be accurately answered without
    watching the videos. This may involve
    identifying poorly crafted distractor
    options or leveraging general
    knowledge and logical reasoning when
    the questions themselves are
    straightforward.

You are STRICTLY expected to choose the
    correct MCQ answer based on your best
    judgment and provide a one-line reason
     for your selection.

<question>
{question}
</question>

<options>
{index2ans}
</options>

DO NOT GENERATE ANSWER SUCH AS 'NOT
    POSSIBLE TO DETERMINE.'
```

C.2. Socratic Model

Socratic model [61] is a framework in which visual perception and textual reasoning are decoupled through a two-stage pipeline. In the first stage, for a video question-answering task, a video captioning model converts the visual input into a language-only caption. In the second stage, a language model receives the generated caption together with the question and answer options, and performs reasoning entirely in the textual domain. Under this setting, the reasoning model does not access raw visual inputs; instead, all perceptual information is compressed through the caption. This framework evaluates how informative the visual input remains when mediated solely through textual descriptions.

In our baseline evaluation, we use GPT-5.2 [43] for both video captioning and caption-based question answering. Videos are sampled at 2 fps, and the sampled frames are stitched into 4×4 image grids, with each grid containing 16 frames. For each video, multiple 4×4 grids are provided to the captioning model to generate a textual description. We follow the same caption generation and evaluation prompts use in HourVideo [6]. Caption generation prompt is provided in §C.2.1. Evaluation prompt using video caption is provided in §C.2.2. Example input for caption generation is provided in §C.2.3.

C.2.1. Caption Generation Prompt

Caption Generation Prompt – part 1

```
MAIN INSTRUCTIONS:

Your task is to analyze video frames
    extracted uniformly from a short ego-
    centric video for a detailed video
    understanding exercise. I will provide
     a sequence of images sampled at 2
    frames per second (2 fps) from this
    video. Examine the video frames
    closely and generate a comprehensive
    caption by strictly following the
    steps below:

Step 1: **Scene Context**
Observe the frames. What is the primary
    setting and activity in the video?

Step 2: **Motion Description**
Identify and describe any significant
    motion or actions taking place across
    the frames.

Step 3: **Spatial Relationship Analysis**
Examine and report the spatial
    relationships between key objects or
    entities. Describe the positioning and
     orientation of each element relative
    to others.

Step 4: **Detailed Object Analysis**
List the key objects and entities.
    Describe visible attributes such as
    color, shape, texture, and other
    notable features with precision (e.g.,
     materials, signage text, tool types,
    object parts).

Step 5: **Temporal Relationship Context**
Describe any observable temporal
    progression or changes across the
    sequence (e.g., before/after changes,
    object movement, state changes, action
     sequences). If no meaningful change
    is visible, state that explicitly.

Step 6: **Additional Detail-Oriented
    Observations**
Add any other concrete, detail-oriented
    observations that could help answer
    fine-grained questions later (e.g.,
    small objects, relative distances,
    occlusions, left/right placement,
    openings/closures, item locations),
    but do not speculate beyond what is
    visible.

Step 7: **Summary**
Provide a concise yet comprehensive
    summary capturing the key elements
    from Steps 1-6.
```

Caption Generation Prompt – part 2

```
GUIDELINES:

1. Return your results in a paragraph
    format with the following fields:
  - Scene Context
  - Motion Description
  - Spatial Relationship Analysis
  - Detailed Object Analysis
  - Temporal Relationship Context
  - Additional Details
  - Summary

2. The total length of the output must not
     exceed 200 words.
```

C.2.2. Socratic Model Evaluation Prompt

Socratic Model Evaluation Prompt

```
You will be provided with a textual
    caption that describes the content of
    a video. The caption is derived from
    the video and reflects its observable
    visual and spatial information.
    Analyze the caption and answer the
    multiple-choice question about the
    video content.

You must use only the information
    contained in the provided caption to
    answer the question; do not rely on
    any external knowledge, assumptions,
    or commonsense beyond what is
    explicitly stated in the caption.

<video_caption>
{video_caption}
</video_caption>

<question>
{question}
</question>

<options>
{index2ans}
</options>

PLEASE ANSWER THE QUESTION USING ONLY THE
    OPTIONS PROVIDED. When answering,
    strictly follow the template below:

"options": <your choice>
"thinking_trace": <your reasoning based
    solely on the caption>
```

C.2.3. Example Inputs for the Socratic Model Baseline

We provide the example input for caption generation in Figure 8.

C.3. Human Evaluation

To establish a human performance upper bound on the benchmark, two graduate students are recruited to complete the full set of evaluation tasks. Each annotator answers all questions independently, with unlimited time and full access to the corresponding videos. Annotators are instructed to rely solely on the visual information provided, without external tools or discussion. A breakdown of individual annotator performance is reported in Table 11. Human evaluation interface in shown in Figure 9.

D. Video Filming Protocol and Meta Information Annotation

D.1. Video Filming Protocol

We define a structured recording protocol to ensure consistent coverage of observer-centric spatial reasoning primitives while maintaining controllable trajectory complexity. Each video is associated with a predefined movement pattern, a set of spatial queries, and deterministic ground-truth answers derived from the recording plan.

We divide the recording protocol into four trajectory categories: (1) **In-place orientation** (§D.1.1), in which the camera wearer remains at a fixed spatial location and only rotates their viewpoint; (2) **Manhattan-style piecewise linear trajectories** (§D.1.2), in which the camera wearer follows a predefined path with two turns; (3) **Simple geometric trajectories** (§D.1.3), in which the camera wearer moves along canonical geometric paths; and (4) **Extra video collections** (§D.1.4), which include additional recordings designed to support the **Spatial Memory** and **Spatial Affordance** tasks.

Figure 8. **Example Input to the Socratic Baseline.** Each 4×4 grid contains 16 frames sampled at 2 fps, with frame indices shown in the bottom-left corner of each frame.

Table 5. **Human Baseline Performance on SAW-BENCH.**

Model	All	Self-Localization	Relative Direction	Route Shape	Reverse Route Plan	Spatial Memory	Spatial Affordance
Annotator 1	91.07	98.50	86.69	98.53	91.70	94.00	76.54
Annotator 2	92.03	89.50	92.09	96.70	94.32	83.00	81.48
Average	91.55	94.00	89.39	97.62	93.01	88.50	79.01

D.1.1. In-Place Orientation

D.1.2. Manhattan-Style Piecewise Linear

D.1.3. Simple Shape Trajectories

D.1.4. Extra Video Collections

In addition to the predefined observer trajectories described in §D.1.1, §D.1.2, and §D.1.3, we instruct video collectors to record additional videos in each scene to support the **Spatial Memory** and **Spatial Affordance** tasks. For **Spatial Memory**, video collectors identify objects that have been moved or displaced within the scene over the course of the trajectory. For **Spatial Affordance**, video collectors assess whether a queried action is physically feasible given the scene's spatial and physical constraints.

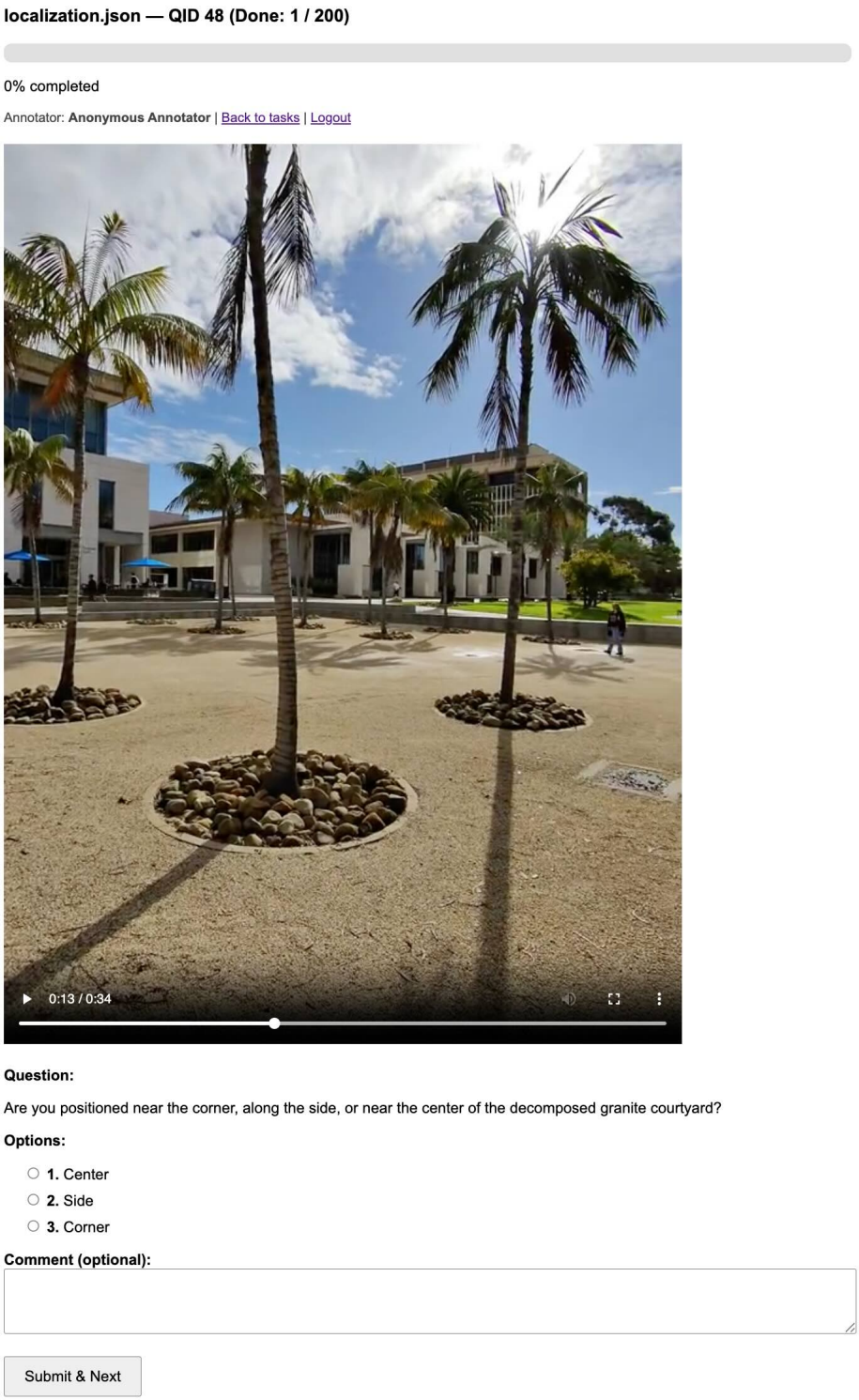

Figure 9. **Human Evaluation Interface.** Each annotator answers all questions independently, with unlimited time and full access to the corresponding videos. Annotators are instructed to rely solely on the visual information provided, without external tools or discussion.

D.2. Meta Information Annotation

Ground-truth answers are derived directly from the predefined trajectory and viewpoint sequence, independent of vi-

Table 6. **Video Filming Protocol for In-Place Orientation.** Camera wearers remain at a fixed spatial location (*i.e.* corner, side, or center of a scene) and only changes their viewing orientation, without introducing translational motion. Whenever possible, we include instances from each spatial location to ensure coverage across different positional contexts. For each video, we asked one **Self-Localization** and one **Route Shape** question.

Spatial Location	Video ID	Q1	A1	Q2	A2
Corner 1	0	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Corner	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Corner 2	1	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Corner	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Corner 3	2	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Corner	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Corner 4	3	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Corner	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Side 1	4	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Side	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Side 2	5	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Side	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Side 3	6	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Side	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Side 4	7	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Side	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation
Center	8	**Self-Localization**: Are you positioned near the corner, along the side, or near the center of the ⟨Scene Name⟩?	Center	**Route Shape**: What's the shape of my moving trajectory?	In-place orientation

sual appearance. Screenshot of annotation interface is provided in Figure 9.

Self-Localization. Ground-truth answers are deterministically defined by the filming location of the observer within the scene (*i.e.* corner, side, or center). Since observer positions are predefined during data collection, labels of **Self-Localization** do not require human annotation.

Relative Direction. Ground-truth answers are annotated by human annotators based on the relative spatial relationship between the observer's starting and ending viewpoints along the trajectory. Each question is independently annotated by two annotators, and inter-annotator agreement is reported to assess annotation consistency. We annotate only trajectory-level meta information; in this way, all videos that share the same trajectory share the same relative-direction labels. Inter-annotator agreement score is reported in Table 9. Any annotation disagreements are resolved through discussion at the end of the annotation process.

Route Shape. Ground-truth answers are deterministically specified by the predefined movement trajectory (*e.g.* straight, L-shape, U-shape, circular). Similar to **Self-Localization**, these labels are directly derived from the filming protocol and do not depend on visual interpretation.

Reverse Route Plan. Ground-truth answers for reverse route planning are defined analogously to **Relative Direction**, describing the sequence of actions required to return from the final viewpoint to the starting viewpoint. These answers are annotated by human annotators to account for egocentric action interpretation, following the same annotation protocol as relative direction. Similarly, we annotate only trajectory-level meta information. Inter-annotator agreement score is reported in Table 9. Any annotation disagreements are resolved through discussion at the end of the annotation process.

Spatial Memory. Ground-truth answers are specified during the filming process by intentionally moving or displacing selected objects along the trajectory. Since object state changes are controlled and recorded at capture time, no additional human annotation is required.

Spatial Affordance. Ground-truth answers are likewise determined during filming by designing queries whose feasibility is unambiguous given the scene's physical layout and constraint. Similar to **Spatial Memory**, no additional human annotation is required.

CVPR #48

Table 7. **Video Filming Protocol for Manhattan-Style Piecewise Linear.** Observers' moving trajectories are visualized using egocentric action arrows. For each video, we asked two **Relative Direction** questions, one **Reverse Route Plan** question, and one **Route Shape** question.

Trajectory	Video ID	Q1	A1	Q2	A2	Q3	A4
↑ ← ↑	18	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front left / back right⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, then go straight. Turn **left**, go straight, then turn **right** and continue straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
← ↓ ↑	19	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨ left / left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **right** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
↑ → ↑	20	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨ front right / back left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
→ ↑ ↓	21	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨ right / right⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **left**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
← ↑ ←	22	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front left / back left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
→ ↑ ←	23	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front / right⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **left**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
← ↓ ←	24	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨ back left / back right⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **left**, go straight, then turn **right** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
← ↓ →	25	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨back / left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **right** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
← ↑ →	26	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨back / left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **right** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
→ ↑ →	27	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front right / back right⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **left**, go straight, then turn **right** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
→ ↓ ←	28	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨back / left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **left**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
→ ↓ →	29	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨back right / back left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
↑ ↓ ←	30	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨ left / right⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **left**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
↓ ← ↓	31	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨ back left / back left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **left** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
↓ → ↑	32	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨ right / left⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **right**, go straight, then turn **right** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear
↓ → ↓	33	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨back right / back right⟩	**Reverse Route Plan**: From my viewpoint at the end of the video, how can I go back to my starting point?	Turn around, go straight. Turn **left**, go straight, then turn **right** and continue go straight.	**Route Shape**: What's the shape of my moving trajectory?	Manhattan-style piecewise linear

Table 8. **Video Filming Protocol for Other Simple Trajectories.** For each video, we asked two **Relative Direction** questions and one **Route Shape** question.

Trajectory	Video ID	Q1	A1	Q2	A2
Straight	34	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨left / right⟩	**Route Shape**: What's the shape of my moving trajectory?	Straight
Straight w/ turning head frequently	35	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front / back⟩	**Route Shape**: What's the shape of my moving trajectory?	Straight
$\rightarrow$ $\uparrow$ (L-Shape)	36	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front right / back right⟩	**Route Shape**: What's the shape of my moving trajectory?	L-Shape
$\leftarrow$ $\uparrow$ (L-Shape)	37	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front left / back left⟩	**Route Shape**: What's the shape of my moving trajectory?	L-Shape
$\subset$ (U-Shape)	38	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front / right⟩	**Route Shape**: What's the shape of my moving trajectory?	U-Shape
$\supset$ (U-Shape)	39	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨back / right⟩	**Route Shape**: What's the shape of my moving trajectory?	U-Shape
$\cup$ (U-Shape)	40	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨right / left⟩	**Route Shape**: What's the shape of my moving trajectory?	U-Shape
$\cap$ (U-Shape)	41	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨right / right⟩	**Route Shape**: What's the shape of my moving trajectory?	U-Shape
Circle (Clockwise)	42	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨same location / same location⟩	**Route Shape**: What's the shape of my moving trajectory?	Circular
Circle (Counter-clockwise)	43	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨same location / same location⟩	**Route Shape**: What's the shape of my moving trajectory?	Circular
Zigzag	44	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨front / back⟩	**Route Shape**: What's the shape of my moving trajectory?	Zigzag
Square	45	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨same location / same location⟩	**Route Shape**: What's the shape of my moving trajectory?	Square
Triangle	46	**Relative Direction**: From my viewpoint at the ⟨beginning / end⟩ of this video, where am I located at the ⟨end / beginning⟩ of the video?	⟨same location / same location⟩	**Route Shape**: What's the shape of my moving trajectory?	Triangle

Table 9. **Inter-Annotator Agreement Measured by Cohen's κ.**

Task	Cohen's κ
Relative Direction	0.987
Reverse Route Plan	0.952

E. Data Analysis

E.1. Video Duration Distribution

We provide video duration distribution of SAW-BENCH in Figure 10.

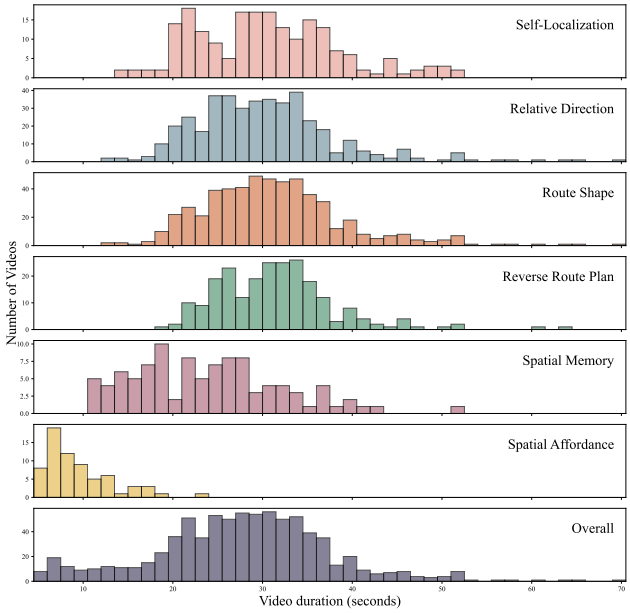

Figure 10. **Video duration distribution of SAW-BENCH.**

E.2. Question Scene Distribution

We provide question scene distribution of SAW-BENCH in Figure 11.

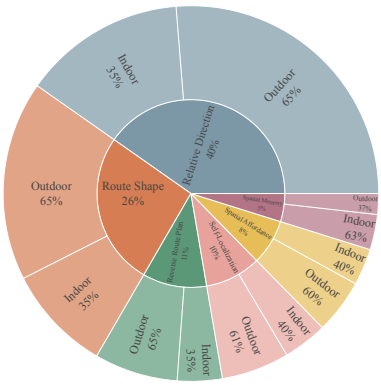

Figure 11. **Question scene distribution of SAW-BENCH.**

E.3. Key Statistics

We summarize key statistics of SAW-BENCH in Table 10.

Table 10. **Key statistics of SAW-BENCH.**

Statistics	Value
Unique Videos	786
Indoor	317 (40.33%)
Outdoor	469 (59.67%)
Number of Scenes	15
Number of Videos per scene (`avg`)	42
Number of Videos in Unassigned Scene	155
Total Questions	2,071

F. Sensitivity Analysis

F.1. Sensitivity to Number of Input Frames

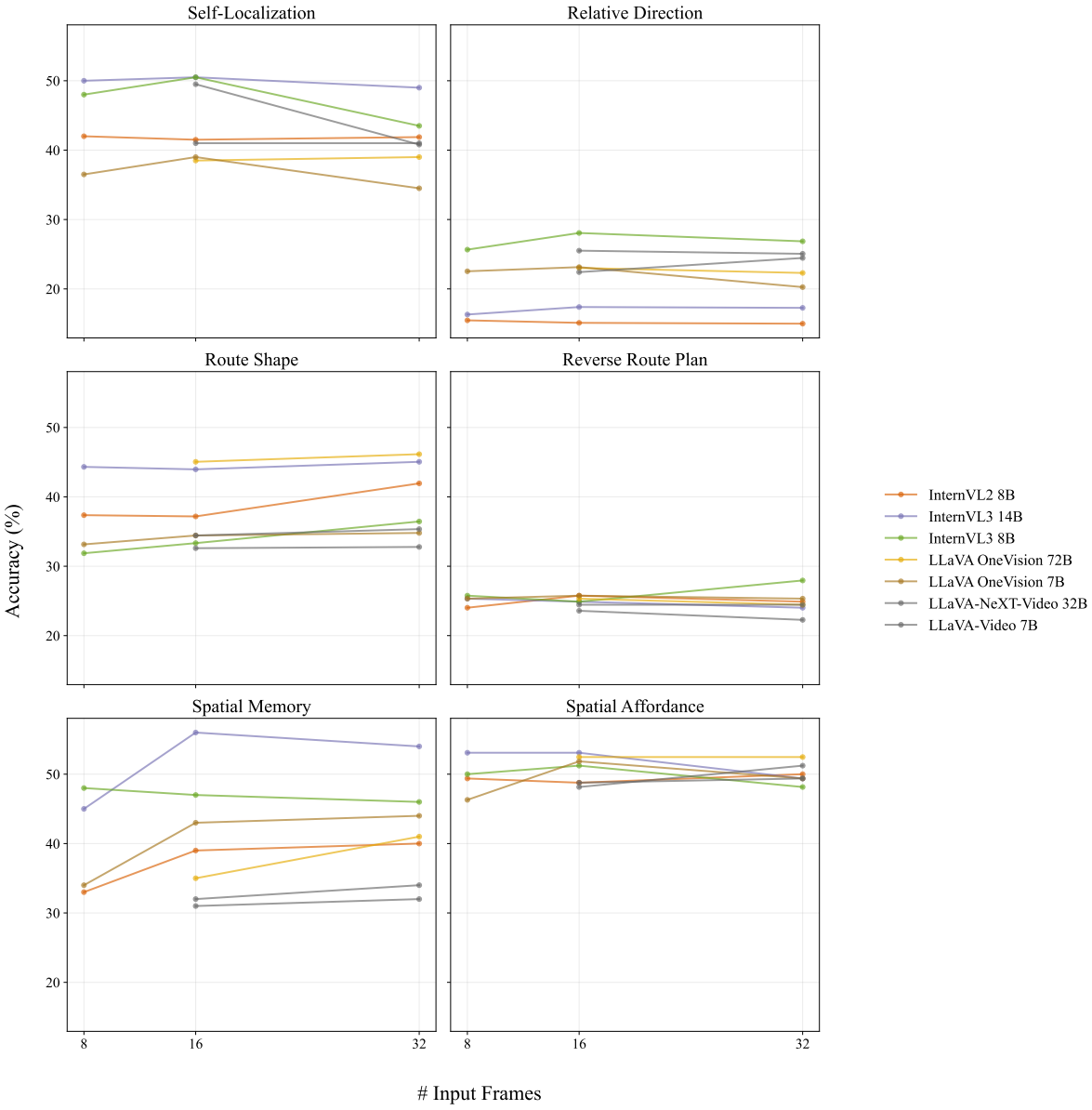

Figure 12. **Sensitivity to the Number of Input Frames Across Reasoning Types.** Each subplot reports accuracy as a function of the number of input frames (8, 16, 32) for a fixed reasoning type. Lines correspond to different open-source MFMs evaluated with varying temporal context lengths. Tasks that require longer-horizon spatial integration, such as **Spatial Memory** and **Route Shape**, generally benefit from increased temporal context, whereas short-horizon tasks (*e.g.* **Self-Localization** and **Relative Direction** exhibit earlier saturation and more limited improvements beyond 16 frames. Nevertheless, additional frames yield only marginal gains across all tasks.

F.2. Sensitivity to Frame Sampling Rate (FPS)

Table 11. **Sensitivity to Frame Sampling Rate (FPS) Across Reasoning Types.** We evaluate the top-performing model, Gemini 3 Flash, on SAW-BENCH by varying the number of input frames through different temporal sampling rates (FPS $\in \{1, 2, 3\}$), and report performance across reasoning tasks. All other configurations are held constant, as detailed in Table 4. Overall, increasing the frame sampling rate does not substantially improve situated spatial reasoning. Notably, performance on **Spatial Memory** decreases at FPS = 3, potentially due to redundant or temporally condensed visual information, while **Reverse Route Plan** performs worst at FPS = 1, suggesting a greater reliance on temporal visual context for this reasoning type.

Model	FPS	All	Self-Localization	Relative Direction	Route Shape	Reverse Route Plan	Spatial Memory	Spatial Affordance
Gemini 3 Flash	1	53.55	45.00	39.93	69.05	55.90	64.00	72.22
Gemini 3 Flash	2	53.89	48.50	41.13	64.84	61.57	66.00	70.99
Gemini 3 Flash	3	51.57	42.50	40.29	60.99	63.76	53.00	70.99