LOST IN THE HALLWAY: BENCHMARKING PERSISTENT SPATIAL MEMORY ACROSS REVISITS

Anonymous authors
Paper under double-blind review

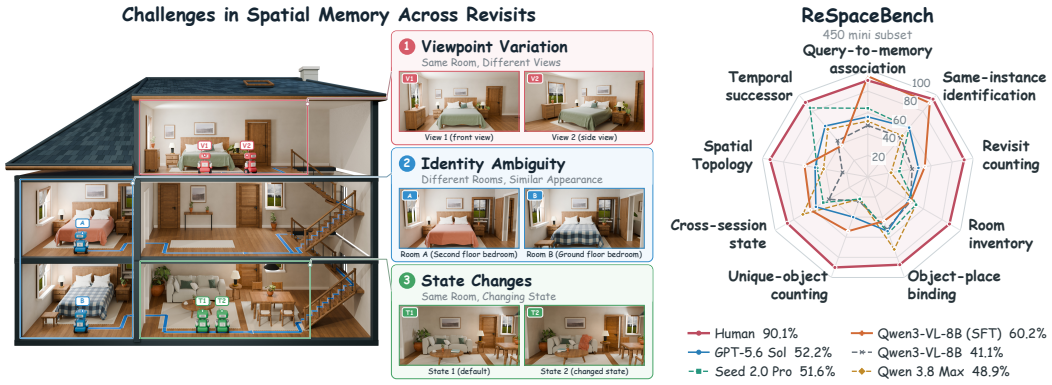

Figure 1: **Challenges in spatial memory across revisits. Left:** Three challenges arise when revisiting a previously explored environment: (1) *Viewpoint Variation*: recognizing the same room from different viewpoints; (2) *Identity Ambiguity*: distinguishing between visually similar but distinct rooms; and (3) *State Changes*: tracking changes within the same room across repeated visits. **Right:** Human and MLLM accuracy across all 9 tasks of RESPACEBENCH-*mini* (450 questions). Even the best-performing model remains substantially below human performance.

ABSTRACT

Persistent spatial understanding is essential for embodied agents to operate reliably in environments they repeatedly encounter. It requires recognizing revisited places despite viewpoint changes, distinguishing them from similar-looking rooms, and updating memories as environments change. Yet existing evaluations often leave physical place identity across separated visits implicit. We introduce RESPACEBENCH, a benchmark comprising 3,244 questions across 590 synthetic and real-home scenes, organized into 9 tasks spanning room association, persistent state tracking, and downstream reasoning. Controlled walkthroughs define visits by physical room entry and deliberately include same-type room distractors and cross-session object changes. Our evaluation of 32 models and 3 agentic video-understanding frameworks reveals substantial limitations: the strongest evaluated zero-shot model achieves 50.5% overall accuracy and 52.2% on a 450-question subset, compared with 90.1% for human annotators. The evaluated agentic frameworks offer no overall improvement over direct inference with the same base model. Analysis of agent execution traces reveals evidence omissions and incorrect cross-visit room associations as major failure modes, suggesting that iterative evidence seeking alone does not ensure consistent spatial memory. Further analyses expose a gap between recognizing familiar rooms and identifying the correct visited room, while annotation-guided sampling highlights the importance of preserving intervening visits and transitions. We also provide RS-30K, a separate training dataset from synthetic sources with no template, scene, route, or video overlap with the benchmark. Fine-tuning Qwen3-VL-8B raises overall accuracy from 35.5% to 60.4%, with the largest gains in association and improvements extending to real-life scenes. RESPACEBENCH and RS-30K provide resources for diagnosing and improving persistent spatial memory across revisits.

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.

1 INTRODUCTION

Embodied agents repeatedly encounter the same physical environments during everyday activities (Hawes et al., 2017; Kunze et al., 2018). A robot or wearable assistant may return to a room from a different viewpoint, after visiting other spaces, or after its contents have changed. To build on past experience during these revisits, it must associate new observations with the correct physical place and update its spatial memory accordingly (Schmid et al., 2024; Labbé & Michaud, 2019; Hughes et al., 2022). Yet *viewpoint variation* can make even a familiar room appear substantially different across visits, complicating place recognition (Garg et al., 2021; Lowry et al., 2016; Zaffar et al., 2021). Conversely, different rooms may look alike, creating *identity ambiguity* about the correct identity of the room being revisited (Lowry et al., 2016; Zhang et al., 2021). Over time, the agent must also track *state changes* in the room's contents and update the corresponding memory (Kurenkov et al., 2023; Narayana et al., 2020). Revisits therefore test whether systems can preserve place identity while tracking changes across temporal observations (Fig. 1).

Recent benchmarks have expanded the evaluation of multimodal large language models (MLLMs) from single-image 3D relations to video-based spatial reasoning and long-horizon memory (Ma et al., 2025; Yang et al., 2025a; Lin et al., 2025; Lang et al., 2026). Yet models still struggle with global cognitive maps, cross-view correspondence, and coherent spatial memory (Yang et al., 2025a; Yeh et al., 2026; Liao et al., 2026; Lin et al., 2025). Cross-visit place association, a task well studied in robotics through loop closure and visual place recognition (Cadena et al., 2016; Lowry et al., 2016), remains largely implicit in the spatial understanding task evaluations. Recent benchmarks, UCS-Bench (Wang et al., 2026a) and SpaMEM (Liao et al., 2026) evaluate persistent spatial reasoning and memory under scene changes, while LongSpace-Bench (Lang et al., 2026) covers scene consistency and long-horizon recall; however, none establishes a dedicated protocol for room re-identification across controlled revisits. OVO-S-Bench (Li et al., 2026) includes revisit detection and counting, but defines visits by whether a location enters the field of view, allowing a room briefly glimpsed during a camera turn to count without physical entry. Its construction also does not systematically control same-type rooms as competing identities. Thus, associating observations with specific physical rooms across controlled revisits remains under-explored.

We present RESPACEBENCH, a benchmark for persistent spatial understanding across revisits. Its 9 tasks spans three levels: *association*, recognizing previously visited rooms and matching them to earlier observations; *persistent state*, maintaining and updating room and object memories; and *downstream reasoning*, recovering temporal order and spatial topology. Walkthroughs follow controlled routes, with visits defined by physical room entry rather than visual observation. We select homes with spaces of the same type to serve as identity distractors, testing whether models recognize a specific place rather than merely remembering its semantic category.

We construct RESPACEBENCH from synthetic homes in 3D-FRONT (Fu et al., 2021) and ProcTHOR (Deitke et al., 2022), real homes scanned with FARO laser scanners (FARO, 2026), and egocentric videos recorded with a PICO headset (PICO, 2026). Using annotated scene metadata and constructed navigation meshes, we use our designed question templates to generate task-conditioned routes: each template specifies both the question and the navigation requirements needed to answer it. The resulting route trajectories include egocentric videos with repeated room visits. We generate answers from metadata, then apply rule-based checks with human verification to remove question–answer pairs that are inconsistent with the video. Combining synthetic environments and real homes, RESPACEBENCH comprises 590 scenes, 3,037 rooms, 353.4 hrs, and 3,244 questions.

Our evaluation of 32 models and three agentic video-understanding frameworks (Section 4) reveals substantial limitations in persistent spatial memory. Proprietary models remains the strongest zero-shot models in our evaluation, while specialized spatial and video models do not consistently outperform open-source general-purpose MLLMs. The evaluated agentic frameworks offer no overall gains over direct inference. On a shared 450-question subset, humans achieve 90.1% accuracy compared with 52.2% for the strongest evaluated zero-shot model GPT-5.6 Sol. Further analysis of the six proprietary models on query-to-memory association shows that wrong-room selections account for 66.5–88.3% of their errors, exposing a gap between recognizing a room as familiar and identifying the correct visited room. To support model development, we also provide 30k training questions from ProcTHOR training scenes with no scene, route, or video overlap with the bench-

Table 1: Comparison with related spatial understanding and memory benchmarks. Symbols reflect explicit evaluation targets: ✓ indicates explicit coverage; ✗ indicates a criterion not explicitly targeted; △ indicates related coverage that does not clearly establish the criterion. *Controlled revisits* refers to planned room re-entries along multi-room trajectories. *Same-type room distractors* refers to distinct rooms of the same category deliberately included as competing place identities.

Benchmark	**Input**	**Evaluation coverage**			**Revisit-focused design**	
		Spatial memory	State changes	Revisit recognition	Controlled revisits	Same-type room distractors
VSI-Bench (Yang et al., 2025a)	Video	✓	✗	✗	✗	✗
VSI-SUPER (Cambrian-S) (Yang et al., 2026c)	Video	✓	✗	✗	✗	✗
MMSI-Video-Bench (Lin et al., 2025)	Video	✓	✓	✗	✗	✗
SpaMEM (Liao et al., 2026)	Img. Sequence	✓	✓	✗	✗	✗
UCS-Bench (Wang et al., 2026a)	Video	✓	✓	✗	✗	✗
LongSpace-Bench (Lang et al., 2026)	Video	✓	✓	△	✗	✗
OVO-S-Bench (Li et al., 2026)	Video	✓	✓	✓	✗	✗
RESPACEBENCH (ours)	Video	✓	✓	✓	✓	✓

mark. Fine-tuning Qwen3-VL-8B on this dataset raises overall accuracy from 35.5% to 60.4%, with the largest gains in association and smaller, less consistent gains in state tracking and reasoning.

Our contributions are as follows.

- **A benchmark for persistent spatial memory.** We introduce RESPACEBENCH, comprising 9 task families of 3 levels across synthetic and real homes. Controlled revisits, same-type room distractors, and cross-session state changes support explicit evaluation of room association, persistent state tracking, and downstream spatial reasoning.
- **A systematic evaluation and error analysis.** We evaluate 32 models and 3 agentic video-understanding frameworks, with a human performance reference on a shared subset. Detailed analyses of room-association errors, agentic failures, and evidence coverage identify recurring limitations in maintaining and using spatial memory across revisits.
- **A training dataset.** We release 30k training samples generated from ProcTHOR scenes, with no template, scene, route or video overlap with the benchmark. A targeted fine-tuning study demonstrates their utility for developing models with stronger persistent spatial memory.

2 RELATED WORK

Long-context visual retrieval. Long-context evaluations examine whether models can retrieve and combine sparse visual evidence from extended multimodal inputs. Visual Haystacks tests visual needle retrieval and cross-image reasoning, revealing sensitivity to evidence position (Wu et al., 2024). MMLongBench broadens evaluation to diverse interleaved vision–language tasks at controlled context lengths (Wang et al., 2025b), while Video-LevelGauge systematically probes positional bias in video contexts (Xia et al., 2025). Recently, agentic video-understanding frameworks such as AVP, DVD, and VideoAgent use tools to retrieve and inspect video evidence (Wang et al., 2026c; Zhang et al., 2025b; Wang et al., 2024). Retrieving relevant observations, however, does not establish whether they refer to the same physical room across separate visits. RESPACEBENCH explicitly evaluates this cross-visit room association over long visual contexts.

Spatial intelligence and long-horizon memory. Recent benchmarks address complementary aspects of spatial intelligence and long-horizon memory. SpaMEM studies spatial belief updates and object-identity continuity under action-conditioned scene changes in ProcTHOR (Liao et al., 2026), while UCS-Bench evaluates continual spatial reasoning relative to users' changing locations in egocentric streams (Wang et al., 2026a). EgoMemReason probes entity, event, and behavioral memory over week-long videos (Wang et al., 2026b), and 3DMem-Bench evaluates memory-dependent question answering and embodied tasks in dynamic 3D environments (Hu et al., 2025). MMSI-Video-Bench includes cross-video spatial reasoning (Lin et al., 2025). In generative world modeling, MemoBench evaluates whether video generators recover objects in their updated states after disappearance and reappearance (Chen et al., 2026). More closely related to our setting, LongSpace-Bench

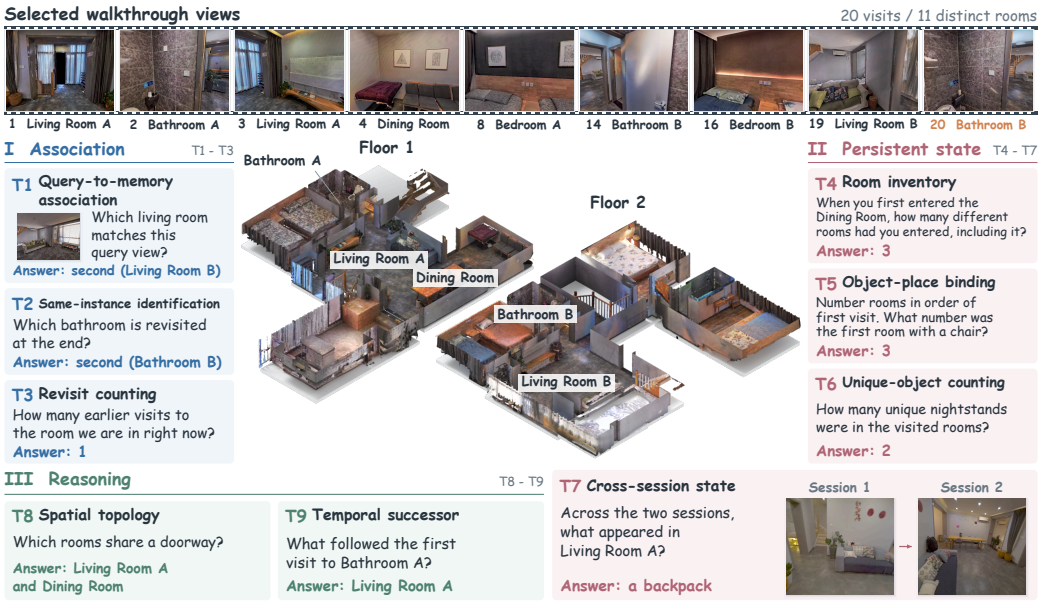

Figure 2: **Benchmark taxonomy:** RESPACEBENCH comprises 9 tasks across association (T1–T3), persistent state (T4–T7), and reasoning (T8–T9). Representative questions and answers illustrate these tasks through room revisits, same-type room distractors, and cross-session state changes.

includes scene consistency, state change, and route recall in long room-tour videos (Lang et al., 2026). OVO-S-Bench evaluates scene revisit detection and counting under a causal, prefix-only protocol, with visits in its counting task defined by a location entering the field of view (Li et al., 2026). RESPACEBENCH defines visits by physical room entry and deliberately includes distinct same-type rooms as competing identities along controlled revisit trajectories.

Persistent association and structured spatial memory. Persistent association underlies place recognition and loop closure in SLAM (Cadena et al., 2016). REMIND uses appearance, spatial context, and joint assignment for indoor object re-identification (Diaz-Pereda et al., 2026), while R4DSG indexes egocentric memory by time, place, persistent objects, and anchor-relative changes (Ma et al., 2026). PBD-AG separates a robot-verified structural baseline from revisable dynamic events (Bao et al., 2026), and GESTO uses event context within a persistent 4D scene graph to refine object associations (Bartoli et al., 2026). Complementing these methods, RESPACEBENCH evaluates whether MLLMs can associate observations with persistent room identities and use visit histories and state information to answer questions across revisits.

3 RESPACEBENCH

RESPACEBENCH evaluates persistent spatial memory through egocentric walkthrough that return to previously visited places. Its tasks span three levels (Fig. 2): *association*, recognizing a revisited room and matching it to an earlier observation; *persistent state*, maintaining room inventories, object–place bindings, and object states across visits; and *reasoning*, inferring temporal order and spatial connectivity from this history. The benchmark evaluates these capabilities using rendered walkthroughs of synthetic and scanned homes.

Task overview. For *association*, T1 (*query-to-memory association*) matches a query room to a visited room or labels it as never entered; T2 (*same-instance identification*) identifies which previously visited same semantic type room is being revisited; and T3 (*revisit counting*) counts prior entries into a specified room. For *persistent state*, T4 (*room inventory*) counts distinct visited rooms, overall or by semantic type, or determines a room's position in the order of first visits; T5 (*object–place binding*) reports first-visit of the first room containing an object of a queried type; T6 (*unique-object counting*) counts distinct objects of a queried type across revisits; and T7 (*cross-session state*) tracks

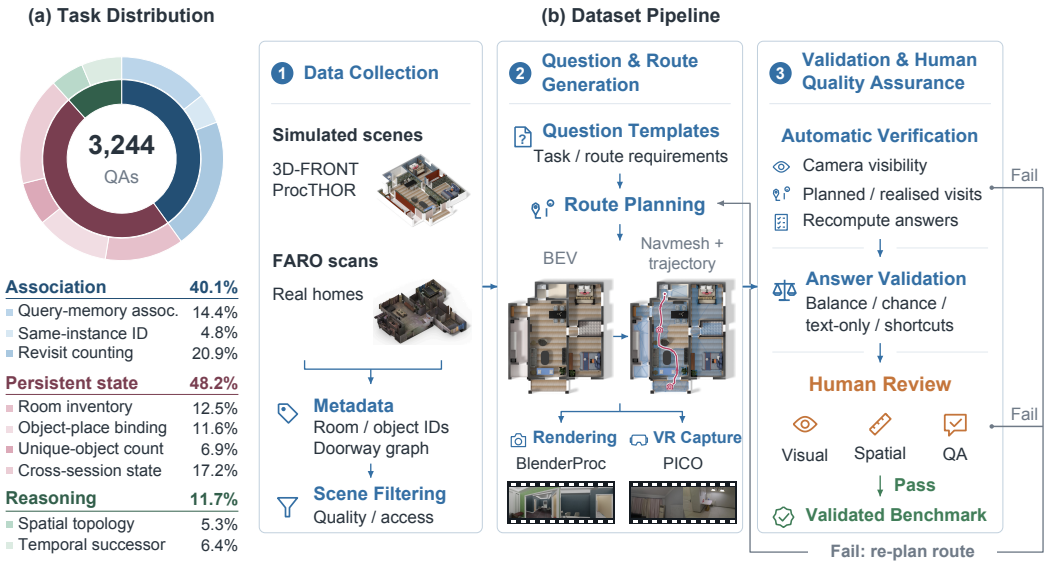

Figure 3: Benchmark statistics and construction pipeline. **(a)** Question distribution across the 3 task levels and 9 task families. **(b)** The 3-stage construction pipeline for synthetic and real homes, with automated checks and human quality control. The camera path is illustrative.

object locations and changes across sessions in the same home. For *reasoning*, T8 (*spatial topology*) identifies room pairs directly connected by a doorway, and T9 (*temporal successor*) identifies the room type entered immediately after a specified visit.

3.1 Benchmark Construction Pipeline

Fig. 3(b) shows the 3 stages of benchmark construction. Scene metadata and camera trajectories connects each question to the rooms and objects observed in its video, providing a shared and consistent basis for both question generation and answer verification.

Scene collection, filtering, and metadata. To ensure that the benchmark poses meaningful spatial-memory challenges, we select large homes with diverse room layouts and multiple rooms of the same semantic type. We source synthetic household environments from 3D-FRONT (Fu et al., 2021), which provides professionally designed furnished apartments, and ProcTHOR (Deitke et al., 2022), which contains procedurally generated homes. To complement these synthetic scenes, we scan real-world residential homes using a FARO laser scanner and record egocentric activities with a PICO headset. For synthetic scenes, we extract room geometry and semantic types, persistent room and object identities, object categories and locations, and doorway connectivity from the source metadata. For real homes, we manually annotate the corresponding information in the registered 3D scenes and align the annotations with the recorded egocentric trajectories.

Route planning and egocentric video generation. We plan walkthrough routes using room layout to balance room coverage and controlled revisits. Task-specific configurations include visits to visually similar rooms of the same semantic type, episodes spanning single and multiple homes, and repeated sessions in the same home with controlled object changes. For both synthetic and scanned homes, we construct navigation meshes using Habitat Simulator (Savva et al., 2019). Then, we adapt the navmesh-based navigation setup of Pred-EQA (Yuan et al., 2026) to generate traversable paths that follow the planned room-entry sequences. We convert each path into a continuous camera trajectory and render egocentric videos using BlenderProc (Denninger et al., 2023), then concatenating videos across multiple homes when needed. Rendered walkthroughs use nominal walking speeds of 0.6–0.8 m/s. At each room entry, the camera pauses for a 270° look-around lasting 6–9 seconds, providing sustained observations of the room. Hallways and stairwells are treated as transit spaces and are not counted as room visits. Each video is paired with camera poses, timestamps, room-entry events, and object-state records to support question generation, answer verification.

Question generation from templates. We instantiate task-specific question templates using scene metadata, ordered room-entry sequences, and object-state records. The templates cover room association, room and object inventories, object to place binding, state changes, temporal order, and room connectivity. Rooms are referred to by semantic type when the type is unique within a scene. Otherwise, we disambiguate them by order of first visit or unique visual cues (e.g., *the second bedroom visited* or *the bedroom with wooden floor*). Ground-truth answers are deterministically derived from the corresponding scene annotations and realized trajectory. For multiple-choice questions, distractors are included to reflect plausible spatial-memory errors, such as confusing distinct rooms of the same type or inferring direct doorway connectivity from consecutive visits.

Automatic and Human Quality Control. We verify each walkthrough by comparing the realized room-entry sequence with the planned route, rejecting trajectories with missing or unintended room entries. Ground-truth answers are then recomputed from the verified entry sequence and object-state records, and questions involving ambiguous visits or insufficient visual evidence are discarded. We further combine task-specific automated checks with manual inspection to assess rendering quality, room and object visibility, and consistency between the video and its metadata. During construction, shortcut diagnostics is implemented examine answer-frequency biases, question and option wording, and room-type priors. These checks guide revisions to question templates and answer balancing, while blind and frequency baselines provide additional diagnostics on the evaluation set.

Training data. To support model development, we additionally provide RS-30K, a training dataset of 30k questions generated from ProcTHOR training scenes. RS-30K uses a separate set of question templates and shares no scenes, routes, or videos with RESPACEBENCH. It provides targeted supervision without reusing evaluation environments or walkthroughs. We assess its utility through supervised fine-tuning of Qwen3-VL-8B, with results reported in Sec. 4.

3.2 BENCHMARK STATISTICS

RESPACEBENCH contains 3,244 evaluation questions, including 3,026 questions derived from synthetic environments and 218 from real homes. Fig. 3(a) shows their distribution across 3 task levels and 9 task families. We additionally use RESPACEBENCH-*mini*, a 450-question subset containing 50 samples per task, for human evaluation and agentic experiments (Sec. 4). RS-30K is a separate training dataset and is therefore excluded from these evaluation statistics.

3.3 EVALUATION PROTOCOL

Each model receives the walkthrough video and a textual question, together with textual answer options for multiple-choice items. T1 additionally provides a short query clip showing a room from a new viewpoint. The model must match this room to a previously entered room or select the "not entered" option. Scene metadata, camera poses, and ground-truth event annotations are only for further analysis and not provided to models under the default evaluation protocol.

Responses use one of two formats: an option letter for *multiple-choice* questions or a value for *numeric* questions. Multiple-choice responses are scored by whether the selected option is correct, while numeric responses are scored by exact match to the ground-truth value. Overall accuracy is computed as the unweighted mean of the 9 task-family accuracies. We report accuracy separately for each task family, alongside the Blind, Chance, and Frequency baselines described in Section 4.

4 EVALUATIONS ON RESPACEBENCH

We evaluate 32 models covering proprietary MLLMs, open-source general-purpose MLLMs, unified models, and specialized spatial and video-understanding models, as well as 3 agentic video-understanding frameworks. Except for targeted fine-tuning, models use their standard inference interfaces without benchmark-specific training. For direct inference, we sample videos at 1 fps by default, retaining approximately 6–9 views per room-entry look-around, and further downsample when required by model-specific context limits. Details are provided in the appendix.

Table 2: Evaluation of 32 models on RESPACEBENCH. Darker green entries denote the best among proprietary models and lighter green entries denote the best among open-source models. †Qwen3-VL-8B fine-tuned on RS-30K training data generated from the ProcTHOR training set.

Methods	Rank	Avg.	Association			Persistent state				Reasoning	
			T1	T2	T3	T4	T5	T6	T7	T8	T9
Baseline											
Chance Level (Random)	–	–	25.0	33.3	–	–	–	–	25.0	25.0	25.0
Chance Level (Frequency)	–	–	25.1	34.0	25.9	27.8	22.4	32.1	25.5	26.7	25.7
Blind (Seed 2.0 Pro (Bytedance Seed, 2026))	–	27.4	23.4	37.0	25.1	25.1	19.2	25.4	21.4	33.1	36.8
RESPACEBENCH-*mini*											
GPT-5.6 Sol (OpenAI, 2026)	–	52.2	56.0	60.0	48.0	44.0	54.0	40.0	56.0	50.0	62.0
Human (Mean of 3)	–	90.1	90.0	94.7	91.3	88.0	87.3	90.0	86.7	92.7	90.7
Proprietary Models											
GPT-5.6 Sol (OpenAI, 2026)	1	50.5	56.2	65.4	34.2	55.4	58.4	30.0	47.0	41.9	66.2
GPT-5.6 Luna (OpenAI, 2026)	17	32.5	38.6	44.2	14.0	31.8	39.0	32.7	34.9	25.0	32.4
Seed 2.1 Pro (ByteDance Seed, 2026)	3	48.0	58.8	59.0	26.7	48.3	54.4	30.0	46.3	44.2	64.7
Seed 2.0 Pro (Bytedance Seed, 2026)	2	49.3	52.1	60.3	29.1	48.8	52.3	31.8	49.7	43.0	76.3
Qwen 3.8 Max Qwen Team (2026a)	4	46.3	50.0	49.4	18.9	48.5	56.0	27.4	61.7	43.0	62.3
Gemini 3.1 Pro Google DeepMind (2026)	12	34.0	44.2	39.7	19.5	25.4	36.9	19.7	44.0	30.8	45.4
Open-source General Purpose Models											
Qwen3.6-27B (Qwen Team, 2026b)	6	40.3	34.5	40.4	27.4	42.6	55.4	33.2	48.8	27.3	53.1
Qwen3.6-35B-A3B (Qwen Team, 2026c)	8	36.8	32.0	48.7	22.6	38.2	35.3	25.6	40.8	32.6	55.6
Qwen3-VL-32B (Qwen Team, 2025)	5	41.3	45.1	44.9	30.1	36.5	49.3	29.6	41.5	36.6	58.0
Qwen3-VL-8B (Qwen Team, 2025)	10	35.5	43.1	36.5	23.6	38.4	42.7	26.9	41.5	25.6	41.5
Qwen3-VL-4B (Qwen Team, 2025)	16	32.7	39.1	35.9	24.5	26.1	32.1	30.9	32.2	28.5	45.4
InternVL3.5-38B (Wang et al., 2025a)	7	38.0	44.2	40.4	25.5	31.3	31.8	34.1	44.5	34.9	55.6
InternVL3.5-30B-A3B (Wang et al., 2025a)	19	30.7	28.3	39.7	24.2	35.5	24.1	29.1	38.5	24.4	32.4
InternVL3.5-8B (Wang et al., 2025a)	21	30.4	25.8	32.7	26.1	29.3	23.1	29.6	39.4	27.3	40.6
GLM-4.6V-106B (Z.ai, 2025)	13	33.1	38.8	37.2	26.0	32.8	12.5	32.7	47.6	26.2	44.4
GLM-4.6V-Flash-9B (Z.ai, 2025)	22	29.0	42.3	34.0	21.1	23.2	15.4	17.0	40.3	24.4	43.5
Gemma-4 26B-A4B (Gemma Team, 2026)	15	32.8	29.8	35.3	25.7	32.3	34.5	26.5	39.0	28.5	44.0
Gemma-4 E4B (Gemma Team, 2026)	29	25.0	33.5	35.9	26.5	13.3	27.3	0.4	31.1	20.9	35.7
Gemma-4 E2B (Gemma Team, 2026)	30	24.2	26.4	30.8	26.4	20.4	22.5	16.1	16.3	23.3	35.7
Cosmos3-Super-64B (Agarwal et al., 2026)	13	33.1	40.8	35.3	21.7	24.9	38.7	27.8	38.5	29.1	41.5
Video Understanding Models											
Video-XL-2-8B (Qin et al., 2025)	24	28.7	25.5	37.8	27.3	10.8	27.3	26.5	36.9	16.9	49.8
VideoChat-Online-4B (Huang et al., 2025)	28	26.2	24.0	35.9	25.1	17.2	21.0	20.6	32.0	20.9	38.6
Flash-VStream-7B (Zhang et al., 2025a)	26	27.1	23.8	32.7	26.3	20.2	22.3	29.1	32.7	23.3	33.8
Mage-VL (Yang et al., 2026b)	9	35.7	34.5	38.5	28.8	29.6	40.1	43.0	36.0	32.0	38.6
RoboBrain2.5-8B (Tan et al., 2026)	20	30.5	32.6	34.6	26.0	28.6	26.5	29.6	36.5	20.9	39.6
LLaVA-OneVision-2-8B (An et al., 2026)	11	34.4	30.5	37.2	25.4	34.2	35.8	37.7	39.7	31.4	38.2
InternVideo3-8B (Yan et al., 2026)	18	30.9	30.3	37.2	27.9	32.5	28.6	23.3	38.8	24.4	35.3
Spatial Understanding Models											
SenseNova-SI-1.5-8B (Cai et al., 2026)	23	28.8	32.8	30.8	25.2	23.9	24.1	30.0	36.0	25.0	30.9
VST-7B-RL (Yang et al., 2025b)	25	28.4	26.8	38.5	23.3	19.5	26.3	29.6	39.4	19.8	32.9
Cambrian-P-7B (Yang et al., 2026a)	27	26.5	28.5	32.7	19.6	9.1	28.1	30.9	26.5	24.4	38.6
Cambrian-S-7B (Yang et al., 2025c)	31	23.8	25.1	35.9	16.2	5.4	18.8	17.5	28.3	26.2	40.6
Spatial-TTT-3B (Liu et al., 2026)	32	22.0	25.5	32.1	0.0	13.1	14.1	32.7	30.4	19.8	30.0
Ours											
Qwen3-VL-8B (Qwen Team, 2025)	–	35.5	43.1	36.5	23.6	38.4	42.7	26.9	41.5	25.6	41.5
Qwen3-VL-8B-SFT (Blind)	–	29.6	26.4	32.7	25.0	23.2	23.3	32.1	34.4	34.3	35.1
Qwen3-VL-8B-SFT (Ours)†	–	60.4	85.6	86.5	63.4	45.3	45.6	52.5	58.5	61.6	44.4

Blind, chance, and frequency baselines. *Blind* evaluates Seed 2.0 Pro with task instructions, questions, and answer options without any visual input. *Chance* averages $1/K$ over multiple-choice questions with K options. *Frequency* predicts the most frequent ground-truth option letter or numeric value within each question subtype, with scores weighted by subtype question counts. It is computed from the evaluation answer distribution to quantify answer imbalance. All model responses follow the task-specific scoring rules (Section 3.3).

4.1 MAIN RESULTS

The task-level results in Tab. 2 reveal uneven performance across the benchmark. Among models evaluated without task-specific fine-tuning, there is a 9.2% gap between GPT-5.6 Sol (50.5%) and Qwen3-VL-32B (41.3%), the best performing proprietary and open-source general-purpose models, respectively. However, rankings varies across tasks. Mage-VL achieves the highest accuracy on unique-object counting (T6) among these baselines, at 43.0%. Its overall accuracy is 35.7%, the best among video-understanding models but falls behind the general-purpose MLLMs. All evaluated spatial-understanding models perform poorly, with overall accuracies ranging from 22.0% to 28.8%.

Table 3: Agentic video understanding frameworks on RESPACEBENCH-*mini*(%). Higher-capability and lightweight backbones are replaced with GPT-5.6 Sol and GPT-5.6 Luna, respectively, with all other configurations unchanged. VA denotes VideoAgent.

Task	Chance	Direct	AVP	DVD	VA
T1	25.0	**56.0**	52.0	28.0	30.0
T2	33.3	**60.0**	54.0	54.0	36.0
T3	–	**48.0**	22.0	30.0	32.0
T4	–	44.0	32.0	**48.0**	16.0
T5	–	**54.0**	40.0	52.0	34.0
T6	–	**40.0**	32.0	18.0	18.0
T7	25.0	56.0	**66.0**	56.0	36.0
T8	25.0	**50.0**	40.0	44.0	24.0
T9	25.0	62.0	58.0	**66.0**	44.0
Avg	–	**52.2**	44.0	44.0	30.0

Table 4: Frame selection strategies ablation on RESPACEBENCH-*mini* (%) using GPT-5.6 Sol. 1FPS denotes sampling the video at 1 frame per second and cov.-aware denotes coverage aware sampling to keep max information. Avg is the mean across nine tasks.

Task	1FPS	Uniform	Weighted	Cov.-aware
T1	56.0	54.0	48.0	**62.0**
T2	60.0	**66.0**	56.0	**66.0**
T3	**48.0**	26.0	44.0	44.0
T4	44.0	50.0	36.0	**58.0**
T5	54.0	**58.0**	44.0	56.0
T6	40.0	22.0	30.0	**42.0**
T7	56.0	52.0	56.0	**60.0**
T8	**50.0**	32.0	44.0	42.0
T9	62.0	68.0	**72.0**	68.0
Avg	52.2	47.6	47.8	**55.3**

Human Evaluation. We evaluate three independent human annotators on RESPACEBENCH-*mini*, with unrestricted video viewing and replay. Their mean accuracy reaches 90.1%, compared with 52.2% for GPT-5.6 Sol, a difference of 37.9%. Human performance remains high across all nine tasks, supporting the benchmark's solvability.

Agentic Video Understanding. We re-implement AVP (Wang et al., 2026c), DVD (Zhang et al., 2025b), and VideoAgent (Wang et al., 2024) on RESPACEBENCH-*mini*. To control for differences in backbone capabilities, we replace higher-capability models (e.g., Gemini 2.5 Pro, GPT-4 and o3) with GPT-5.6 Sol and lightweight models (e.g., Mini variants) with GPT-5.6 Luna, while retaining all other framework-specific configurations. Agents can access the full video and revisit segments through their respective tools. Under these configurations, none of the three frameworks outperform over direct inference. AVP and DVD each achieve 44.0%, while VideoAgent achieves 30.0%, compared with 52.2% for GPT-5.6 Sol (Tab. 3).

Targeted Fine-Tuning. Fine-tuning Qwen3-VL-8B on our separate 30k-question training set raises overall accuracy from 35.5% to 60.4%, a gain of 24.9 percentage points (Tab. 2). Accuracy improves on all nine tasks, with the largest gains in association and smaller, less consistent gains in persistent state and reasoning. The higher ProcTHOR accuracy may partly reflect similar design and furnishing styles within the dataset, despite no scene, route, or video overlap between training and evaluation. Average accuracy also improves on 3D-FRONT and real-world scenes, which are absent from fine-tuning, by 21.9 and 14.8 percentage points, respectively (Tab. 6). These results demonstrate the utility of targeted supervision, while the uneven gains motivate methods that more consistently maintain spatial state and use it for reasoning across revisits.

4.2 FURTHER ANALYSIS

Can models distinguish rooms across revisits? We analyze T1 errors across the six proprietary models, including the strongest zero-shot models. Errors comprise *wrong room* (choosing another visited room), *false merge* (labeling an unentered room as visited), and *false split* (labeling an entered room as unentered). Wrong-room selections dominate, accounting for 66.5–88.3% of errors (Figure 4(a)). False-split rates on visited-room queries and false-merge rates on unvisited-room queries are below their respective uniform-random baselines of 25.0% and 75.0% (Figure 4(b)). However, models still select an incorrect visited room in 44.8–58.0% of visited-room queries. These results suggest that models can distinguish visited from unvisited rooms to some extent, but often fail to identify which previously visited room the query depicts.

Why do agentic frameworks fail? We manually inspect failed trajectories from AVP, DVD, and VideoAgent on RESPACEBENCH-*mini*, reviewing their reasoning traces, tool calls, observed frames,

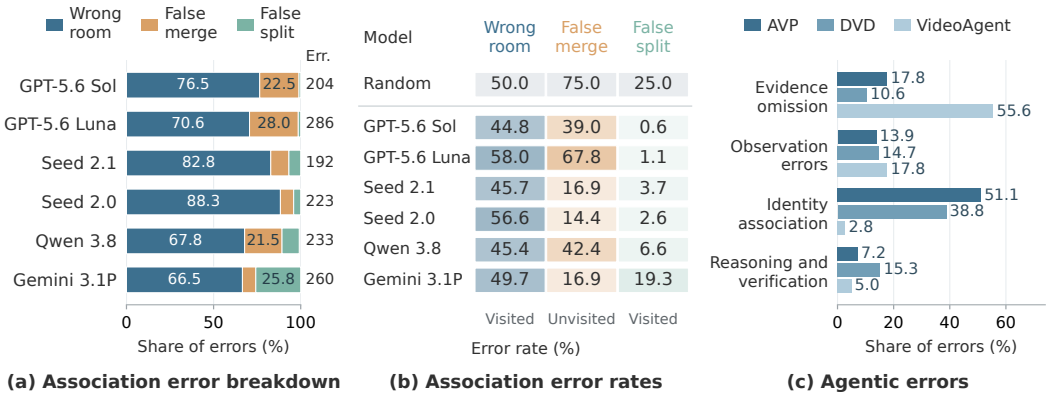

Figure 4: **Association errors and agentic failure patterns.** (a) Error distributions for six proprietary models on 466 T1 task queries, normalized by each model's total number of errors (Err.). (b) Error rates on visited-room queries ($n = 348$; wrong room and false split) and unvisited-room queries ($n = 118$; false merge). *Random* denotes uniform guessing among the options. (c) Shares of failures in four major categories on ReSpaceBench-*mini*, based on manual inspection of agent execution traces. Gemini 3.1P denotes Gemini 3.1 Pro.

and final answers (Fig. 4(c)). We identify four recurring major failure types: missing evidence, inaccurate observations, incorrect room-identity associations, and reasoning or verification errors. Trajectories with multiple problems are classified by the first observable error. AVP and DVD most often make errors in room identity association across revisit, whereas VideoAgent more frequently misses relevant evidence from video sequence. These findings suggest that evidence verification should check not only whether the necessary evidence has been retrieved, but also whether it has been interpreted correctly and linked to the correct room and visit.

Does preserving event coverage improve performance? We compare uniform, most relevant event weighted, and event coverage aware sampling on ReSpaceBench-*mini* using GPT-5.6 Sol, with identical prompts and a fixed budget of 128 frames. Both event-based strategies use walk-through route annotations: *event-weighted sampling* assigns $4\times$ the sampling weight to relevant room-entry events and *coverage-aware sampling* retains at least one frame from each room-entry, look-around, and hallway transitions interval. Coverage-aware sampling achieves 55.3%, compared with 47.6% for uniform and 47.8% for event-weighted sampling; the separate 1 fps baseline scores 52.2% (Table 4). Omitting an intervening bedroom visit, for example, can make two bathroom visits appear continuous. These annotation-guided results suggest that preserving intervening visits and transitions helps distinguish revisits from continuous stays, highlighting the importance of event coverage alongside query-relevant observations.

5 Conclusion

We introduced ReSpaceBench for evaluating persistent spatial memory across controlled revisits, together with RS-30K for targeted model training. Across synthetic and real homes, the benchmark evaluates room association, persistent state tracking, and downstream reasoning under viewpoint variation, identity ambiguity, and state changes. Our evaluations reveal a substantial human–model gap and no overall advantage from the evaluated agentic frameworks. The analyses expose a gap between recognizing a room as familiar and linking it to the correct spatial memory, while annotation-guided sampling demonstrates the value of retaining intervening visits and transitions. Fine-tuning on RS-30K improves Qwen3-VL-8B from 35.5% to 60.4%, with gains extending to scene sources absent from training, but state tracking and reasoning benefit less consistently than association. These findings motivate spatial memory mechanisms that preserve room identity across viewpoints, distinguish visually similar rooms, and update memories as room and object states change, while retaining visit histories for downstream reasoning. We hope that ReSpaceBench, RS-30K, and together with the analysis can provide a useful foundation for systematic evaluation, failure diagnosis, and targeted model development, and ultimately support further research toward more robust and generalizable persistent spatial memory and reasoning across revisits.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding and code review, and to generate illustrative icons and visual assets for figures to improve clarity of presentation. Additionally, we used generative AI tools to polish the language of selected paragraphs without altering their scientific meaning or the reported results. We have reviewed all AI-assisted work. Specifically, we manually reviewed the AI-assisted code and code review suggestions, checked the polished text for consistency with the original meaning, and reviewed the generated visual assets to ensure that they appropriately represented the intended concepts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuo Bao, Wei Dong, Shuyue Zhang, Ming Shang, Yuchen Huang, Han Yu, Chengjie Xu, Yiheng Bi, Kai Sun, Fuchun Sun, and Xinzhou Wang. PBD-AG: Persistent baseline-delta active graphs with uncertainty-aware inspection for long-horizon service robots. *arXiv preprint arXiv:2608.10449*, 2026.

Ermanno Bartoli, Buwei He, Dennis Rotondi, Sebastian Koch, Federico Tombari, Kai O. Arras, Patric Jensfelt, Yixi Cai, and Iolanda Leite. GESTO: Human-centric spatio-temporal memory for reasoning in dynamic scenes. *arXiv preprint arXiv:2608.10886*, 2026.

Bytedance Seed. Seed2.0 Model Card: Towards Intelligence Frontier for Real-World Complexity. *arXiv preprint arXiv:2607.00248*, 2026. URL https://arxiv.org/abs/2607.00248.

ByteDance Seed. Seed2.1 Officially Released: Advancing AI Productivity, June 2026. URL https://seed.bytedance.com/en/blog/seed2-1-officially-released-advancing-ai-productivity.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling Spatial Intelligence with Multimodal Foundation Models. In *CVPR*, pp. 7879–7890, 2026.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, Weiyue Li, Alan Yuille, Paul Pu Liang, and Yilun Du. MemoBench: Benchmarking world modeling in dynamically changing environments. *arXiv preprint arXiv:2606.27537*, 2026.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *NeurIPS*, volume 35, 2022.

Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. BlenderProc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023.

Pablo Diaz-Pereda, Alejandro Rodriguez-Ramos, David Perez-Saura, and Pascual Campoy. REMIND: RE-identification with memory for INDoor navigation. *arXiv preprint arXiv:2607.09267*, 2026.

FARO. FARO Focus Laser Scanning Solution, 2026. URL https://www.faro.com/en/Products/Hardware/Focus-Laser-Scanners. Accessed: 2026-09-23.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3d furnished rooms with layouts and semantics. In *ICCV*, pp. 10933–10942, 2021.

Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? *arXiv preprint arXiv:2103.06443*, 2021.

Gemma Team. Gemma 4 Technical Report, 2026. URL https://arxiv.org/abs/2607.02770.

Google DeepMind. Gemini 3.1 Pro Model Card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Nick Hawes, Christopher Burbridge, Ferdian Jovan, Lars Kunze, Bruno Lacerda, Lenka Mudrova, Jay Young, Jeremy Wyatt, Denise Hebesberger, Tobias Kortner, Rares Ambrus, Nils Bore, John Folkesson, Patric Jensfelt, Lucas Beyer, Alexander Hermans, Bastian Leibe, Aitor Aldoma, Thomas Faulhammer, Michael Zillich, Markus Vincze, Eris Chinellato, Muhannad Al-Omari, Paul Duckworth, Yiannis Gatsoulis, David C. Hogg, Anthony G. Cohn, Christian Dondrup, Jaime Pulido Fentanes, Tomas Krajnik, Joao M. Santos, Tom Duckett, and Marc Hanheide. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics and Automation Magazine*, 24(3):146–156, 2017. doi: 10.1109/MRA.2016.2636359.

Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan Bitton, Idan Szpektor, and Kai-Wei Chang. 3DLLM-Mem: Long-term spatial-temporal memory for embodied 3D large language model. *arXiv preprint arXiv:2505.22657*, 2025.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online Video Understanding: OVBench and VideoChat-Online. In *CVPR*, pp. 3328–3338, June 2025.

Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022.

Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, 3(4):4023–4030, 2018. doi: 10.1109/LRA.2018.2860628.

Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Emily Jin, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martın-Martın. Modeling dynamic environments with scene graph memory. In *ICML*, pp. 17976–17993. PMLR, 2023.

Mathieu Labbé and François Michaud. Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of field robotics*, 36(2):416–446, 2019.

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.

Chih-Ting Liao, Xi Xiao, Chunlei Meng, Zhangquan Chen, Yitong Qiao, Weilin Zhou, Tianyang Wang, Xu Zheng, and Xin Cao. SpaMEM: Benchmarking dynamic spatial reasoning via perception-memory integration in embodied environments. *arXiv preprint arXiv:2604.22409*, 2026.

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.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-TTT: Streaming Visual-based Spatial Intelligence with Test-Time Training. *arXiv preprint arXiv:2603.12255*, 2026. URL https://arxiv.org/abs/2603.12255.

Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J. Leonard, David Cox, Peter Corke, and Michael J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1): 1–19, 2016.

Ke Ma, Yamin Mao, Weiming Li, Shuai Tan, Yijie Zhong, Hao Chen, Haofen Wang, and Meng Wang. R4DSG: Relative 4D scene graph memory for object-centric question answering in long egocentric video. In *ACM MM*, 2026.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *ICCV*, pp. 6924–6934. IEEE, 2025.

Manjunath Narayana, Andreas Kolling, Lucio Nardelli, and Phil Fong. Lifelong update of semantic maps in dynamic environments. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6164–6171. IEEE, 2020.

OpenAI. GPT-5.6: Frontier Intelligence that scales with your ambition, July 2026. URL https://openai.com/index/gpt-5-6/.

PICO. PICO Mixed Reality, 2026. URL https://www.picoxr.com/global. Accessed: 2026-09-23.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-XL-2: Towards Very Long-Video Understanding Through Task-Aware KV Sparsification. *arXiv preprint arXiv:2506.19225*, 2025. URL https://arxiv.org/abs/2506.19225.

Qwen Team. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Qwen Team. Qwen3.8-Max: A New Bar for Coding and Cowork, August 2026a. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *ICCV*, 2019.

Lukas Schmid, Marcus Abate, Yun Chang, and Luca Carlone. Khronos: A Unified Approach for Spatio-Temporal Metric-Semantic SLAM in Dynamic Environments. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.081.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, et al. Robobrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, 2024.

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. In *ICML*, 2026a.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. MMLongBench: Benchmarking long-context vision-language models effectively and thoroughly. *arXiv preprint arXiv:2505.10610*, 2025b.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. EgoMemReason: A memory-driven reasoning benchmark for long-horizon egocentric video understanding. In *COLM*, 2026b.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *CVPR Findings*, 2026c.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv preprint arXiv:2407.13766*, 2024.

Hou Xia, Zheren Fu, Fangcan Ling, Jiajun Li, Yi Tu, Zhendong Mao, and Yongdong Zhang. VideoLevelGauge: Investigating contextual positional bias in large video language models. *arXiv preprint arXiv:2508.19650*, 2025.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. InternVideo3: Agentify Foundation Models with Multimodal Contextual Reasoning. *arXiv preprint arXiv:2606.12195*, 2026. URL https://arxiv.org/abs/2606.12195.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, pp. 10632–10643. IEEE, 2025a.

Jihan Yang, Zifan Zhao, Xichen Pan, Shusheng Yang, Junyi Zhang, Bingyi Kang, Hu Xu, Shang-Wen Li, and Saining Xie. Cambrian-P: Pose-Grounded Video Understanding. *arXiv preprint arXiv:2605.22819*, 2026a. URL https://arxiv.org/abs/2605.22819.

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

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, Jinghao Guo, Yifei Shen, Xinjie Zhang, Xiaoyi Zhang, Haoqing Wang, Xiao Li, Peng Zhang, et al. Mage-VL: An Efficient Codec-Native Streaming Multimodal Foundation Model. *arXiv preprint arXiv:2607.24904*, 2026b.

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*, 2025c. URL https://arxiv.org/abs/2511.04670.

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 *ICLR*, volume 2026, 2026c.

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*, volume 40, pp. 12000–12008, 2026.

Bowen Yuan, Sisi You, and Bing-Kun Bao. Predict before you explore: Predictive planning with specialized memory for embodied question answering. In *CVPR*, pp. 29610–29619, 2026.

Mubariz Zaffar, Sourav Garg, Michael Milford, Julian Kooij, David Flynn, Klaus McDonald-Maier, and Shoaib Ehsan. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *IJCV*, 129(7):2136–2174, 2021.

Z.ai. GLM-4.6V: Open Source Multimodal Models with Native Tool Use, December 2025. URL https://z.ai/blog/glm-4.6v.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *ICCV*, pp. 21059–21069. IEEE, 2025a.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *NeurIPS*, 2025b.

Xiwu Zhang, Lei Wang, and Yan Su. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, 113:107760, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107760. URL https://www.sciencedirect.com/science/article/pii/S003132032030563X.

A Implementation Details

A.1 Model Evaluation Settings

We resize video frames to 640×480 pixels and initially sample at 1 fps for default setting for all evaluated models. Due to model-specific context-length limits, we further uniformly downsample the sampled sequence when necessary, retaining the maximum number of frames that each model can accommodate alongside the question, answer options, reference inputs, and reserved response tokens. Sequences that fit are used in full. All retained frames preserve their chronological order. For multi-home visit questions, the context budget covers all sequences jointly.

For the models that can't fit our 1 fps video input, our evaluation use 320 for VST-7B-RL, GLM-4.6V-Flash, and GLM-4.6V (106B); 480 for Gemma-4 E2B, Gemma-4 E4B, Cambrian-S, and Cambrian-P; 150 for InternVL3.5-8B, InternVL3.5-38B, and InternVL3.5-30B-A3B; and 120 for SenseNova-SI-1.5.

A.2 Supervised Fine-Tuning

We fine-tune Qwen3-VL-8B-Instruct on RS-30K, comprising 30,018 questions from ProcTHOR training scenes. The language backbone is fully updated; the vision encoder and multimodal projector are frozen. Targets are the answer option letters or numeric answers. Table 5 lists the optimization settings. Both training and testing follow the video-input protocol in Section A.1, including the resolution, and 1 fps

Table 5: Supervised fine-tuning settings for Qwen3-VL-8B.

Hyperparameter	**Value**
Training epochs	1
Global batch size	32
Gradient accumulation	4 steps
Optimizer	AdamW
Learning rate	1×10^{-5}
Learning-rate schedule	Cosine decay
Warmup ratio	0.1
Adam (β_1, β_2)	$(0.9, 0.95)$
Weight decay	0.1
Gradient clipping (max. norm)	1.0
Numerical precision	bfloat16
Random seed	42

A.3 Results by Scene Source

Table 6 compares the base and fine-tuned Qwen3-VL-8B models on the same 1 fps evaluation questions, split into ProcTHOR, 3D-FRONT, and real-world scenes. Each subset's score averages its task accuracies with equal weight. Fine-tuning uses only ProcTHOR training scenes and improves the average on all three sources, including the real scenes.

Table 6: Per-task average accuracy (%) of base and SFT Qwen3-VL-8B at 1 fps, grouped by scene source.Δ is the SFT gain in percentage points.

Scene source	**Questions**	**Base Avg.**	**SFT Avg.**	**Δ**
ProcTHOR	1,471	34.7	74.5	+39.8
3D-FRONT	1,555	33.2	55.1	+21.9
Real-world	218	40.4	55.2	+14.8
All	3,244	35.5	60.4	+24.9

B Data Construction and Question Definitions

Room entries and persistent identity. Room entries are recovered from the rendered camera trajectory and its room labels. A visit requires entering a room; viewing it through a doorway does not count. Re-entry into the same physical room creates a new visit while preserving its identity. The initial countable room is included as the first visit, and hallways, corridors, stairs, and other transit spaces are excluded. Room inventories and first-encounter ordinals count each physical room once, whereas revisit counts retain separate entry events.

Ground-truth verification. Answers are recomputed from the verified room-entry sequence and object-state records. For entry-dependent questions, we check the annotated entry order against the recorded look-arounds. Questions with inconsistent annotations or ambiguous room references are excluded. Object counts use persistent instance identities to avoid counting the same object again on a revisit.

Task-specific evidence validation. To ensure disambiguity in synthetic scene annotations, room labels are checked against the furnishings and fixtures present after scene preparation. T1 (query-to-memory association) verifies whether the query room was entered, and T9 (temporal successor) uses the realized order of countable-room entries. Object related questions additionally applies geometric visibility and image-quality checks to the queried objects over multiple frames, with further framing and occlusion checks for reference images. Only visible objects in the video are counted towards the ground-truth.

B.1 Task Subtypes and Templates

Table 7 summarizes the task definitions and gives representative question wording from the evaluation set. Task numbers and names follow the taxonomy in Section 3.

Task	Subtypes / target	Representative template
T1 Query-to-memory association	Query clip matched to a previously entered room, or never entered (MCQ).	The video above shows one or more walks through homes. Did you walk into this room in the short clip during your walk(s)? If so, which one was it?
T2 Same-instance identification	Current room matched to an earlier room of the same type (MCQ).	Which of the rooms you entered earlier is the one you are in now?
T3 Revisit counting	Prior entries into the current physical room (numeric).	In the room you're currently in, how many times have you been to THIS EXACT room before?
T4 Room inventory	Distinct-room count by type; first-encounter ordinal (numeric).	You walked through a home. How many different bathrooms did you walk into? / Number the distinct rooms in the order you first walked into them. Which number is the kitchen?
T5 Object–place binding	First room containing an object category, in distinct-room order (numeric).	Number the distinct rooms in the order you first walked into them. What is the number of the first room you entered with a refrigerator in it?
T6 Unique-object counting	Unique object instances across entered rooms and revisits (numeric).	How many different sofas did you see in total for the rooms you entered?
T7 Cross-session state	First-to-second visit change; first/last locations; last observed room before disappearance; four-visit location history (MCQ).	The picture shows a chest of drawers you saw on your first visit. Comparing your first visit with your second visit, what happened to it? / Where was a potted plant in the living room on each of your four visits?
T8 Spatial topology	Room pair directly connected by a doorway (MCQ).	Which two of these rooms are directly connected to each other by a doorway (with no other room or passages in between)?
T9 Temporal successor	Next countable room type after a specified entry (MCQ).	Right after you left the kitchen for the first time, which kind of room did you walk into next? (Hallways and other passages you just walk through never count.)

Table 7: Question families and representative evaluation templates. The real-video cross-session extension also asks which objects are newly present on the second visit.