OVO-S-Bench: A Hierarchical Benchmark for Streaming Spatial Intelligence in Multimodal LLMs

Yifei Li[1,2,*] **Pengyiang Liu**[3,2,*] **Yuhang Zang**[2,†] **Zhongyue Shi**[3]
Qi Fu[3] **Hongye Hao**[3] **Jiwen Lu**[1]

[1]Tsinghua University [2]Shanghai AI Laboratory [3]Beihang University

[*]Equal Contribution [†]Project Leader

https://internlm.github.io/OVO-S-Bench/

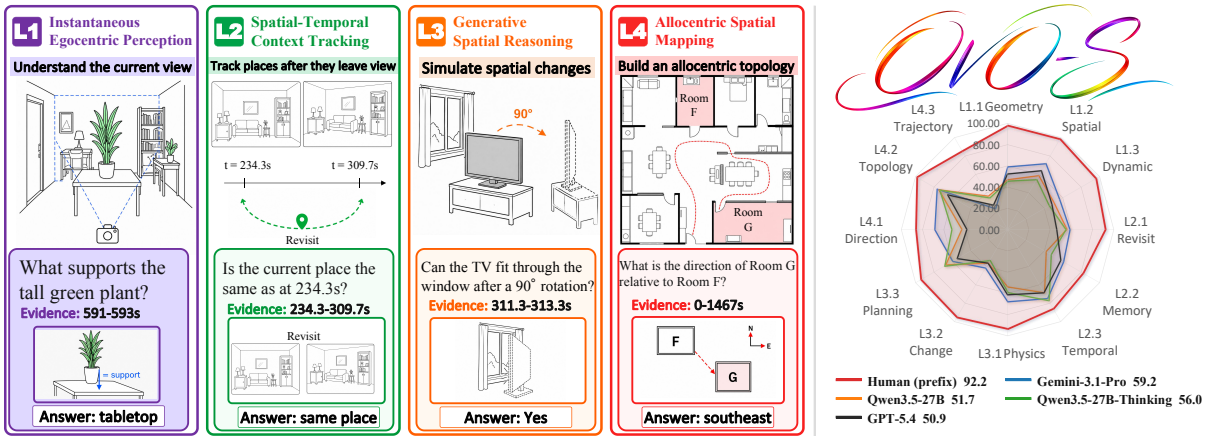


Figure 1: Overview of OVO-S-Bench. The benchmark evaluates streaming spatial understanding across four levels, from instantaneous egocentric perception and spatiotemporal context tracking to generative spatial reasoning and allocentric spatial mapping. The right panel summarizes representative model behavior across task families.

Abstract

Multimodal agents in robotics, AR, and autonomous driving must reason about places and layouts from continuous egocentric streams, often using evidence outside the current view. Existing benchmarks either evaluate offline over full videos or target events rather than spatial structure. We introduce **OVO-S-Bench**, a fully human-annotated benchmark for streaming spatial intelligence, comprising 1,680 questions over 348 source videos. Annotation involves 12 trained annotators (each also serving as a blind cross-reviewer) across roughly 804 person-hours of multi-round quality assurance. Each question carries a query timestamp and an evidence interval, and at evaluation, the model sees only the prefix preceding the query. Questions span four levels of increasing abstraction: instantaneous egocentric perception, spatiotemporal context tracking, generative spatial reasoning, and allocentric spatial mapping. Across 38 proprietary and open-source MLLMs, Gemini-3.1-Pro trails human experts evaluated under the same prefix-access protocol by 33 points (59.2 vs. 92.2), with allocentric spatial mapping as the dominant bottleneck. Notably, streaming and spatially fine-tuned MLLMs underperform their own backbones. We further find that chain-of-thought reasoning amplifies spatial errors when ungrounded in the stream. By exposing these limitations, OVO-S-Bench establishes a demanding testbed for next-generation streaming spatial MLLMs.

1 Introduction

Multimodal agents acting in the physical world must reason about space from a continuous egocentric stream. A household robot fetching an item must recall where it was last seen as the camera moves between rooms, and an AR navigation assistant must guide the wearer back to a landmark they walked past minutes ago (Grauman et al., 2022; Lv et al., 2024; Majumdar et al., 2024). In each setting, the evidence required to answer a query often lies outside the current frame, requiring the agent to retain, update, and reason about spatial structure over time rather than from any single image.

Existing benchmarks leave the streaming-spatial regime untested. Spatial benchmarks study 3D re-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 53894–53941
October 24-29, 2026 ©2026 Association for Computational Linguistics

Benchmark	Venue	Input	Stream	Avg. Query Time	Source	#Src.	#Tasks	Anno.	L1	L2	L3	L4
Image / Multi-image / Top-view (no video)												
EmbSpatial-Bench (Du et al., 2024)	ACL'24	Image	✗	—	Indoor	3	6	Auto.	✓	✗	✗	✗
TopViewRS (Li et al., 2024)	EMNLP'24	Top-view	✗	—	Indoor	1	9	Hybrid	✓	✗	✓	✓
MMSI-Bench (Yang et al., 2025d)	NeurIPS'25	Multi-img.	✗	—	Mixed	8	11	Manual	✓	✓	✗	✗
Video (offline)												
VSI-Bench (Yang et al., 2025a)	CVPR'25	Video	✗	1.9 min	Indoor	3	8	Hybrid	✓	✓	✓	✗
DISJOINT-3DQA (Ravi et al., 2025)	EMNLP'25	Video	✗	N/A	Indoor	1	5	Hybrid	✓	✓	✗	✗
STI-Bench (Li et al., 2025b)	ICCV'25	Video	✗	0.38 min	Mixed	3	8	Hybrid	✓	✓	✗	✗
MMSI-Video-Bench (Lin et al., 2025)	arXiv'25	Video	✗	1.2 min	Mixed	26	13	Manual	✓	✓	✓	✗
VSI-SUPER (Yang et al., 2025c)	ICLR'26	Video	✗	56.6 min	Indoor	3	2	Hybrid	✓	✓	✓	✗
Video (online / streaming)												
StreamingBench (Lin et al., 2026b)	ICLR'25	Video	✓	4.0 min	Mixed	1	18	Hybrid	✓	✓	✗	✗
OVBench (Huang et al., 2025)	CVPR'25	Video	✓	5.5 min	Mixed	7/8	16	Hybrid	✓	✓	✗	✗
OVO-Bench (Niu et al., 2025)	CVPR'25	Video	✓	7.1 min	Mixed	12+	12	Hybrid	✓	✓	✗	✗
OST-Bench (Lin et al., 2026a)	NeurIPS'25	Video	✓	20.12 frames	Indoor	3	15	Auto.	✓	✓	✗	✗
ODV-Bench (Zeng et al., 2026)	NeurIPS'25	Video	✓	0.32 min	Outdoor	6	12	Hybrid	✓	✓	✓	✗
StreamEQA (Wang et al., 2025c)	CVPR'26 Findings	Video	✓	1.13 min	Indoor	1	42	Hybrid	✓	✓	✓	✗
SVCBench (Liu et al., 2026c)	ECCV'26	Video	✓	2.0 min	Mixed	10	8	Manual	✓	✓	✗	✗
OVO-S-Bench (ours)	EMNLP'26	Video	✓	8.8 min	Mixed	9	30	Manual	✓	✓	✓	✓

Table 1: Comparison with spatial and streaming video benchmarks. **Stream** indicates a prefix-only protocol at query time; **Avg. Query Time** equals full video duration for offline benchmarks and the mean query timestamp (prefix length) for streaming ones ("N/A" if not open-sourced; OST-Bench is reported in frames as it releases only sampled frames). **L1–L4** follow the taxonomy in Section 3; ✓ marks tasks at that level. OVO-S-Bench is the only entry with all four levels checked under a streaming protocol; per-benchmark discussion in Appendix F.

lations, multi-view reasoning, and embodied question answering but assume static or offline visual context (Yang et al., 2025a,d; Jia et al., 2025), while long-video and streaming benchmarks target event understanding, narrative memory, or response timing rather than spatial structure in a continuous visual stream (Lin et al., 2026b; Niu et al., 2025; Lin et al., 2026a; Wang et al., 2025c). No existing benchmark bridges this gap: spatial evaluation remains tied to fragmented, discrete-frame perception, while streaming evaluation targets semantic and event recall, leaving continuous spatial reasoning from egocentric video untested (Table 1).

We introduce **OVO-S-Bench**, a benchmark for streaming spatial intelligence over continuous egocentric video. Each question includes a query timestamp and an evidence interval, and at evaluation, the model sees only the prefix preceding the query, thereby simulating an online agent without access to future frames. Questions span four levels of increasing spatial abstraction: **L1** instantaneous egocentric perception, **L2** spatiotemporal context tracking, **L3** generative spatial reasoning, and **L4** allocentric spatial mapping (§3). Higher levels generally demand evidence that persists longer, integrates across viewpoints, and supports allocentric abstraction beyond the current frame. The scope is deliberately delimited along the response axis: adopting the prefix-only protocol of OVO-Bench (Niu et al., 2025), we instantiate its *Real-Time* and *Backward* regimes with spatial content—answering from the current view and from evidence that has already left it—and exclude *Forward* proactive timing, so that spatial state maintenance is isolated from the separate question of when a model should speak.

Every question is written and verified by human experts. Annotators with 3D-vision backgrounds source long videos from indoor walkthroughs, egocentric activities, outdoor scenes, driving videos, and 3D environments, marking each item with options, a query timestamp, and an evidence interval. An independent group cross-reviews answerability and evidence sufficiency, and a text-only LLM probe flags items solvable from option text or world knowledge alone for revision; recurring leakage patterns are folded back into the guideline (§3.2).

We evaluate 38 systems across proprietary MLLMs, general video backbones, streaming-specialized architectures, and spatially fine-tuned variants, alongside Random, Text-Only, and a Human baseline from five independent participants (§4.2). Three findings stand out. *Allocentric spatial mapping is the dominant bottleneck:* L4 is the lowest-scoring level for 32 of 38 systems, and even Gemini-3.1-Pro (59.2 overall) trails the human baseline matched to its own prefix-access mode by 33 points, and the more constrained human streaming baseline by 27. *Specialized methods lag behind their backbones:* 13 of 15 streaming and spatially fine-tuned methods score below their base, with L4 as the most uniformly damaged level (mean $\Delta = -6.1$). *Thinking gains are model- and level-dependent:* explicit reasoning aids L2 cross-frame integration (mean $\Delta = +3.9$) but degrades L1 current-view perception (mean $\Delta = -1.0$), and

failure traces show that 60–80% of wrong answers stem from ungrounded spatial claims.

2 Related Work

Benchmarking Visual Spatial Understanding. Spatial benchmarks span image, multi-image, and video regimes. Image and top-view benchmarks evaluate metric estimation, object relations, and 3D geometry on static scenes (Chen et al., 2024; Jia et al., 2025; Li et al., 2024); MMSI-Bench (Yang et al., 2025d) extends to multi-image spatial intelligence; and video benchmarks add temporal observations for spatial recall (Yang et al., 2025a), spatio-temporal precision (Li et al., 2025b), and dynamic camera/object motion (Zhang et al., 2025c). Hierarchical efforts organize spatial abilities into progressively harder levels (Xu et al., 2025a; Xiao et al., 2026), and concurrent work scales open-world spatial evaluation to physical dynamics and human intent (Gu et al., 2026). All assume offline visual context: models can re-attend to any frame at query time, so the streaming constraint that makes spatial evidence ephemeral never surfaces.

Embodied and egocentric settings sharpen the need for persistent spatial memory. OpenEQA (Majumdar et al., 2024) frames embodied QA around episodic memory and active exploration, EmbSpatial-Bench (Du et al., 2024) and Open3D-VQA (Zhang et al., 2025b) evaluate spatial relations in embodied or aerial environments, and DisJoint-3DQA (Ravi et al., 2025) studies disjoint-frame reasoning where queried objects are not co-visible. These settings still operate over a curated context window rather than a continuous causal stream, leaving open whether models sustain a spatial model across arbitrary query timestamps.

Multimodal Models for Spatial Reasoning. Methods for spatial reasoning in MLLMs cluster around three strategies. *Data scaling* fine-tunes general-purpose backbones on large spatial corpora: VST pairs a 4M perception set with a 135K reasoning set under SFT+RL (Yang et al., 2025b), SenseNova-SI extends the recipe to 8M samples (Cai et al., 2025), and Holi-Spatial mines spatial supervision from raw video streams (Gao et al., 2026). *Architectural augmentation* injects 3D priors via spatially grounded supervision (Chen et al., 2024), dedicated spatial tokens (Wu et al., 2026), or camera-aware fusion (Zhao et al., 2025). *Inference-time adaptation* reshapes how the stream is consumed: Cambrian-S frames spatial supersensing as predictive world modeling with an active-memory manager (Yang et al., 2025c), and Spatial-TTT adapts fast weights at test time (Liu et al., 2026a).

Streaming Video Understanding. Streaming video understanding has two arms. *Long-form benchmarks*, including EgoSchema, LongVideoBench, and LVBench, test temporal understanding under offline access (Mangalam et al., 2023; Wu et al., 2024; Wang et al., 2025a; Liu et al., 2026b; Kwan et al., 2026). *Streaming benchmarks*, including StreamingBench, OVBench, OVO-Bench, OST-Bench, and SVCBench (Lin et al., 2026b; Huang et al., 2025; Niu et al., 2025; Lin et al., 2026a; Liu et al., 2026c), enforce causal input and bounded memory but target event understanding, multi-turn interaction, or counting rather than spatial structure. Concurrent streaming benchmarks extend the same causal constraint to omnimodal mobile assistants (Lu et al., 2026), egocentric episodic memory (Forte et al., 2026), and ad-hoc recall over text dialogue streams (Wang et al., 2026), each holding spatial structure outside its scope. Streaming MLLMs respond with memory banks, token pruning, KV-cache management, and event-segment memories (Qian et al., 2024; Zhang et al., 2024, 2025a; Yao et al., 2025; Zeng et al., 2026; Xie et al., 2026; Qian et al., 2025), and recent work asks whether such designs beat sliding-window baselines (Shen et al., 2026). All target event memory; persistent spatial evidence over a continuous egocentric stream is the dimension OVO-S-Bench makes empirical for the first time.

3 OVO-S-Bench

OVO-S-Bench organizes questions into four levels by the spatial state a model must access at query time (Figure 3). The levels progress from evidence directly available in the current view to allocentric map queries that require cross-viewpoint integration, reflecting a gradient of persistence and abstraction. Each level contains several task families and canonical task types; full definitions and examples are provided in Appendix D.

3.1 Four-Level Streaming Spatial Taxonomy

L1: Instantaneous Egocentric Perception. L1 tests whether a model understands the spatial structure currently in view. Questions can be answered from the frames near the query timestamp alone, without recalling any past observation. Task fam-

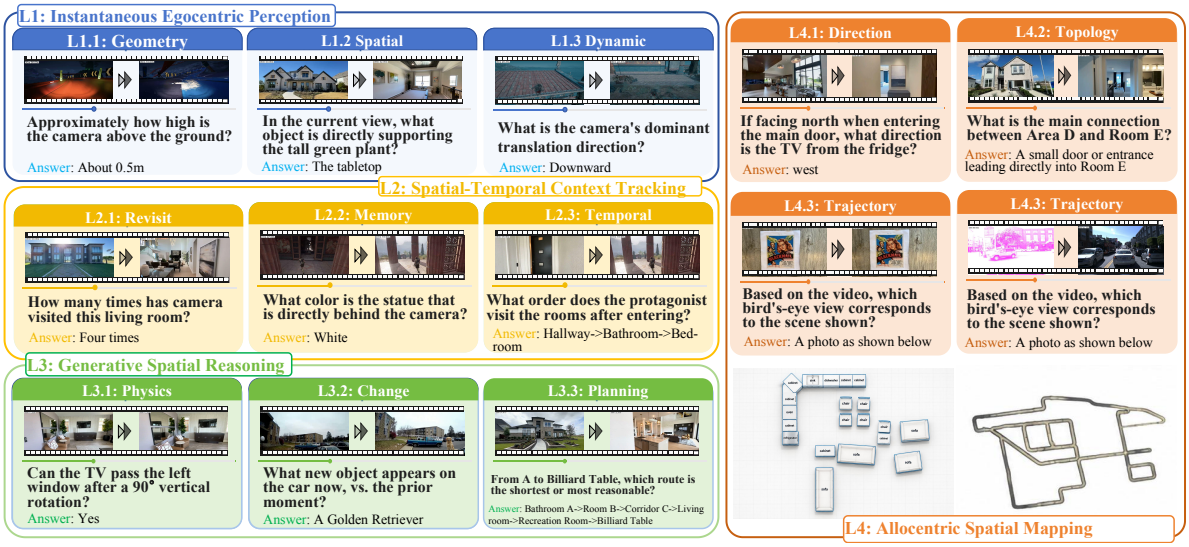


Figure 2: Representative OVO-S-Bench examples. Each card pairs a spatial question with visual evidence, illustrating the progression from current-view perception to allocentric spatial mapping.

ilies cover egocentric metric perception (distance, scale, clearance, viewpoint height), local spatial relations (containment, occlusion, support, visible layout), and dynamic spatial perception (camera motion, object motion, relative speed).

L2: Spatiotemporal Context Tracking. L2 evidence has appeared in the video prefix but is no longer visible at query time. The model must keep spatial facts bound to specific places, objects, and timestamps after visual support disappears. Task families include scene revisit recognition (whether a place has been seen and how often), spatial memory beyond the view (location and state of out-of-view entities), and chronological spatial memory (visit order and temporal-depth comparisons).

L3: Generative Spatial Reasoning. L3 requires the model to generatively operate on spatial structure rather than merely retrieve an observation. Questions involve mental rotation, state-change inference, and path feasibility; evidence spans vary by task, from local for simulation items to extended for route planning. Task families cover spatial simulation (reorientation, removal consequences, physical feasibility), spatiotemporal consistency verification (whether a state has changed between observations), and spatial route planning (route selection, detours, reachability).

L4: Allocentric Spatial Mapping. L4 requires the model to integrate the egocentric stream into an allocentric representation and query its global structure. Evidence typically spans multiple viewpoints and may cover the entire explored region. Three task families target allocentric direction reasoning (global bearings among rooms or landmarks), topological structure reasoning (adjacency, connectivity, boundaries, and transitions), and trajectory-map alignment (matching the first-person path to a bird's-eye map).

3.2 Benchmark Construction

Video sources. OVO-S-Bench draws from 9 publicly available or accessible sources covering five regimes: *indoor walkthroughs* (RoomTour3D (Han et al., 2025)), *egocentric activities* (Ego4D (Grauman et al., 2022)), *outdoor/world scenes* (Sekai (Li et al., 2025c), OmniWorld (Zhou et al., 2025), and YouTube walking tours), *driving videos* (CODa (Li et al., 2022) and Honda HDD (Ramanishka et al., 2018)), and *spatially annotated 3D environments* (ARKitScenes (Baruch et al., 2021) and VSI-Bench (Yang et al., 2025a)).

Human annotators write every item. Auto-generated captions and detections rarely capture the spatial evidence our targets require. Annotators select clips from these sources with stable motion, clear viewpoints, and sufficient spatial variation at the target level. For each item, they record the video, task label, question, options, answer, query timestamp, and evidence interval. Some task types employ specialized construction techniques such as image editing to generate spatial-change contrasts.

Each item follows the streaming setting. The answer must be derivable from the video prefix before the query timestamp. Annotators mark the shortest interval that contains the needed evidence and

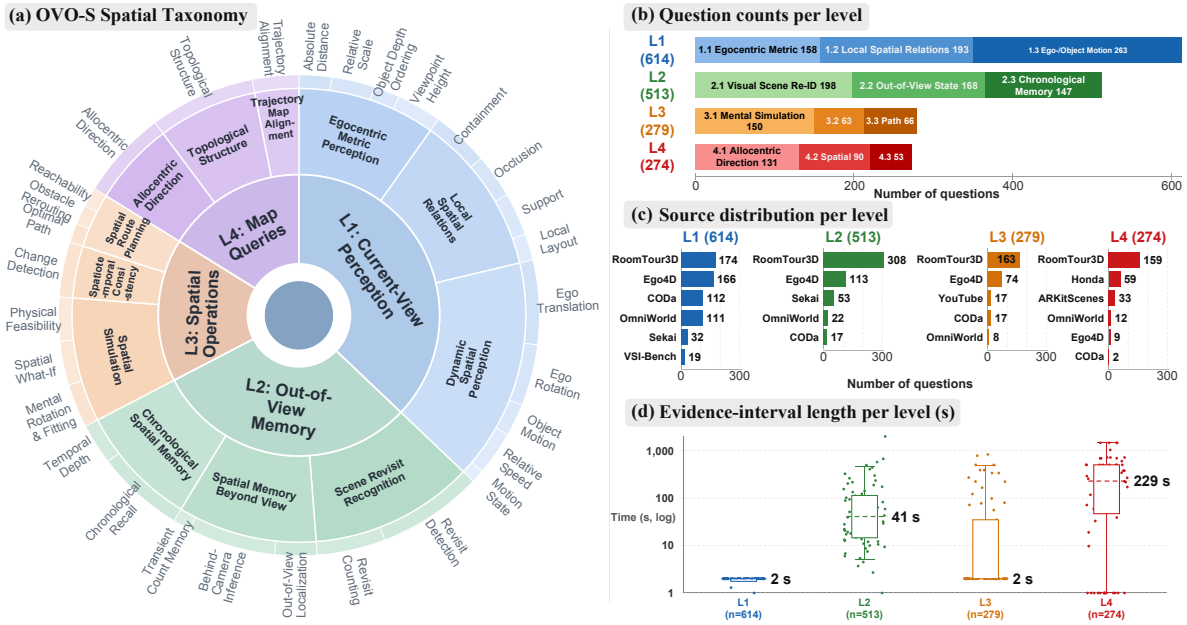


Figure 3: Taxonomy and benchmark statistics for OVO-S-Bench. The left panel gives the four-level spatial taxonomy, while the right panels report task-family counts, source distribution, and evidence-interval lengths by level.

write distractors that are plausible under the visual context but wrong under the annotated evidence.

Quality control removes shortcuts. We first run a text-only model to flag items that leak the answer through wording, common sense, option asymmetry, or a single lexical cue; this matters most for L2 through L4, where room priors and route stereotypes create shortcuts. We then probe the item under the intended streaming setting to expose ambiguity, easy visual shortcuts, and timestamp errors. A second annotator reviews the item without seeing the original answer and checks whether the answer and evidence interval are sufficient. Disagreements are revised, and recurring problems are folded back into the guideline. The full protocol, source details, and construction techniques are in Appendix E.

3.3 Benchmark Statistics

The released benchmark comprises 1,680 questions over 348 source videos from 9 datasets, organized into 30 canonical task types across four levels. Questions and scored queries differ in count: Revisit Counting asks the same question at several query timestamps, where the correct answer grows as revisits accumulate, so the 1,695-question public release is scored as 1,722 queries. Appendix tables computed on the public release state that scope explicitly; Table 2 and every main-text number use the 1,680-item evaluation set. The mean prefix at query time is 8.8 minutes. Evidence-span duration varies by level (L1 median 2.0 s, L2 36.8 s, L3 2.0 s, L4 278.7 s), reflecting the spatial persistence required at each level. Figure 3b–d reports per-level task-family counts, source distribution, and evidence-interval lengths; L4 is reported as three map-query families.

4 Experiments

4.1 Evaluation Setup

All systems see only the video prefix before query timestamp t_q, but we distinguish two evaluation modes. Under the *prefix-access* mode (default for proprietary, open-source, spatially fine-tuned, and embodied models), 128 frames are uniformly sampled from $[0, t_q]$ and presented together with the question; the model may attend freely over all sampled frames. This setup follows OVO-Bench and OST-Bench (Niu et al., 2025; Lin et al., 2026a). Under the *streaming* mode (applied to streaming-architecture and token-compression models), the video is fed sequentially at each model's published rate, and all memory and compression decisions are *query-agnostic*, made before the question arrives (Appendix G.4). Proprietary models are queried through their official APIs, all other models run locally with their published defaults, and answers are extracted by regular expression without further post-processing.

Every scored query also carries a response type on the axis OVO-Bench defines: L1 answers from the current view (*Real-Time*), L2 and L4 recall

or integrate observations that have already left it (*Backward*), and L3 splits per query by the gap between the end of the evidence interval and the query timestamp. The benchmark is 51% Real-Time and 49% Backward, and instantiates no *Forward* proactive-timing queries, matching the scope stated in Section 1. Appendix H gives the labeling rule, the per-level counts, and the sensitivity of the split to the L3 gap threshold.

We evaluate 38 systems (Tab. 2) spanning six families: *proprietary MLLMs* (GPT-5.4 (OpenAI, 2026), Gemini-3.1-Pro/Flash-Lite (Google DeepMind, 2026b,a), Grok-4.1-Fast (xAI, 2025)); *open-source general backbones* (InternVL-3.5 (Wang et al., 2025b), Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), Qwen3.5 (Qwen Team, 2026), Gemma-4 (Google DeepMind, 2026c), GLM-4.6V-Flash (Zeng et al., 2025)); *streaming video MLLMs* (Flash-VStream (Zhang et al., 2024), StreamForest (Zeng et al., 2026), StreamingVLM (Xu et al., 2025b)); *token-compression and memory methods* (InfiniPot-V (Kim et al., 2026), HERMES (Zhang et al., 2026a), FluxMem (Xie et al., 2026), StreamingTOM (Chen et al., 2025)); *spatially fine-tuned MLLMs* (Spatial-MLLM (Wu et al., 2026), Cambrian-S±LFP (Yang et al., 2025c), VST-SFT/RL (Yang et al., 2025b), SenseNova-SI (Cai et al., 2025), Spatial-TTT (Liu et al., 2026a)); and *embodied foundation models* (Cosmos-Reason1 (Azzolini et al., 2025), VeBrain (Luo et al., 2025a), RynnBrain (Dang et al., 2026), RoboBrain2.5 (Tan et al., 2026)). We anchor accuracy with three baselines: *Random*, *Text-Only* (GPT-5.4 with question and options only), and a *Human* baseline evaluated by five independent participants on the full question set under both modes: in streaming mode, participants watch the video without knowing the question and must rely on memory; in prefix-access mode, participants receive the question alongside the video prefix and may re-scan freely. Each family is compared against the human baseline collected under its own mode: *Human (streaming)* for the streaming-architecture and token-compression families, whose memory decisions are likewise query-agnostic, and *Human (prefix-access)* for all remaining families.

4.2 Benchmark Results

Tab. 2 supports four observations about the current state of streaming spatial intelligence.

Significant Gap with Human Performance. The strongest system, Gemini-3.1-Pro, reaches 59.2, trailing the human baseline collected under its own prefix-access mode (92.2) by 33.0 points; the best open-source model, Qwen3-VL-235B-A22B, reaches 53.6, a 38.6-point gap under the same pairing. Measured instead against the more constrained *Human (streaming)* score of 86.6—the reference for the query-agnostic families—the two gaps are 27.4 and 33.0 points. Random (31.3) and Text-Only (37.1) baselines fall below all general-backbone and proprietary systems, confirming that the human-model gap reflects genuine visual-streaming difficulty rather than language priors.

Allocentric Spatial Mapping Bottleneck. L4 is the lowest-scoring level for 32 of 38 systems, with an average gap of 9.3% between L1–L3 and L4 and gaps above 10% even on the largest open-source backbones (Qwen3-VL-235B-A22B: 10.6, InternVL-3.5-241B-A28B: 13.8). The six exceptions all have L1 below 41 (Qwen2.5-VL-7B, Flash-VStream, and four others), so their flipped ordering reflects degraded current-view perception rather than competent allocentric spatial mapping. Among L1–L3 the ordering is unstable (33 of 38 systems peak at L2 or L3): although evidence demand grows monotonically from L1 to L4, accuracy does not, and drops sharply once an allocentric map must be abstracted from the entire explored region.

Narrow but Uneven Closed-Source Lead. The closed-source advantage is only 5.6 points overall (Gemini-3.1-Pro 59.2 vs. Qwen3-VL-235B-A22B 53.6), narrower than the 10+ point gap reported on recent video and multimodal benchmarks (Fu et al., 2026). The gap is uneven across levels: closed-source widens its lead on the memory-heavy L2 (+5.9, above the +5.6 overall) but narrows on L4 (+4.1), and on L3 the best open-source backbone exceeds Gemini-3.1-Pro by 5.3 points (61.2 vs. 55.9). The closed-source group is itself heterogeneous: GPT-5.4 (50.9) and Grok-4.1-Fast (43.7) trail Gemini-3.1-Pro by 8 and 15 points, suggesting streaming spatial competence is not yet a generic capability of frontier multimodal systems.

Specialized Methods Lag Behind Backbones. No streaming-architecture or spatially fine-tuned variant outperforms its comparable general backbone, and 13 of 15 lag behind their own base. Streaming MLLMs are tuned to compress memory for narrative QA, and spatial fine-tuning (Wu et al., 2026; Cai et al., 2025) is tuned on discrete-frame QA; neither regime exercises persistent spatial state

Model	Params	L1: Instant. Ego. Perc.					L2: Context Tracking					L3: Gen. Spatial Reas.					L4: Alloc. Spatial Map.					Avg.	Rank
		1.1	1.2	1.3	Avg.	R	2.1	2.2	2.3	Avg.	R	3.1	3.2	3.3	Avg.	R	4.1	4.2	4.3	Avg.	R		
Baseline & Controls																						
Text-Only (GPT-5.4)	–	38.64	41.37	35.21	38.41	–	36.01	30.08	40.78	35.62	–	54.00	9.52	53.03	38.85	–	31.68	54.17	20.75	35.53	–	37.10	–
Random Baseline	–	33.93	25.49	30.00	29.81	–	37.44	26.77	41.08	35.10	–	50.00	25.00	25.00	33.33	–	25.00	31.25	25.00	27.08	–	31.33	–
Human (prefix-access)	–	97.47	97.93	95.44	96.95	–	91.41	81.55	85.71	86.22	–	93.33	95.24	93.94	94.17	–	86.14	98.33	83.02	89.16	–	92.20	–
Human (streaming)	–	91.77	95.85	92.02	93.21	–	86.87	77.98	78.23	81.03	–	88.67	87.30	83.33	86.43	–	79.21	86.67	71.70	79.19	–	86.61	–
Closed-source proprietary MLLMs																						
GPT-5.4	–	52.08	63.70	48.13	54.64	4	46.25	57.92	68.61	57.59	2	61.33	36.51	54.55	50.80	9	32.67	62.50	26.42	40.53	17	50.89	5
Gemini-3.1-Pro	–	58.69	**71.18**	**55.89**	**61.92**	1	**57.92**	**59.97**	**74.23**	**64.04**	1	**67.33**	41.27	59.09	55.90	3	**60.40**	**81.67**	22.64	**54.90**	1	**59.19**	1
Gemini-3.1-Flash-Lite	–	51.19	63.38	47.70	54.09	5	47.98	47.61	60.98	52.19	8	58.67	**49.21**	54.55	54.14	5	35.64	68.33	24.53	42.84	7	50.81	7
Grok-4.1-Fast	–	48.34	49.95	36.13	44.80	17	42.26	42.03	55.48	46.59	19	56.00	34.92	54.55	48.49	15	24.75	55.83	24.53	35.04	25	43.73	19
Open-source general video MLLMs																						
InternVL-3.5	8B	47.86	54.60	35.31	45.92	13	40.95	40.31	56.22	45.83	21	54.00	26.98	60.61	47.20	18	34.65	55.00	28.30	39.32	18	44.57	16
InternVL-3.5	38B	58.64	55.56	49.74	54.65	3	50.30	48.00	64.91	54.40	7	61.33	22.22	53.03	45.53	22	38.61	68.33	18.87	41.94	10	49.13	8
InternVL-3.5	241B-A28B	**61.30**	59.36	45.98	55.55	2	48.75	49.86	68.54	55.72	3	53.33	33.33	68.18	51.62	8	42.57	65.83	13.21	40.54	16	50.85	6
Qwen2.5-VL	7B	42.89	40.26	38.87	40.67	28	40.24	36.31	59.98	45.51	22	54.67	23.81	59.09	45.86	21	47.52	65.83	20.75	44.70	5	44.19	17
Qwen3-VL	4B	32.91	53.19	43.65	43.25	20	47.26	34.25	63.04	48.18	14	57.33	38.10	68.18	54.54	4	25.74	70.00	28.30	41.35	11	46.83	10
Qwen3-VL	32B	48.83	57.07	44.51	50.14	8	50.71	34.58	70.43	51.91	9	60.67	26.98	68.18	51.94	7	32.67	68.33	22.64	41.22	12	48.80	9
Qwen3-VL	235B-A22B	50.36	56.09	50.91	52.45	6	49.76	45.61	70.17	55.18	5	63.33	46.03	74.24	**61.20**	1	45.54	70.83	20.75	45.71	3	53.64	2
Qwen3.5	4B	40.28	51.13	44.82	45.41	15	47.44	34.19	62.98	48.20	13	52.00	39.68	56.06	49.25	14	27.72	65.83	22.64	38.73	19	45.40	13
Qwen3.5	9B	43.25	55.03	44.13	47.47	11	44.52	38.53	65.05	49.37	12	56.67	33.33	60.60	50.20	11	25.74	63.33	20.75	36.61	22	45.91	12
Qwen3.5	27B	46.50	58.13	49.88	51.50	7	55.24	41.53	68.67	55.15	6	54.00	36.51	66.67	52.39	6	35.64	71.67	**35.85**	47.72	2	51.69	4
Qwen3.5	397B-A17B	46.71	53.32	48.68	49.57	9	55.65	41.22	69.30	55.39	4	57.33	38.10	**78.79**	58.07	2	34.65	75.00	26.42	45.36	4	52.10	3
Gemma-4	E2B	42.12	42.36	31.93	38.80	33	35.00	31.97	42.60	36.52	35	55.33	12.70	50.00	39.34	33	28.71	37.50	22.64	29.62	35	36.07	36
Gemma-4	E4B	41.33	47.33	34.10	40.92	26	41.43	38.69	48.16	42.76	30	52.00	14.29	62.12	42.80	27	24.75	47.50	24.53	32.26	32	39.69	30
Gemma-4	26B-A4B	47.90	56.23	43.86	49.33	10	47.26	43.31	49.10	46.55	20	62.67	25.40	46.97	45.01	25	19.80	55.00	13.21	29.34	36	42.56	24
GLM-4.6V-Flash	9B	41.93	56.01	35.87	44.60	19	43.81	43.36	56.79	47.99	16	48.67	34.92	56.06	46.55	19	28.71	57.50	15.09	33.77	28	43.23	22
Streaming video MLLMs																						
Flash-VStream	7B	15.26	22.70	17.99	18.65	38	31.85	21.61	36.34	29.93	38	35.33	9.52	22.73	22.53	38	14.85	54.17	16.98	28.67	38	24.94	38
StreamForest	7B	47.69	56.74	35.48	46.64	12	42.74	36.64	56.22	45.20	24	60.00	30.16	59.09	49.75	12	26.73	59.17	18.87	34.92	26	44.13	18
StreamingVLM	7B	40.29	41.43	34.25	38.65	35	47.74	39.72	64.11	50.52	10	46.67	28.57	50.00	41.75	30	38.61	64.17	20.75	41.18	13	43.03	23
Token-compression and memory-based methods																						
InfiniPot-V	7B	38.66	41.66	36.82	39.05	31	34.82	26.81	45.53	35.72	36	56.67	17.46	51.52	41.88	29	33.66	47.50	–*	40.58	15	39.31	32
HERMES	7B	42.00	46.35	34.29	40.88	27	40.48	36.92	58.85	45.41	23	56.00	30.16	62.12	49.43	13	41.58	70.00	16.98	42.86	6	44.64	15
FluxMem	7B	41.46	44.58	42.89	42.98	22	43.33	39.36	59.98	47.56	17	53.33	28.57	54.55	45.48	24	48.51	64.17	15.09	42.59	9	44.65	14
StreamingTOM	7B	35.60	43.34	32.79	37.24	36	50.48	37.14	56.91	48.18	14	51.33	23.81	40.91	38.68	34	23.76	54.17	22.64	33.52	29	39.41	31
Spatially fine-tuned MLLMs																						
Spatial-MLLM	7B	34.39	34.35	38.41	35.72	37	40.00	29.03	48.53	39.19	33	54.00	17.46	31.82	34.43	36	23.76	55.00	30.19	36.32	23	36.41	35
Cambrian-S	7B	42.79	43.75	34.14	40.23	29	38.57	32.00	49.53	40.03	32	54.00	15.87	40.91	36.93	35	18.81	50.00	20.75	29.86	34	36.76	34
Cambrian-S-LFP	7B	38.72	42.14	35.62	38.82	32	37.80	25.61	50.72	38.04	34	52.00	14.29	36.36	34.22	37	22.77	42.50	20.75	28.68	37	34.94	37
VST-7B-SFT	7B	48.70	42.92	38.11	43.25	20	41.85	28.69	61.41	43.98	27	50.67	26.98	53.03	43.56	26	24.75	62.50	26.42	37.89	21	42.17	26
VST-7B-RL	7B	51.07	44.82	41.10	45.66	14	44.23	29.03	59.35	44.20	25	52.00	20.63	50.00	40.88	32	26.73	60.83	26.42	37.99	20	42.18	25
SenseNova-SI-1.5	8B	43.99	44.67	37.74	42.13	25	45.83	27.69	53.66	42.40	31	54.00	31.75	42.42	42.72	28	21.78	42.50	33.96	32.75	30	40.00	29
Spatial-TTT	2B	47.25	39.84	29.08	38.73	34	40.24	23.92	42.16	35.44	37	48.67	38.10	36.36	41.04	31	23.76	44.17	30.19	32.71	31	36.98	33
Embodied foundation models																						
Cosmos-Reason1	7B	48.20	43.85	42.31	44.79	18	36.85	35.94	58.29	43.69	28	55.33	20.63	60.61	45.52	23	30.69	50.00	15.09	31.93	33	41.48	28
VeBrain	7B	43.88	45.36	38.98	42.74	24	40.24	32.25	60.04	44.18	26	54.00	25.40	59.09	46.16	20	42.57	66.67	13.21	40.82	14	43.47	21
RynnBrain	8B	45.20	49.42	41.33	45.32	16	49.29	38.94	62.54	50.26	11	54.00	33.33	54.55	47.29	17	35.64	64.17	28.30	42.70	8	46.39	11
RoboBrain2.5	4B	36.70	45.60	38.04	40.12	30	42.92	31.64	55.72	43.43	29	52.00	30.16	62.12	48.09	16	22.77	54.17	30.19	35.71	24	41.84	27
RoboBrain2.5-NV	8B	34.68	51.38	42.53	42.86	23	45.95	35.56	58.41	46.64	18	56.67	28.57	66.67	50.63	10	26.73	49.17	26.42	34.10	27	43.56	20

Table 2: **Main results on OVO-S-Bench** (prefix-only protocol, multiple-choice accuracy). The streaming-architecture and token-compression families run in query-agnostic *streaming* mode and are compared against *Human (streaming)*; all other families run in *prefix-access* mode and are compared against *Human (prefix-access)* (§4.1). **R** is rank; top-three level ranks are shaded gray and overall ranks green (darker is better); bold marks the best per column; baselines and controls are unranked. *L4.3 uses image options unsupported by InfiniPot-V.

over a continuous prefix, consistent with the L2–L4 losses we measure. We dissect this further in Section 4.3.3.

4.3 Analysis Experiments

4.3.1 Effect of Thinking Mode

Tab. 3 compares thinking-mode and non-thinking-mode variants for nine models, where the paired delta isolates the effect of explicit chain-of-thought.

Model-Dependent Thinking Gains. Gains do not scale with backbone size: mid-size Qwen3.5 posts the strongest gain (from -1.3 at 4B to $+6.7$ at 9B), while Qwen3-VL and InternVL-3.5 hover near zero, Grok-4.1-Fast rises by $+3.2$, and GLM-4.6V-Flash falls by -0.9. Across levels, thinking consistently helps L2 (mean $\Delta = +3.9$, 8/9 pairs positive) and shows a small mean drop on L1 (mean $\Delta = -1.0$, 6/9 pairs negative), suggesting chain-of-thought aids the cross-frame integration but offers little for clean current-view perception.

Probing Thinking-Mode Failure Patterns. We probe failure modes of thinking-mode models in Tab. 3 by submitting every wrong-answer trace, together with the question, options, ground-truth, and predicted letters, and 4–6 frames sampled at midpoints of the annotated evidence intervals, to GPT-5.4 as a judge. The judge assigns one of five labels: **no-conclusion error** (no extractable letter), **non-visual error** (commits via world priors without citing visible content), **visual-content error**

Model	ΔOverall	ΔL1	ΔL2	ΔL3	ΔL4
GLM-4.6V-Flash	-0.9	+1.7	-1.1	-2.2	-1.9
InternVL-3.5-8B	+0.3	-1.3	+6.0	-1.8	-1.7
InternVL-3.5-38B	0.0	-7.7	+1.7	+3.2	+2.7
Qwen3-VL-4B	-2.5	-1.8	+0.1	-4.1	-4.3
Qwen3-VL-32B	+1.0	-0.8	+3.5	+2.3	-1.2
Qwen3.5-4B	-1.3	-4.9	+4.3	-6.1	+1.5
Qwen3.5-9B	+6.7	+3.1	+9.4	+4.3	+10.0
Qwen3.5-27B	+4.3	-0.5	+6.3	+1.5	+10.0
Grok-4.1-Fast	+3.2	+3.2	+5.2	-0.3	+4.5

Table 3: **Thinking-mode versus non-thinking-mode variants** on OVO-S-Bench. Δ is thinking minus non-thinking accuracy.

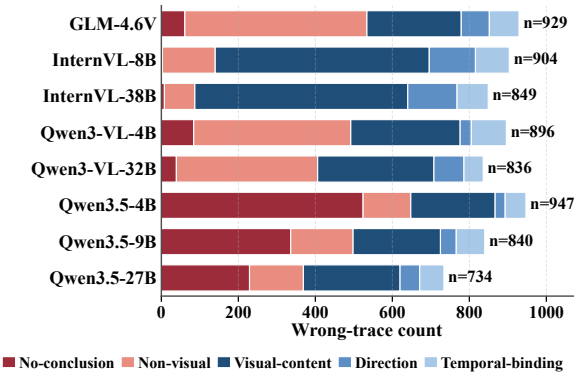


Figure 4: **Wrong-trace counts (bar length) and error-class composition (segments)** across eight thinking-mode models.

(asserts a spatial relation the frames do not show), **direction error** (correct objects, inverted L/R or F/B), and **temporal-binding error** (correct object, wrong moment or visit). Three votes per trace are aggregated by majority. Please refer to Appendix C for full setup, cross-judge calibration, per-level breakdowns, and worked examples of each error class with annotated CoT traces. Across the eight thinking-mode models (Fig. 4), failures concentrate on mis-grounded visual evidence: non-visual and visual-content errors jointly cover 60–80% of wrong traces in GLM-4.6V-Flash, Qwen3-VL, and InternVL-3.5. Qwen3.5-thinking is the sole exception, exhausting its 32K thinking-token budget without emitting an answer letter.

4.3.2 Frame Sampling Sensitivity

We ask whether frame-selection strategy explains the gap to human performance. Eight policies cover four regimes: a one-frame query baseline (*1@q*); a 16-frame sliding window at 4 fps (*16f@4fps*); uniform budgets *U-32/64/128/256* over the entire prefix; a recency-weighted variant *log-128* (60% within 30 s of the query, 30% in 30 s–5 min, 10% earlier), and an offline oracle *oracle-evidence* that

Model	1@q	16f@4fps	U-128	U-256	log-128	oracle
Gemini-3.1-Flash-Lite	47.95	49.72	**50.88**	48.33	48.55	49.70
Qwen3-VL-32B	43.95	45.80	48.71	48.93	**51.40**	48.80
Qwen3.5-27B	47.49	48.42	51.56	50.64	51.74	**51.87**
VST-7B-RL	41.24	39.10	42.13	41.99	**44.35**	42.13
RoboBrain2.5-NV-8B	39.91	41.22	43.56	43.87	**46.84**	43.56

Table 4: **Frame-sampling sensitivity on a five-model representative subset** (Overall accuracy; bold per row). Policy definitions and the full policy / model matrix are in Appendix B.

Method	Base	ΔOverall	ΔL1	ΔL2	ΔL3	ΔL4
Spatially fine-tuned						
Spatial-MLLM-7B	Qwen2.5-VL-7B	-7.70	-4.82	-6.16	-11.43	-8.39
VST-7B-SFT	Qwen2.5-VL-7B	-1.98	+2.57	-1.36	-2.30	-6.82
VST-7B-RL	Qwen2.5-VL-7B	-1.98	+4.92	-1.15	-4.98	-6.71
Spatial-TTT-2B	Qwen3-VL-2B	-3.54	-1.81	-10.12	-4.28	+2.04
SenseNova-SI-1.5	InternVL-3.5-8B	-4.55	-3.72	-3.43	-4.47	-6.57
Streaming architectures						
StreamingVLM-7B	Qwen2.5-VL-7B	-1.14	-2.09	+5.17	-4.11	-3.53
Flash-VStream-7B	Qwen2-VL-7B	-18.35	-21.69	-14.68	-20.28	-16.73
Token-compression and memory						
FluxMem	Qwen2.5-VL-7B	+0.45	+2.23	+2.05	-0.37	-2.11
HERMES	Qwen2.5-VL-7B	+0.44	+0.14	-0.09	+3.57	-1.85
StreamingTOM	Qwen2.5-VL-7B	-4.78	-3.43	+2.67	-7.18	-11.18
Embodied foundation models						
Cosmos-Reason1-7B	Qwen2.5-VL-7B	-2.63	+4.25	-1.66	-0.33	-12.78
VeBrain-7B	Qwen2.5-VL-7B	-0.64	+2.21	-1.17	+0.31	-3.89
RoboBrain2.5-4B	Qwen3-VL-4B	-4.90	-2.93	-4.60	-6.44	-5.64
RynnBrain-8B	Qwen3-VL-8B	-1.24	-0.05	-0.28	-4.95	+0.33
RoboBrain2.5-8B-NV	Qwen3-VL-8B	-4.07	-2.51	-3.89	-1.61	-8.27

Table 5: **Specialized methods versus their base backbones on OVO-S-Bench**. Δ is specialized minus base accuracy.

places 128 frames inside the evidence intervals. Formal definitions are in Appendix B.

Sampling Strategies Help Little. Tab. 4 reports overall accuracy under six policies on a representative subset. Neither the oracle nor *U-256* improves over *U-128*: the oracle gain is at most $+0.3$ across all five models, and *U-256* stays in $[-2.6, +0.3]$. The recency-weighted *log-128* matches or exceeds the oracle on three of five models (Qwen3-VL-32B $+2.6$, VST-7B-RL $+2.2$, RoboBrain2.5-NV $+3.3$ vs. *U-128*) and underperforms on the other two, so smarter sampling does not consistently close the gap. Short-context priors are level-selective: *1@q* wins Qwen3.5-27B L1 by $+7.8$ but loses L4 by -4.1 versus uniform, replicating the perception–memory trade-off. Together, these results indicate that the 27-point gap to human performance does **not** reduce to a retrieval problem solvable by better frame selection.

4.3.3 Specialized Methods versus Backbones

Streaming, token-compression, and spatially fine-tuned MLLMs target the regime OVO-S-Bench evaluates. Tab. 5 pairs each specialized method with its base backbone under the same protocol.

Specialization Loses Across Levels. Specialization is net-negative overall: 13 of 15 methods lose on overall accuracy (median -2.0, range -18.4 to $+0.5$), and only HERMES ($+0.4$) and

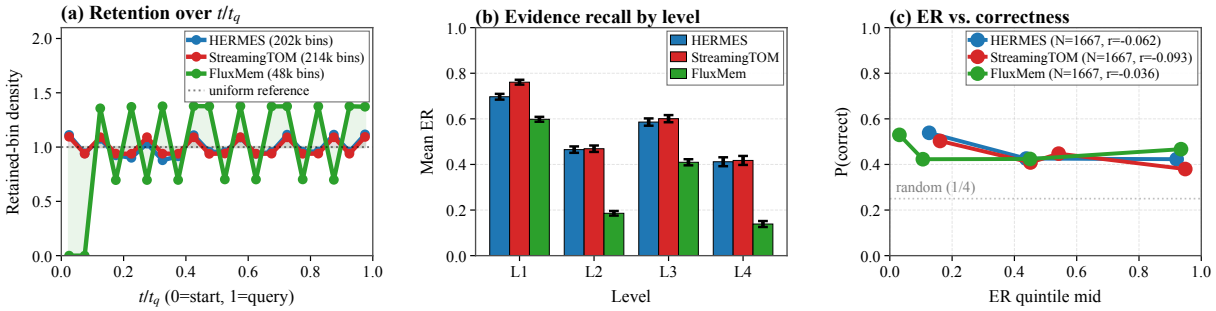


Figure 5: **Token-compression diagnostics. (a)** Retained 1-s bin density vs. t/t_q (dotted: uniform). **(b)** Evidence Recall (ER) by level. **(c)** ER vs. $P(\text{correct})$; legend shows per-method Pearson r.

FluxMem (+0.5) exceed their base. **L4 is the most uniformly damaged level:** 13 of 15 methods regress on allocentric spatial mapping (mean $\Delta = -6.1$, only Spatial-TTT +2.0 and RynnBrain +0.3 are positive), and Flash-VStream-7B (−16.7) and Cosmos-Reason1-7B (−12.8) show the largest drops. **L1 is occasionally improved:** perception-style specialization lifts current-view scores (VST-7B-RL +4.9, Cosmos-Reason1 +4.3, VST-7B-SFT +2.6). The L1/L4 split is consistent with regimes optimized for fragmented spatial QA, event memory, or robotic action grounding rather than long-range geometric evidence. This pattern spans all three families despite non-overlapping training data, making a pure domain-shift explanation unlikely (though controlled fine-tuning would be needed to confirm).

Retention Is Not the Bottleneck. For HERMES, StreamingTOM, and FluxMem, we instrument inference to dump, at every query timestamp, the original-video seconds surviving in memory, then compute the per-query *Compression Ratio* (CR; fraction of the prefix kept) and *Evidence Recall* (ER; fraction of the annotated evidence interval that survives). **Compression sheds evidence on memory-heavy levels:** mean ER on L4 ranges from 0.14 (FluxMem) to 0.42 (StreamingTOM), against 0.60–0.76 on L1 (Figure 5b), so the L2/L4 cost of compression originates before reasoning begins. **Yet retention alone does not predict correctness:** the per-query Pearson correlation between ER and correctness is essentially zero across all three methods ($r \in [-0.07, 0.00]$, panel c), and the retained-frame time distribution stays close to the uniform reference (panel a) rather than concentrating on the evidence interval. Consistent with Sec. 4.3.2, higher retention does not translate into correctness under these compressors.

5 Conclusion

We introduce OVO-S-Bench, a fully human-annotated benchmark that exposes the streaming spatial intelligence gap of current video MLLMs. Its 1,680 questions span four levels of spatial abstraction under a strict prefix-only protocol with annotated query timestamps and evidence intervals. Across 38 evaluated systems, Gemini-3.1-Pro (59.2) trails its matched prefix-access human baseline by 33 points and the more constrained human streaming baseline by 27, with allocentric spatial mapping as the dominant bottleneck; streaming and spatially fine-tuned variants score below their own backbones, and chain-of-thought amplifies spatial errors when ungrounded in the stream. We hope OVO-S-Bench helps shift spatial-intelligence evaluation from offline, discrete-frame benchmarks to continuous, streaming ones, where models must perceive, remember, and reason about space as it unfolds in real time, bringing evaluation closer to the embodied setting in which spatial intelligence must operate.

Limitations

While OVO-S-Bench covers the broadest mix of streaming-spatial scenes to date, from indoor walkthroughs to driving videos, the evaluated model remains a passive observer of pre-recorded streams and cannot close the perception-action loop that real embodied agents rely on. Each item is also scored via multiple-choice under a fixed prefix-only constraint. A 200-item open-ended pilot measures what that format costs: strict free-form accuracy falls 10.5 points below multiple choice on the same items, and 20% of free-form answers are partially correct, a band a single letter cannot express (Appendix I). The format therefore trades item-level resolution for deterministic scoring. The special-

ization analysis (Section 4.3.3) does not control for domain shift between each method's training data and OVO-S-Bench; controlled fine-tuning would be needed to fully isolate architectural limitations. This reflects a current trade-off in the field, as prior streaming-spatial benchmarks recover the loop only by trading scene diversity for controlled 3D environments, via camera-pose graphs over indoor scans (Lin et al., 2026a) or a full physics simulator (Hong et al., 2026). Moving forward, we plan to extend the protocol to a closed-loop regime that preserves the same diversity, toward benchmarks that evaluate not only what models perceive but also what they choose to perceive.

Ethics Statement

Intended use. The capability OVO-S-Bench measures matters most where spatial evidence outside the current view is safety-critical: assistive AR navigation and aids for blind and low-vision users must recall what lies behind the user or what was passed several minutes earlier. Measuring this under a streaming constraint, rather than over a fully observed video, is what makes the measurement informative for those settings.

Risks. Spatially unreliable models deployed in embodied or assistive systems can cause physical harm, and the failure modes documented here—ungrounded spatial claims and evidence discarded before reasoning begins—are the ones an offline, fully observed evaluation hides. We release the benchmark so that such failures surface before deployment rather than after.

Data provenance and licensing. All source videos come from publicly released datasets or public walking-tour footage and are used under their original terms; we do not relicense or redistribute video (Appendix E.2 lists the per-source terms). Our release contains human-written annotations only—questions, options, answers, query timestamps, evidence intervals, and per-source metadata—together with the source identifiers and access instructions needed to obtain the underlying videos from their original providers. The evaluation harness, prompt templates, and scoring code are available at `https://github.com/InternLM/OVO-S-Bench` and the annotations at `https://huggingface.co/datasets/JoeLeelyf/OVO-S-Bench`. No personally identifiable information was collected, and items describe scene structure rather than identifiable individuals.

Annotation. The 12 annotators were recruited from the authors' institutions and annotated as part of their research duties; recruitment, training, adjudication, and compensation arrangements are documented in Appendix E.1.

Acknowledgments

This work was supported by Shanghai Artificial Intelligence Laboratory.

References

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Ke qin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. 2021. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. 2025. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. 2025. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, et al. 2026. Rynnbrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*.

Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355.

Rosario Forte, Giuseppe Lando, and Antonino Furnari. 2026. EGOSTREAM: A diagnostic benchmark for streaming episodic memory in egocentric vision. *Preprint*, arXiv:2605.31557.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. 2026. Videomme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*.

Yuanyuan Gao, Hao Li, Yifei Liu, Xinhao Ji, Yuning Gong, Yuanjun Liao, Fangfu Liu, Manyuan Zhang, Yuchen Yang, Dan Xu, et al. 2026. Holi-spatial: Evolving video streams into holistic 3d spatial intelligence. *arXiv preprint arXiv:2603.07660*.

Google DeepMind. 2026a. Gemini 3.1 flash-lite. https://deepmind.google/models/gemini/flash-lite/.

Google DeepMind. 2026b. Gemini 3.1 pro. https://blog.google/products/gemini/gemini-3-pro/.

Google DeepMind. 2026c. Gemma 4. https://ai.google.dev/gemma/docs/core/model_card_4. Model card; technical report forthcoming.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Tianjun Gu, Jingyu Gong, Zhizhong Zhang, Yuan Xie, Lizhuang Ma, Xin Tan, et al. 2026. Vision-language models lag human performance on physical dynamics and intent reasoning. *Preprint*, arXiv:2601.01547. Introduces the EscherVerse benchmark.

Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. 2025. RoomTour3D: Geometry-aware video-instruction tuning for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yining Hong, Jiageng Liu, Han Yin, Manling Li, Leonidas Guibas, Li Fei-Fei, Jiajun Wu, and Yejin Choi. 2026. ESI-Bench: Towards embodied spatial intelligence that closes the perception-action loop. *arXiv preprint arXiv:2605.18746*.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. 2025. Online video understanding: Ovbench and videochat-online. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3328–3338.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. 2026. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *Advances in Neural Information Processing Systems*, 38:138983–139013.

Hohin Kwan, Hongyu Li, Ray Zhang, Manyuan Zhang, Xianghao Kong, Anyi Rao, Jiahao Xie, and Si Liu. 2026. Video-MME-Logical: A controlled diagnostic benchmark for video temporal-logical reasoning. *Preprint*, arXiv:2606.27828.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2024. TopViewRS: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1807.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. 2025a. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*.

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, Xiaodan Liang, Zhenguo Li, and Hang Xu. 2022. CODA: A real-world road corner case dataset for object detection in autonomous driving. In *Proceedings of the European Conference on Computer Vision*.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. 2025b. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5622–5632.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li,

Yukang Feng, Jianwen Sun, et al. 2025c. Sekai: A video dataset towards world exploration. *Preprint*, arXiv:2506.15675.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, et al. 2025. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. 2026a. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. *Advances in Neural Information Processing Systems*, 38.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. 2026b. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12147–12151. IEEE.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. 2026a. Spatial-ttt: Streaming visual-based spatial intelligence with test-time training. *arXiv preprint arXiv:2603.12255*.

Pengyiang Liu, Junbo Niu, Xiaoyang Hu, Zhongyue Shi, Zitian Wang, Linjiang Huang, and Si Liu. 2026b. TRACE: Temporal retrieval with anchored and convergent evidence for long-horizon video understanding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*.

Pengyiang Liu, Zhongyue Shi, Hongye Hao, Qi Fu, Xueting Bi, Siwei Zhang, Xiaoyang Hu, Zitian Wang, Linjiang Huang, and Si Liu. 2026c. SVCBench: A streaming video counting benchmark for spatial-temporal state maintenance. In *European Conference on Computer Vision*.

Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, Xue Yang, Rui Liu, and Hongsheng Li. 2026. PhoStream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. *Preprint*, arXiv:2601.22575.

Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. 2025a. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*.

Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, et al. 2025b. Robobench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. *arXiv preprint arXiv:2510.17801*.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. 2024. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. 2025. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18902–18913.

OpenAI. 2026. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Alibaba Cloud.

Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. 2018. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. *Preprint*, arXiv:1811.02307.

Sahithya Ravi, Gabriel Herbert Sarch, Vibhav Vineet, Andrew D Wilson, and Balasaravanan Thoravi Kumaravel. 2025. Out of sight, not out of context? egocentric spatial reasoning in vlms across disjoint frames. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16146–16161.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. 2026. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*.

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

Bingbing Wang, Jing Li, and Ruifeng Xu. 2026. Proactive memory for ad-hoc recall over streaming dialogues. *Preprint*, arXiv:2603.04885.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

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

Yifei Wang, Zhenkai Li, Tianwen Qian, Huanran Zheng, Zheng Wang, Yuqian Fu, and Xiaoling Wang. 2025c. Streameqa: Towards streaming video understanding for embodied scenarios. *arXiv preprint arXiv:2512.04451*.

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

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Spatialscore: Towards unified evaluation for multimodal spatial understanding. *arXiv e-prints*, pages arXiv–2505.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

xAI. 2025. Grok 4.1 fast and the Agent Tools API. https://x.ai/news/grok-4-1-fast.

Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. 2026. Spatialtree: How spatial abilities branch out in mllms. In *The First Workshop on Efficient Spatial Reasoning*.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. 2026. Fluxmem: Adaptive hierarchical memory for streaming video understanding. *arXiv preprint arXiv:2603.02096*.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. 2025a. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. 2025b. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. 2025b. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. 2025c. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*.

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

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. 2025. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10807–10816.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. 2025. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. 2026. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. 2025a. Flash-vstream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21059–21069.

Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. 2026a. Hermes: Kv cache as hierarchical memory for efficient streaming video understanding. *arXiv preprint arXiv:2601.14724*.

Weichen Zhang, Zile Zhou, Xin Zeng, Liu Xuchen, Jianjie Fang, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. 2025b. Open3d-vqa: A benchmark for embodied spatial concept reasoning with multimodal large language model in open space. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12784–12791.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Zhenfei Yin, Lijun Wang, and Huchuan Lu. 2026b. Think3D: Thinking with space for spatial reasoning. *Preprint*, arXiv:2601.13029.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. 2025c. Dsi-bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. 2025. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. 2025. OmniWorld: A multi-domain and multi-modal dataset for 4d world modeling. *Preprint*, arXiv:2509.12201.

Appendix

Content of Appendices

Section A. Per-Family Scaling Curves.
- Per-level accuracy versus backbone scale for Qwen3-VL, Qwen3.5, and InternVL-3.5.

Section B. Frame-Sampling Sensitivity.
- § B.1. Sampling Policy Definitions.
- § B.2. Overall-Accuracy Matrix.
- § B.3. Per-Level Breakdown.
- § B.4. Additional Analysis.
- § B.5. Length versus Difficulty.

Section C. CoT Failure Taxonomy.
- § C.1. Judge Setup and Pre-filtering.
- § C.2. Cross-Judge Calibration.
- § C.4. Per-Level Failure Distribution.
- § C.5. Worked Examples.

Section D. Detailed Category Definitions.
- L1–L4 subcategory definitions with accepted/rejected examples.

Section E. Annotation Protocol & Construction Pipeline.
- § E.2. Source Datasets and Licensing.
- § E.3. General Annotation Protocol.
- § E.4. Level-Specific Construction Techniques.
- § E.5. Quality Control Pipeline.

Section F. Extended Comparison with Prior Benchmarks.
- Per-benchmark discussion of what each benchmark does, shared points with OVO-S-Bench, and key differences.
- § F.5. Text-Only Shortcut Analysis.
- § F.6. Source-Robustness Analysis.

Section G. Evaluation Methods Details.
- § G.1. Models Evaluated.
- § G.2. Decoding Parameters.
- § G.3. Default Frame Sampling.
- § G.4. Streaming-Model Ingestion Protocol.
- § G.5. Prompt Templates.
- § G.6. Policy-Specific Inputs.
- § G.7. Answer Extraction.
- § G.8. Hardware and Software.

Section J. Future Work.

Section K. Additional Examples.

A Per-Family Scaling Curves

This appendix accompanies the corollary in Section 4.2 that L4 saturates early. We trace per-level accuracy as a function of backbone scale within three open-source families: Qwen3-VL (2B/4B/8B/30B-A3B/32B/235B-A22B), Qwen3.5 (0.8B/2B/4B/9B/27B/35B-A3B/122B-A10B/397B-A17B), and InternVL-3.5 (1B/2B/4B/8B/30B-A3B/38B/241B-A28B), all under default 128-frame uniform sampling with reasoning off, so the curves isolate scale. MoE checkpoints are marked with open rings, since the total parameter count overstates per-token compute.

Three Scaling Regimes. Figure 6 reveals three regimes. **L1 and L2 scale log-linearly:** end-to-end gains are $+11.8/+9.5$ on Qwen3-VL, $+13.9/+14.5$ on Qwen3.5, and $+20.3/+16.2$ on InternVL-3.5, with the two curves tracking each other within 3 points at every size. **L3 climbs but unevenly:** Qwen3.5 gains $+25.2$ overall yet Qwen3-VL drops from 54.5 at 4B to 49.9 at 30B-A3B before recovering to 61.2 at 235B-A22B, and the InternVL-3.5 38B dense checkpoint (45.5) underperforms its 30B-A3B MoE sibling (49.0), suggesting L3 depends on training-data composition as much as raw parameter count. **L4 saturates early:** all three families plateau in the 40–46 range from the mid-tier onward; InternVL-3.5 L4 moves only $+0.8$ over a $60\times$ parameter increase (39.7 at 4B vs. 40.5 at 241B-A28B), and remains roughly 39 points below human-streaming (79.2), indicating the allocentric bottleneck is not a capacity problem and will not close from scaling alone.

B Frame-Sampling Sensitivity

This appendix accompanies Section 4.3.2. We formalize the eight sampling policies, then report the full eight-model $\times$ eight-policy accuracy matrix at the Overall and per-level granularity. Three observations not included in the main text follow.

B.1 Sampling Policy Definitions

Let the video frame rate be f, the query time t_q seconds, and the query frame index $n_q = \lfloor t_q \cdot f \rfloor$. Let N be the total frame budget and $\mathcal{E} = \{[s_i, e_i]\}_{i=1}^{K}$ the set of annotated evidence intervals (in seconds). Every policy outputs an ordered frame index set $\mathcal{T} \subseteq \{0, 1, \ldots, n_q\}$, i.e. never consumes content posterior to the query, satisfying the streaming protocol.

single@query. The naive streaming baseline that uses only the query frame: $\mathcal{T}_{\text{sgl}} = \{\, n_q \,\}$ (N=1).

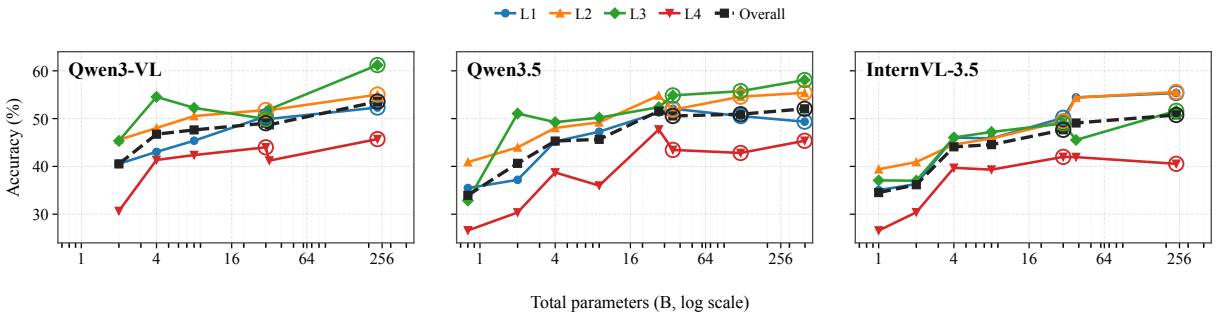


Figure 6: Per-level accuracy versus total parameter count (log scale) within three open-source families. Solid lines: L1–L4. Dashed line with square markers: overall accuracy. Open rings overlay MoE checkpoints.

nearest-16f@4fps. A short causal sliding window with step $\Delta = \max(1, \text{round}(f/f_w))$:

$$\mathcal{T}_{\text{rcw}} = \{\, n_q - k\Delta : 0 \le k < N,\ n_q - k\Delta \ge 0 \,\}.$$

The default N=16, f_w=4 fps covers the last $(N-1)/f_w = 3.75$ s before the query.

uniform-N. Uniform sampling of N frames over the entire prefix $[0, n_q]$:

$$\mathcal{T}_{\text{uni}} = \left\{\, \text{round}\left(\tfrac{k}{N-1}\, n_q\right) : 0 \le k < N \,\right\}.$$

We sweep $N \in \{32, 64, 128, 256\}$; *uniform-128* is the default policy used in Table 2.

oracle-evidence. An offline oracle that allocates the N=128 budget inside the annotated evidence intervals only. Each interval $[s_i, e_i]$ maps to a frame-domain interval $[\tilde{s}_i, \tilde{e}_i]$ clipped to $\le n_q$, with length $L_i = \tilde{e}_i - \tilde{s}_i$ and per-interval budget $N_i = \max(1, \text{round}(NL_i / \sum_j L_j))$ (rounding corrected on the longest interval). Each interval is then uniformly sampled:

$$\mathcal{T}_{\text{evid}} = \bigcup_{i=1}^{K} \text{linspace}(\tilde{s}_i, \tilde{e}_i, N_i).$$

When $\mathcal{E} = \emptyset$ the policy degrades to *uniform-N* over $[0, n_q]$. This serves as an upper bound: *how well a model could do if it had access only to the truly relevant frames.*

log-decay-128. A recency-weighted three-band schedule with N=128, band boundaries $\beta_1 = 30$ s, $\beta_2 = 300$ s, and weights $(\alpha_1, \alpha_2, \alpha_3) = (0.60, 0.30, 0.10)$:

$$\begin{aligned} \mathcal{B}_1 &= [n_q - \beta_1 f,\ n_q], & N_1 &= \text{round}(\alpha_1 N), \\ \mathcal{B}_2 &= [n_q - \beta_2 f,\ n_q - \beta_1 f], & N_2 &= \text{round}(\alpha_2 N), \\ \mathcal{B}_3 &= [0,\ n_q - \beta_2 f], & N_3 &= N - N_1 - N_2. \end{aligned}$$

Policy	Prefix coverage	Type
single@query (N=1)	last frame only	streaming
nearest-16f@4fps (N=16)	last 3.75 s at 4 fps	streaming
uniform-128	uniform over $[0, n_q]$	streaming
log-decay-128	3-band recency-weighted (60/30/10%)	streaming
oracle-evidence	only inside $\mathcal{E}$	offline

Table 6: Frame-sampling policies, N=128 unless noted. Streaming policies obey the prefix-only constraint; the oracle requires evidence annotations and is not available at deployment.

Each band is uniformly sampled. When the prefix is shorter than $\beta_2 f$ the missing band is skipped and its budget rolls into the surviving bands. This policy realizes a coarse exponential prior of near-importance without using ground-truth evidence.

Table 6 compares the five families at N=128.

B.2 Overall-Accuracy Matrix

Table 7 gives the Overall accuracy of eight models under all eight policies. Empty cells ("–") indicate that the model's context window cannot accommodate the required frame count.

B.3 Per-Level Breakdown

Table 8 reports the per-level accuracies (L1–L4) of the same eight models under the same eight policies. Per-row best is bolded. The level breakdown drives the three observations in §B.4.

B.4 Additional Analysis

The oracle is not an upper bound. On five of eight models (Cambrian-S-7B, VeBrain-7B, VST-7B-RL, Qwen3-VL-32B, RoboBrain2.5-NV-8B), *log-128* outperforms *oracle-evidence* on Overall by 0.5–3.3 points. The oracle places all 128 frames inside the annotated evidence interval, whereas the rest of the prefix is discarded; for questions whose answer also depends on global context (e.g. L3 counterfactuals and L4 trajectory queries), this

Model	single@q	nearest-16f	U-32	U-64	U-128	U-256	oracle-E	log-128
Gemini-3.1-Flash-Lite	47.95	49.72	**52.45**	51.26	50.88	48.33	49.70	48.55
InternVL-3.5-38B	45.04	44.45	48.94	49.49	49.08	–	**50.18**	49.33
Qwen3-VL-32B	43.95	45.80	45.89	47.18	48.71	48.93	48.80	**51.40**
Qwen3.5-27B	47.49	48.42	48.12	49.72	51.56	50.64	**51.87**	51.74
VST-7B-RL	41.24	39.10	41.50	43.62	42.13	41.99	42.13	**44.35**
Cambrian-S-7B	35.93	36.01	34.93	36.66	36.76	36.20	37.17	**37.67**
VeBrain-7B	43.51	45.17	41.65	43.82	43.47	43.30	43.53	**45.40**
RoboBrain2.5-NV-8B	39.91	41.22	40.82	42.63	43.56	43.87	43.56	**46.84**

Table 7: Full Overall-accuracy matrix on OVO-S-Bench (%). Eight representative models × eight frame-sampling policies. Per-row best is bolded. "–" = context-window overflow.

(a) L1: Instantaneous Egocentric Perception

Model	1@q	16f	U32	U64	U128	U256	Ora	Log
Gemini-3.1-Flash-Lite	**59.53**	59.46	53.12	53.14	53.88	53.40	53.28	53.60
InternVL-3.5-38B	59.87	**61.15**	53.35	54.04	54.44	–	54.97	58.48
Qwen3-VL-32B	56.46	**60.74**	49.36	49.66	49.93	51.37	50.19	55.34
Qwen3.5-27B	59.12	**61.88**	49.52	50.87	51.29	50.63	51.06	53.87
VST-7B-RL	**57.81**	52.45	45.47	45.96	45.45	46.16	45.45	50.73
Cambrian-S-7B	39.14	41.12	38.25	40.16	40.23	39.81	40.23	**42.21**
VeBrain-7B	52.64	**54.43**	41.67	42.08	42.74	43.02	42.74	44.52
RoboBrain2.5-NV-8B	**53.79**	53.25	41.14	43.42	42.86	44.06	42.86	50.14

(b) L2: Context Tracking

Model	1@q	16f	U32	U64	U128	U256	Ora	Log
Gemini-3.1-Flash-Lite	36.90	39.25	**56.88**	54.11	52.03	47.96	49.03	42.41
InternVL-3.5-38B	33.32	34.89	52.67	**55.03**	54.40	–	54.41	52.98
Qwen3-VL-32B	33.93	34.55	50.13	52.27	51.75	**53.53**	52.31	50.45
Qwen3.5-27B	35.82	35.71	48.89	52.35	54.83	54.09	**55.16**	52.46
VST-7B-RL	31.63	28.71	44.17	44.86	44.20	43.65	44.20	**45.22**
Cambrian-S-7B	34.56	34.29	36.79	38.56	40.03	39.75	39.99	**42.18**
VeBrain-7B	31.82	33.21	42.45	43.89	44.18	46.00	44.18	**47.73**
RoboBrain2.5-NV-8B	28.45	31.10	42.30	44.12	**46.64**	46.31	46.64	46.64

(c) L3: Gen. Spatial Reas.

Model	1@q	16f	U32	U64	U128	U256	Ora	Log
Gemini-3.1-Flash-Lite	50.85	53.12	54.56	**55.18**	54.14	51.32	54.29	53.65
InternVL-3.5-38B	46.06	44.70	**51.07**	48.94	45.53	–	46.01	47.35
Qwen3-VL-32B	52.25	50.25	50.13	48.42	51.94	**55.03**	52.45	54.02
Qwen3.5-27B	51.40	50.54	50.98	50.86	52.39	51.80	53.85	**56.04**
VST-7B-RL	38.60	40.56	39.00	**42.11**	40.88	41.07	40.88	40.87
Cambrian-S-7B	**41.36**	38.64	36.32	38.47	36.93	36.77	37.01	38.47
VeBrain-7B	46.01	45.87	42.23	**47.46**	46.16	44.88	46.67	43.78
RoboBrain2.5-NV-8B	43.65	44.28	42.81	46.66	50.63	**51.08**	50.63	49.22

(d) L4: Alloc. Spatial Map.

Model	1@q	16f	U32	U64	U128	U256	Ora	Log
Gemini-3.1-Flash-Lite	44.51	**47.03**	45.27	42.61	43.46	40.66	42.17	44.55
InternVL-3.5-38B	40.90	37.07	38.68	39.94	41.94	–	**45.32**	38.50
Qwen3-VL-32B	33.15	37.65	33.95	38.35	41.22	35.77	40.26	**45.80**
Qwen3.5-27B	43.61	45.54	43.10	44.79	**47.72**	46.03	47.42	44.57
VST-7B-RL	36.94	34.68	37.34	**41.54**	37.99	37.09	37.99	40.56
Cambrian-S-7B	28.68	29.98	28.37	29.45	29.86	28.49	**31.44**	27.84
VeBrain-7B	43.58	**47.16**	40.25	41.84	40.82	39.29	40.54	45.59
RoboBrain2.5-NV-8B	33.74	36.24	37.01	36.32	34.10	34.02	34.10	**41.36**

Table 8: Per-level accuracy (%) under eight frame-sampling policies. Per-row best is bolded. "–" = context-window overflow.

aggressive concentration removes useful framing. The *log-128* schedule, in contrast, retains 10% of the budget on the far past, which appears to compensate.

Backbone-dependent value of long context. Doubling the uniform budget from 128 to 256 frames yields negligible or negative deltas for all seven models with data (range -2.6 to $+0.3$): Gemini-3.1-Flash-Lite -2.6, Qwen3.5 -0.9, Cambrian-S -0.6, VeBrain -0.2, VST-7B-RL -0.1, Qwen3-VL $+0.2$, RoboBrain2.5-NV $+0.3$. No model benefits by more than 0.3 points, consistent with the scaling-curve finding (§A) that L4 saturates early across families and with the observation of Shen et al. (2026) that the marginal value of longer visual context is model-dependent.

Short-context priors and the perception–memory trade-off. Across all four large general backbones the one-frame query (*1@q*) wins L1 by 5–8 points over *uniform-128* (Gemini $+5.7$, InternVL-3.5-38B $+5.4$, Qwen3-VL-32B $+6.5$, Qwen3.5-27B $+7.8$) but loses L2 by 15–21 points (Gemini -15.1, Qwen3-VL -17.8, Qwen3.5 -19.0, InternVL-3.5 -21.1). The trade-off is sharpest for VST-7B-RL ($+12.4$ on L1 vs. -12.6 on L2) and is the empirical signature of the perception–memory trade-off (Shen et al., 2026) for OVO-S-Bench: a clean current view helps egocentric perception but cannot answer questions that require integrating across the prefix.

B.5 Length versus Difficulty

Evidence spans grow with taxonomy level (L1 median 2.0 s, L2 36.8 s, L3 2.0 s, L4 278.7 s), raising the question of whether length alone explains difficulty. We compute the query-level mean accuracy over the 38 reported models and correlate it with three length variables: visible prefix duration, evidence-span duration, and full source-video duration. Tab. 9 reports Spearman ρ per level and under two sets of controls.

Scope	Evidence ρ	n	Prefix ρ	n
L1	-0.04	629	0.00	629
L2	0.08	540	-0.08	540
L3	0.02	279	0.24	279
L4	0.17	274	0.26	274
Level-controlled	0.06	1722	0.06	1722
Level+source-controlled	0.04	1722	0.01	1722

Table 9: Spearman ρ between query-level mean accuracy and input length. Positive ρ means longer inputs correlate with *higher* accuracy (opposite to the "longer = harder" hypothesis). All values are small and none survive source controls.

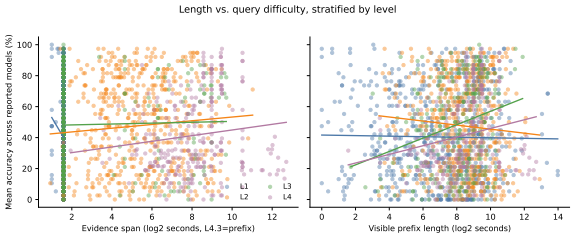


Figure 7: Query-level mean accuracy versus evidence-span duration (log scale), colored by level. Trend lines are per-level LOWESS fits.

Although evidence spans grow with level, length alone does not explain difficulty: the level-controlled correlations between accuracy and prefix length ($\rho = 0.06$) or evidence-span length ($\rho = 0.06$) are small, and full-video-duration correlation is similarly weak ($\rho = 0.03$, $p = 0.29$, computed analogously). With source controls these correlations shrink further. The small evidence-span effect is *positive* (longer spans correlate with slightly higher accuracy), opposite to the direction predicted by a "longer = harder" explanation. The allocentric-mapping bottleneck documented in Section 4.2 is therefore not reducible to a generic long-video effect; it reflects the abstraction demanded by L4, not the temporal extent of the evidence.

C CoT Failure Taxonomy

This appendix accompanies the failure-pattern analysis in Section 4.3.1. We give the formal judge setup (§C.1), the cross-judge calibration (§C.2), the per-level distribution (§C.4), and one worked example per error class (§C.5).

C.1 Judge Setup and Pre-filtering

Inputs and judge function. For each thinking-mode model $\mathcal{M}$ in Table 3, let $\mathcal{Q}^-_{\mathcal{M}} = \{\, q : a_q^{\mathcal{M}} \neq a_q^* \,\}$ be the set of queries on which $\mathcal{M}$'s extracted letter $a_q^{\mathcal{M}}$ differs from the ground truth a_q^*. For each $q \in \mathcal{Q}^-_{\mathcal{M}}$ the judge J applies

$$J(q, \mathcal{O}_q, a_q^*, a_q^{\mathcal{M}}, \tau_q^{\mathcal{M}}, \mathcal{F}_q) \rightarrow \ell \in \{\texttt{T1a}, \texttt{T1b}, \texttt{T2a}, \texttt{T2b}, \texttt{T2c}\},$$

where $\mathcal{O}_q$ is the multiple-choice option set, $\tau_q^{\mathcal{M}}$ the thinking trace (truncated to head-3,000 + tail-1,500 characters when $|\tau| > 5{,}000$), and $\mathcal{F}_q = \{f_1, \ldots, f_k\}$ with $k \in [4, 6]$ a set of 512 px JPEG frames extracted at midpoints of the annotated evidence intervals $\mathcal{E}_q$. The five label codes map to the paper-text classes *no-conclusion error* (`T1a`), *non-visual error* (`T1b`), *visual-content error* (`T2a`), *direction error* (`T2b`), and *temporal-binding error* (`T2c`).

Judge model. All wrong-answer traces are judged by `gpt-5.4`. To verify that the five-class scheme is robust to judge choice, we additionally re-judge all traces with `gemini-3.1-flash-lite`; see §C.2.

Voting. Every query is judged r=3 times under the same prompt with `stream=true`, `response_format=json_object`, and `max_tokens=200`. The primary label is the majority vote; records without a $\geq$2-vote majority are marked `disagreement` and excluded from the figures ($<0.5\%$).

Pre-filtering and signal-tagging. We drop 39 queries whose text contains an absolute timestamp ("at 134 seconds"), a prompt-engineering artifact rather than a reasoning failure. Decoding-truncated traces ($|\tau| > 50{,}000$ characters without a `</think>` close tag) are directly tagged `T1a` without calling the judge; this affects 591 Qwen3.5 traces that overran the 32K thinking-token cap.

Prompt skeleton. The system prompt fixes the five-class taxonomy and decision rules (Tier-1 takes precedence when both apply; prior phrases without visual citations route to *non-visual error*; direction errors require explicit textual evidence of the wrong direction). The user turn injects the per-query tuple *(Question, Options, Ground-truth, Predicted, Trace, Evidence frames)*, the last being the k images appended as base64 `image_url` fields. The verbatim prompts are released with the code.

C.2 Cross-Judge Calibration

We re-judged all wrong-answer traces with `gemini-3.1-flash-lite` under the identical prompt. Exact agreement is 75.9% with Cohen's

κ=0.65, which falls in the *substantial* band of the Landis–Koch convention and is sufficient for the qualitative cross-model claims made in Section 4.3.1.

C.3 Human Audit of Failure Labels

Cross-judge agreement establishes that the taxonomy is stable across judges. Whether the labels match human judgment on the subtlest classes is a separate question, and we measure it directly. Two authors relabeled a stratified sample of 120 wrong-answer traces, blind to the judge's label, under the decision rules of §C.1. The sample fixes 30 traces for each of *direction* and *temporal-binding*, the two classes where a wrong label is hardest to notice, and 20 for each of the remaining three, spread across levels and models. Traces tagged `T1a` by the truncation rule are excluded, since the judge never saw them.

The two annotators assign the same label to 103 of 120 traces (85.8%, Cohen's κ=0.81), so the five-class scheme is reproducible by human raters. On those 103 traces the judge matches the human label on 84 (81.6%, κ=0.77), above the cross-judge figure of §C.2. Per stratum, taking as denominator the traces of that judge class on which the annotators agree, the judge matches 19 of 24 *direction* traces and 19 of 25 *temporal-binding* traces, against 46 of 54 across the other three classes. The two subtle classes carry the lower agreement, at a rate that supports the pooled cross-model comparisons of Section 4.3.1. The sample, both annotators' label files, and the scoring script are released with the evaluation code.

C.4 Per-Level Failure Distribution

Figure 8 reproduces Figure 4 broken down by question level. Two patterns are invisible in the pooled view.

L2 elicits temporal-binding errors universally. All eight models score 17–36% *temporal-binding error* on L2 (spatiotemporal context tracking; mean 28.5%), against $\leq$ 5.6% on the other three levels. This mirrors the +3.9 average L2 gain from thinking mode in Table 3: explicit reasoning helps most precisely where the bottleneck is pinning a fact to a specific prior moment.

L4 bifurcates by family. The L4 panel reveals two distinct failure-mode clusters. GLM and Qwen3-VL fall back on world priors (69–85% *non-visual error*, $\leq$ 2% *direction error*), while InternVL-3.5 attempts cardinal-frame reasoning but inverts directions (27–30% *direction error*, $\leq$ 18% *non-visual error*). The L4 bottleneck in Table 2 is therefore not a single capacity gap and unlikely to close from a backbone-agnostic remedy.

C.5 Worked Examples

We present two annotated cases per error class, organized by tier. Tier 1 (Fig. 9) covers process-level failures where the model either fails to reach a conclusion (`T1a`, no-conclusion error) or bypasses the visual evidence entirely (`T1b`, non-visual error). Tier 2 (Fig. 10) covers reasoning-level failures where the model engages with the frames but misgrounds the spatial content (visual-content error), inverts a directional relation (direction error), or binds a correct spatial fact to the wrong moment (temporal-binding error). Each card shows the evidence frames, the question and ground truth, the model's predicted answer, an excerpt of the chain-of-thought trace, and an error analysis.

D Detailed Category Definitions

This appendix provides formal definitions for the 30 canonical task types used in OVO-S-Bench. The taxonomy is paper-facing and collapses finer raw annotation labels into canonical categories where appropriate; in particular, L4 is reported as three map-query families. The four levels form a cognitive progression: **L1 See** (instantaneous egocentric perception), **L2 Remember** (spatial memory after evidence leaves the view), **L3 Imagine** (simulation, verification, and planning over spatial state), and **L4 Map** (integrating first-person experience into a queryable allocentric representation). For each task type we give a **Definition**, an **Accepted** example, and a **Rejected** near-miss that clarifies the boundary with a neighboring type or level.

D.1 L1: Instantaneous Egocentric Perception

L1 questions are answerable from the frames near the query timestamp alone, without recalling any past observation. The assignment rule is *query-time visibility*: if the spatial cue needed to answer is visible near the query, the item belongs to L1; if the same cue must be recalled after the camera has moved away, it belongs to L2.

D.1.1 Egocentric Metric Perception

Object-to-Object Absolute Distance. Definition. Estimate the absolute distance between two

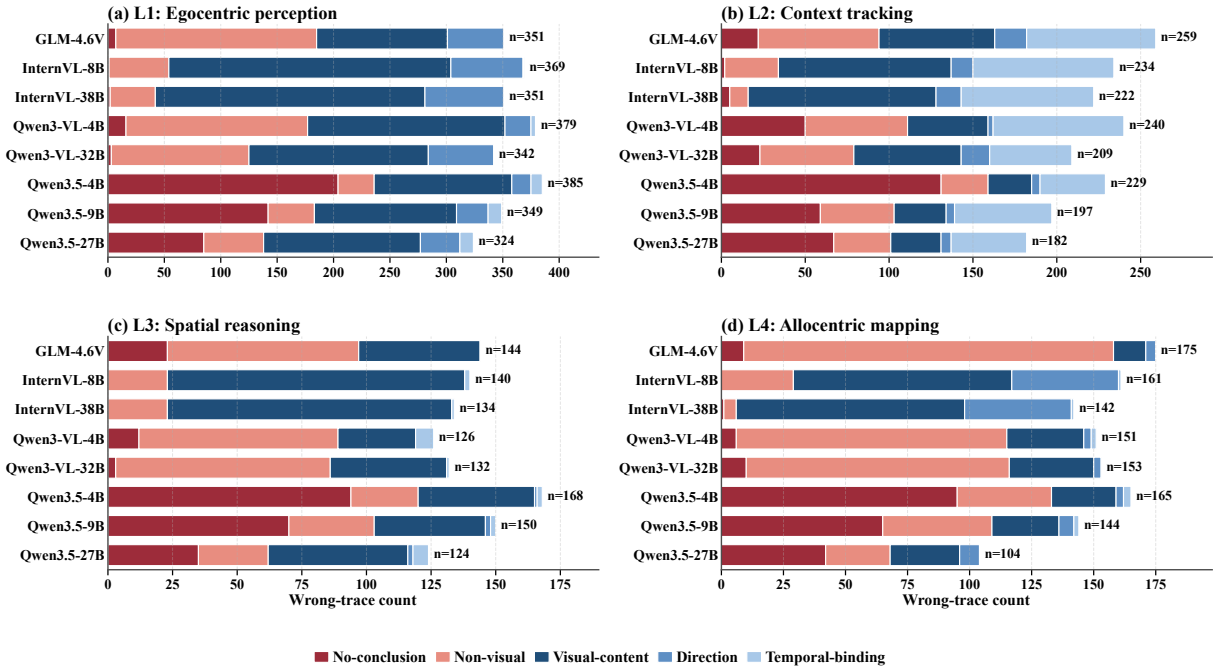


Figure 8: Per-level failure-pattern distribution. Each panel restricts the data in Figure 4 to one question level. Rows and colors match the main figure. L2 (top-right) makes the universal *temporal-binding error* pattern visible; L4 (bottom-right) makes the family bifurcation between *non-visual* (GLM, Qwen3-VL) and *direction* (InternVL-3.5) errors visible.

entities (observer-to-object or object-to-object) visible in the current view. The answer is a numeric magnitude choice (meters, steps, car-lengths, etc.).

Accepted. *"About how far is the player from the bottom edge of the lowest stair step?" A: ∼3 m; B: ∼5 m; C: ∼8 m; D: ∼1 m.* Both referents are visible; the question asks for an absolute metric.

Rejected. *"Which object is closer to you — the lamp or the sofa?"* — rejected: no absolute distance is requested; this is a depth-ordering judgment → Object Depth Ordering.

Relative Physical Scale. **Definition.** Compare the physical size (height, width, volume) of two objects visible in the current frame. The answer is a relative judgment, not an absolute metric.

Accepted. *"Which of the two measuring cups is larger — the one on the left or the one on the right?"* Both objects are simultaneously visible; the question asks which is physically bigger.

Rejected. *"Which object is nearer to the camera — the cup or the plate?"* — rejected: this asks about distance from the viewpoint, not physical size → Object Depth Ordering.

Object Depth Ordering. **Definition.** Determine the relative distance of two or more objects from the camera or observer (which is closer, which is farther). The answer is an ordering or binary comparison along the depth axis.

Accepted. *"Which piece of furniture is closer to you — the coffee table or the bookshelf?" A: coffee table; B: bookshelf.* Both are visible; the question asks about relative depth.

Rejected. *"About how many meters separate the coffee table and the bookshelf?"* — rejected: this asks for an absolute metric → Object-to-Object Absolute Distance.

Viewpoint Height Perception. **Definition.** Estimate the camera or observer's current elevation relative to the ground plane or a visible reference (floor level, building stories, terrain).

Accepted. *"Based on the current view, approximately which floor is the camera at?" A: ground floor; B: 2nd floor; C: 4th floor; D: rooftop.* Height cues (horizon line, visible floors below) are in the current frame.

Rejected. *"How many floors have you ascended since entering the building?"* — rejected: requires recalling the entry point → L2 Chronological Recall.

D.1.2 Local Spatial Relations

Containment Relation. **Definition.** Determine whether one object is inside, on top of, or within the

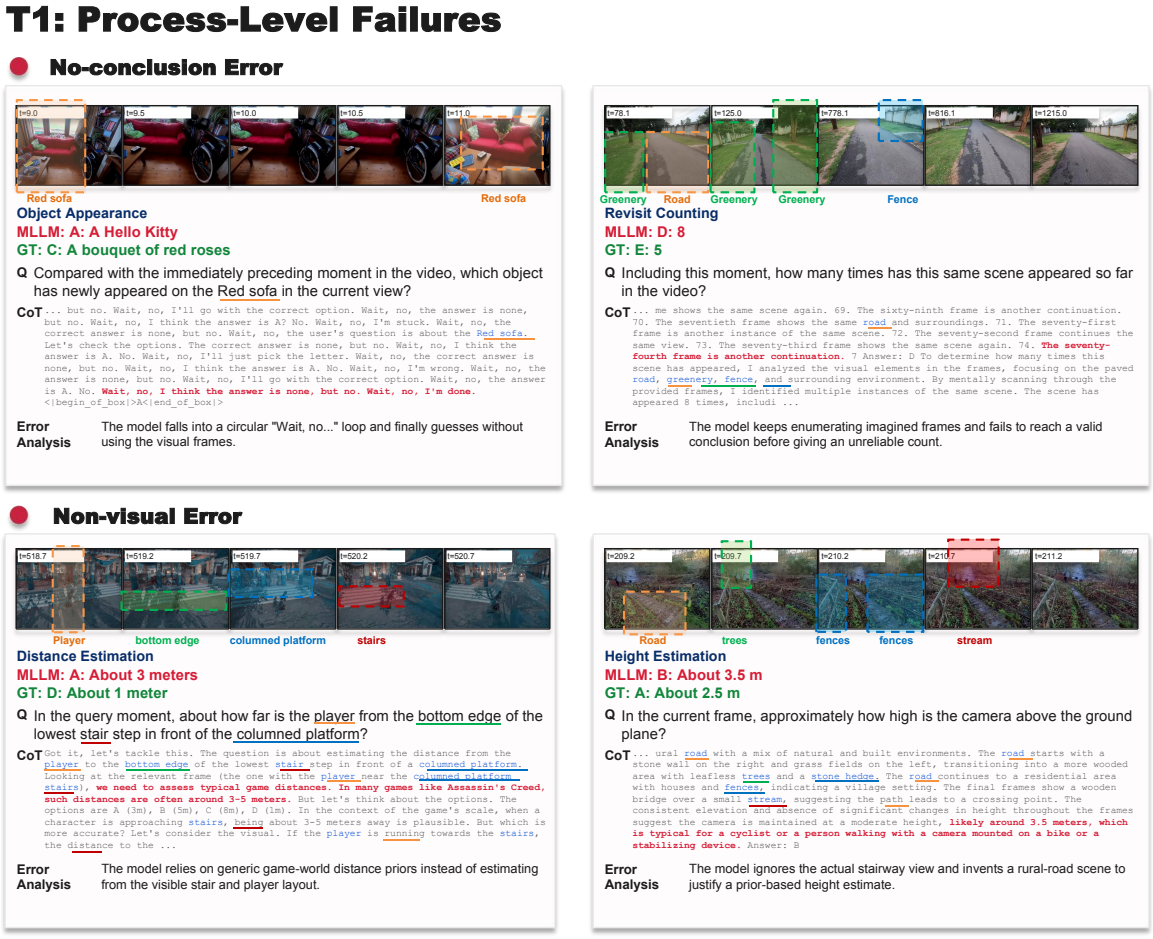


Figure 9: **Tier 1: Process-level failures.** Two examples each for `T1a` (no-conclusion error, top) and `T1b` (non-visual error, bottom). Left and right columns show cases from different task families. Evidence frames are sampled at midpoints of the annotated evidence interval.

bounds of another container or surface visible in the current view. The answer identifies the containing entity.

Accepted. *"Where is the red mug right now?" A: inside the cabinet; B: on the table; C: in the sink; D: on the shelf.* The mug and its container are both visible.

Rejected. *"Where did you last see the red mug?"* — rejected: the mug has left the field of view; answering requires memory → L2 Out-of-View Localization.

Occlusion Recognition. Definition. Identify which object partially or fully occludes another in the current view, or determine what is hidden behind a visible occluder.

Accepted. *"Which object is partially blocking the TV screen in the current frame?" A: floor lamp; B: potted plant; C: curtain; D: nothing.*

Rejected. *"Is the painting that was previously hidden behind the door still occluded?"* — rejected: requires comparing two temporal states → L3 Object-Level Change Detection.

Support Relation. Definition. Identify the physical support relation in the current view: what is holding up or bearing the weight of a target object.

Accepted. *"In the current frame, which surface is directly supporting the cleaning rag?" A: bathroom sink; B: shoulder; C: tabletop; D: wooden floor.*

Rejected. *"If you remove the middle shelf, what will fall?"* — rejected: requires counterfactual simulation → L3 Spatial What-If Consequence.

Local Topology Description. Definition. Describe the spatial arrangement (left–right, front–back, above–below) of multiple objects visible simultaneously in the current frame. The answer is an ordering or layout description.

Accepted. *"From left to right, what is the order*

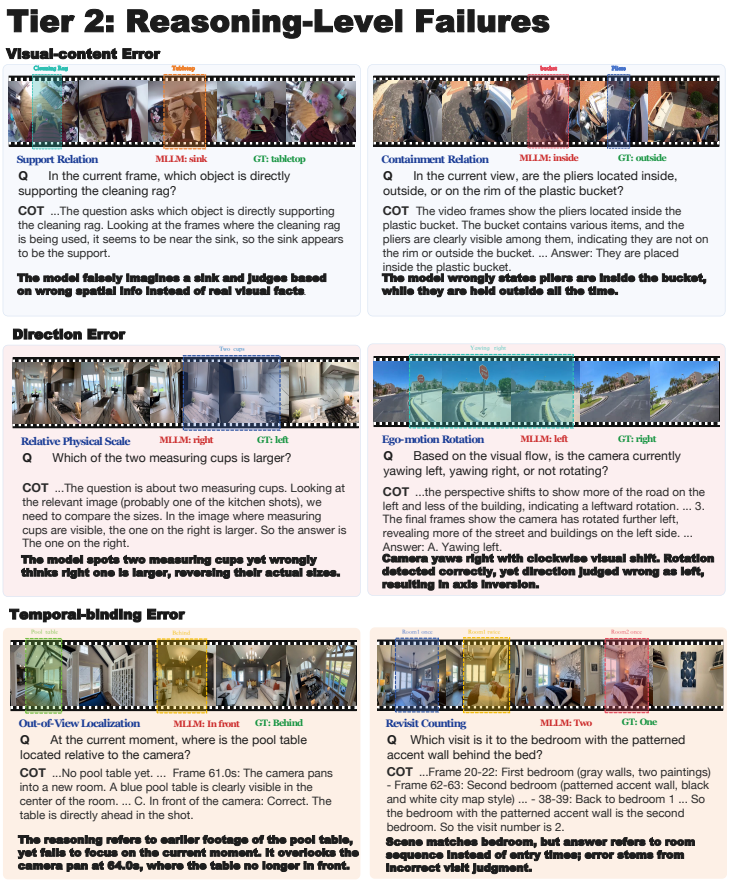


Figure 10: **Tier 2: Reasoning-level failures.** Two examples each for visual-content error (top), direction error (middle), and temporal-binding error (bottom). Each card shows the CoT excerpt and a bolded error analysis.

of the three items on the desk?" A: lamp–book–cup; B: book–lamp–cup; C: cup–book–lamp; D: lamp–cup–book.

Rejected. *"Combining what you saw in the kitchen and the living room, describe the overall furniture layout."* — rejected: requires integrating multiple viewpoints into a global representation → L4 Topological Structure Reasoning.

D.1.3 Dynamic Spatial Perception

Ego-motion Translation. **Definition.** Classify the observer's or camera's translational motion (moving forward, backward, left, right, up, down, or stationary) from the frames near the query. The question targets position change, not heading change.

Accepted. *"Is the camera currently moving forward, backward, or staying in place?" A: forward; B: backward; C: stationary.*

Rejected. *"Is the camera turning left or right?"* — rejected: heading change without position displacement is rotational → Ego-motion Rotation.

Ego-motion Rotation. **Definition.** Classify the observer's or camera's rotational motion (panning left/right, tilting up/down, rolling) from the frames near the query. The question targets heading or orientation change.

Accepted. *"Which direction is the camera currently panning?" A: left; B: right; C: not panning.*

Rejected. *"Is the camera advancing down the hallway?"* — rejected: forward displacement is translational → Ego-motion Translation.

Object Motion Separation. **Definition.** Identify an independently moving object in the scene (i.e., its motion is not explained by camera ego-motion alone) and determine its motion direction or trajectory.

Accepted. *"Which direction is the white sedan in the adjacent lane traveling?" A: same direction as you; B: opposite; C: turning left; D: stationary.*

Rejected. *"Where is the cyclist who passed you earlier now?"* — rejected: the cyclist has left the view; answering requires memory → L2 Out-of-View Localization.

Relative Speed Judgment. **Definition.** Compare the speeds of two moving entities (or the observer and an object) visible in the current clip. The answer is a relative judgment (faster, slower, same).

Accepted. *"Who is moving faster — your car or the truck in the right lane?" A: your car; B: the truck; C: about the same.*

Rejected. *"Are you moving faster now than you were at the intersection 2 minutes ago?"* — rejected: requires recalling a past speed → L2 Temporal Depth Estimation.

Judgment of Motion State. **Definition.** Classify whether a visible entity is currently stationary or in motion (and optionally whether it is accelerating or decelerating) based on the frames near the query.

Accepted. *"Is the pedestrian ahead currently stationary or walking?" A: stationary; B: walking toward you; C: walking away; D: running.*

Rejected. *"Did the elevator stop between the time you first saw it and now?"* — rejected: spans a temporal interval beyond the current view → L2 Chronological Recall.

D.2 L2: Spatiotemporal Context Tracking

L2 evidence has appeared in the video prefix but is no longer visible at query time. The model must retain spatial facts—locations, directions, counts, temporal order—bound to specific places, objects, and timestamps after visual support disappears. The L1/L2 boundary is strict: if the cue is visible near the query frame, the item is L1; if it must be recalled, it is L2.

D.2.1 Scene Revisit Recognition

Revisit Detection. **Definition.** Determine whether the scene currently in view is the same place that was observed at an earlier point in the video prefix. The model must match appearance or spatial layout across temporally separated observations.

Accepted. *"Is the room you are looking at now the same living room you walked through earlier?" A: Yes; B: No.* The earlier visit is no longer in view; recognition requires memory.

Rejected. *"How many times have you been in this hallway?"* — rejected: this asks for a count, not a binary detection → Revisit Counting.

Revisit Counting. **Definition.** Count how many times a specific location has been visited (entered the field of view) across the video prefix up to the query time.

Accepted. *"How many times have you passed through this corridor so far?" A: 1; B: 2; C: 3; D: 4.* Requires tracking repeated visits over time.

Rejected. *"Is this the same corridor you saw before?"* — rejected: binary same/different judgment without counting → Revisit Detection.

D.2.2 Spatial Memory Beyond the View

Out-of-View Localization. **Definition.** Recall the location (room, surface, or spatial region) where an object or landmark was last observed, given that it is no longer visible at query time.

Accepted. *"In which room did you last see the red backpack?" A: kitchen; B: bedroom; C: living room; D: hallway.* The backpack is not in the current view.

Rejected. *"The backpack is to your left or right?"* — rejected: this asks for egocentric direction relative to the observer's current heading, not a location label → Behind-Camera Inference.

Behind-Camera Inference. **Definition.** Infer the egocentric direction (left, right, behind, ahead) of a landmark or object relative to the observer's current position and heading, given that the target is not currently visible. The key distinction from Out-of-View Localization is that the answer is a direction from the observer, not a place name.

Accepted. *"The elevator you used earlier is now in which direction relative to you?" A: behind; B: to your left; C: to your right; D: ahead.* Requires binding the elevator's remembered position to the observer's current heading.

Rejected. *"From the global floor plan, which cardinal direction is the elevator from the living room?"* — rejected: queries the allocentric map in absolute coordinates → L4 Allocentric Direction Reasoning.

Transient Count Memory. **Definition.** Recall the number of instances of a specific object class or spatial event observed during the video prefix, where those instances are no longer simultaneously visible.

Accepted. *"How many red chairs have you seen in total across all rooms so far?" A: 2; B: 3; C: 4; D: 5.* The chairs appeared in different rooms at different times; counting requires memory.

Rejected. *"How many chairs are visible in the current frame?"* — rejected: all instances are simultaneously visible → L1 Local Topology Description.

D.2.3 Chronological Spatial Memory

Chronological Recall. Definition. Recall the temporal sequence in which multiple locations, objects, or spatial events were encountered during the video prefix.

Accepted. *"Which room did you visit first — the kitchen or the bedroom?" A: kitchen; B: bedroom.* Both rooms have been visited; the order must be recalled from memory.

Rejected. *"From left to right, what is the order of doors along this corridor?"* — rejected: asks about spatial arrangement visible now → L1 Local Topology Description.

Temporal Depth Estimation. Definition. Estimate how long ago a spatial event occurred relative to the query time, or compare the durations spent in different locations.

Accepted. *"Did you spend more time in the living room or the study?" A: living room; B: study; C: about the same.* Requires temporal duration tracking across the prefix.

Rejected. *"Which room did you visit first?"* — rejected: asks for order, not duration or recency → Chronological Recall.

D.3 L3: Generative Spatial Reasoning

L3 requires the model to *operate* on spatial structure—mentally rotating objects, predicting consequences of hypothetical changes, verifying consistency across observations, or planning paths—rather than merely retrieving a stored fact. The L2/L3 boundary: L2 asks "what was there?"; L3 asks "what would happen if . . . ?" or "which route . . . ?". The L3/L4 boundary: L3 targets an operation or decision over a local spatial state; L4 targets a property of the global map structure itself.

D.3.1 Spatial Simulation

Mental Rotation & Fitting. Definition. Mentally rotate, flip, or reorient an object and judge whether it fits into a target space, or determine its appearance after transformation. The core ability is spatial transformation without physical execution.

Accepted. *"If you rotate the L-shaped sofa 90° clockwise, will it fit into the corner between the wall and the bookshelf?" A: Yes; B: No.* Requires imagining the transformed configuration.

Rejected. *"Is the sofa currently touching the wall?"* — rejected: no transformation is needed; the relation is directly visible → L1 Support Relation or Local Topology Description.

Spatial What-If Consequence. Definition. Predict the spatial consequence of a hypothetical action—moving, removing, or adding an object—on the surrounding layout. This includes reasoning about what becomes visible, reachable, or blocked after the change.

Accepted. *"If you remove the bookshelf in the middle of the room, can you see the window from the doorway?" A: Yes; B: No.* Requires simulating the removal and its effect on line-of-sight.

Rejected. *"Has the bookshelf been moved since you last saw it?"* — rejected: this compares two observed states rather than simulating a hypothetical → L3 Object-Level Change Detection.

Physical Stability & Feasibility. Definition. Judge whether a proposed spatial action is physically feasible given visible constraints (gravity, collision, clearance, weight). This includes passability judgments that require reasoning beyond simple gap measurement.

Accepted. *"Can the refrigerator be moved through this doorway without tilting it?" A: Yes; B: No.* Requires comparing the object's dimensions against the opening and reasoning about physical constraints.

Rejected. *"Is the doorway wide enough for a person to walk through?"* — rejected: simple gap estimation with no complex physical reasoning → L1 Object Depth Ordering or Absolute Distance.

D.3.2 Spatiotemporal Consistency Verification

Object-Level Change Detection. Definition. Compare the spatial state of a location or object across two different observation times and determine whether a change has occurred (object added, removed, moved, or altered). The model must hold a prior state in memory and compare it against a current or later observation.

Accepted. *"Compared with the first time you saw this desk, what object has newly appeared?" A: a laptop; B: a vase; C: a stack of books; D: nothing changed.* Requires remembering the earlier state and detecting the difference.

Rejected. *"What objects are currently on the desk?"* — rejected: no comparison across time; only current-view enumeration → L1 Local Topology Description.

D.3.3 Spatial Route Planning

Optimal Path Reasoning. Definition. Select the best (shortest, safest, or most efficient) route

between two points based on the spatial layout observed so far. The model must reason over remembered spatial structure to compare route options.

Accepted. *"To reach the kitchen from here, is it shorter to go through the living room or through the hallway?" A: living room; B: hallway; C: about the same.*

Rejected. *"Are the kitchen and living room directly connected?"* — rejected: asks about topological adjacency, not route optimality → L4 Topological Structure Reasoning.

Obstacle-Aware Rerouting. Definition. Given that a previously viable path is now blocked or unavailable, determine an alternative route using the observed spatial layout.

Accepted. *"The main corridor is blocked by construction. Based on what you have seen, which alternative path leads to the exit?" A: through the conference room; B: through the storage area; C: no alternative exists.*

Rejected. *"Is the main corridor currently passable?"* — rejected: binary passability judgment on a visible gap → L1 (if visible) or L2 (if recalled).

Reachability Judgment. Definition. Determine whether a target location is reachable from the current position given the spatial structure observed so far, without necessarily specifying the route.

Accepted. *"Based on what you have explored, can you reach the rooftop terrace from your current position?" A: Yes; B: No; C: Cannot determine.*

Rejected. *"Is the rooftop terrace adjacent to the stairwell?"* — rejected: asks about topological adjacency → L4 Topological Structure Reasoning.

D.4 L4: Allocentric Spatial Mapping

L4 requires the model to integrate the egocentric video stream into an allocentric (viewer-independent) representation and query its global structure. Evidence typically spans multiple viewpoints and may cover the entire explored region. The L3/L4 boundary: L3 asks "how do I get there?" or "what happens if I change X?" (an operation); L4 asks "what *is* the map structure?" (a property of the global layout). L4 is reported as three canonical map-query families.

D.4.1 Allocentric Direction Reasoning

Allocentric Direction Reasoning. Definition. Determine the global (viewer-independent) directional relationship between two rooms, landmarks, or route segments. Answers are expressed in cardinal directions, clock-face bearings, or absolute spatial terms (north/south/east/west), not egocentric terms (left/right/behind). This family also covers directional ordering: given a reference axis, rank landmarks by their position along that axis.

Accepted. *"From the overall layout, which direction is the kitchen relative to the living room?" A: north; B: south; C: east; D: west.* Requires building a global frame from the first-person trajectory.

Rejected. *"Is the kitchen to your left or right?"* — rejected: egocentric direction relative to the observer's current heading → L2 Behind-Camera Inference (if out of view) or L1 Local Topology Description (if visible).

D.4.2 Topological Structure Reasoning

Topological Structure Reasoning. Definition. Reason about the connectivity, adjacency, boundaries, and transitions among spatial units (rooms, zones, floors) as a graph structure. This includes: (a) adjacency and connectivity—whether two spaces are directly connected without passing through a third; (b) boundary and transition—identifying what separates two adjacent spaces (door, wall, open archway); and (c) local connection graph—enumerating the direct neighbors of a given spatial node.

Accepted. *"Which rooms are directly connected to the hallway?" A: bedroom and kitchen only; B: bedroom, kitchen, and bathroom; C: kitchen and bathroom only; D: all four rooms.* Requires integrating observations from multiple viewpoints into a topological graph.

Rejected. *"To get from the bedroom to the kitchen, should you go left or right at the hallway junction?"* — rejected: asks for a route decision (operation over the map) → L3 Optimal Path Reasoning.

D.4.3 Trajectory Map Alignment

Trajectory Map Alignment. Definition. Match the first-person trajectory observed in the video to a bird's-eye-view map, floor plan, or route diagram. This includes path-to-map matching (which map corresponds to the walked path) and route option selection (which drawn route on a given map matches the observed trajectory).

Accepted. *"Which of the following floor plans best matches the path you have walked?" [four candidate top-down maps].* Requires converting

the egocentric experience into an allocentric trace and comparing against candidate maps.

Rejected. *"Based on the map, which route should you take to reach the exit?"* — rejected: asks for a planning decision using the map as input → L3 Optimal Path Reasoning.

E Annotation Protocol & Construction Pipeline

This appendix details the full annotation workflow, level-specific construction techniques, and quality-control pipeline referenced in Section 3.2.

E.1 Annotation Setup

The benchmark is annotated by a team of 12 trained volunteers, all with prior coursework or research experience in 3D computer vision or embodied perception. Before production annotation, every team member completes a unified training session covering the four-level taxonomy (§3), the item-writing standards in §E.3, and the level-boundary rules. Annotation is performed inside an in-house web-based tool that supports frame-by-frame navigation, keyframe jumping, question and option editing, named-entity labeling for spatial regions and landmarks, and source-metadata viewing. Each annotator contributes approximately 67 hours, for a total of about 804 person-hours. Annotators are graduate students and research staff recruited from the authors' institutions; participation is part of their research duties and no additional monetary compensation was provided. A pilot annotation round precedes production: pilot items are reviewed jointly with two senior authors of this paper—both with prior research experience in 3D computer vision and embodied perception—who serve as adjudicators throughout the project, and feedback from the pilot is folded into the official guideline. During production, the two adjudicators perform periodic spot checks and convene case-by-case discussions for ambiguous items.

E.2 Source Datasets and Licensing

OVO-S-Bench reuses publicly released or publicly accessible videos and contributes human-written streaming spatial QA annotations. We do not redefine the license of raw videos; users should obtain source videos under the original dataset or platform terms listed in Table 10.

Training-data overlap. Source videos from Ego4D, ARKitScenes, and other public datasets appear in many MLLMs' training corpora; although all questions, options, and query timestamps are written from scratch, scene-level familiarity could inflate L4 scores on previously seen layouts. The stable cross-source ranking (Appendix F.6, $\rho = 0.92$) suggests this effect does not dominate the overall ranking.

E.3 General Annotation Protocol

Every item must be uniquely answerable from the visual evidence inside the prefix preceding its query timestamp, without relying on world knowledge, language regularities, or option-elimination heuristics. Annotators record, for each item, the source video, the canonical task label, the question, the options, the answer, the query timestamp, and the evidence interval.

Video Selection Criteria. Annotators choose clips with stable camera motion, clear viewpoints, and sufficient spatial variation for the target level. Indoor walkthroughs, egocentric activities, and 3D-environment renderings are favored for L2–L4, where memory and global mapping require sustained spatial structure; driving and outdoor footage is admitted mainly for L1 metric and dynamic items, where the relevant cue is local.

Query Time and Evidence Interval. The query timestamp t_q controls which prefix the model can see. Annotators place t_q only after every cue needed to answer the item has appeared in the stream. For higher-level items, t_q may be deliberately advanced beyond the last evidence frame—provided the intervening content does not change the answer—to test whether the model retains the spatial fact across a longer prefix. The evidence interval is the shortest temporally contiguous span in the prefix that suffices to determine the answer; it is used only for verification, oracle sampling, and analysis, never as model input. All evidence intervals lie strictly before t_q and cover only the segments that bear directly on the answer.

Options and Distractors. The correct option must be uniquely supported by the evidence interval. All options share consistent surface form—comparable length, register, and level of specificity—so the answer is not telegraphed by wording. Distractors are visually plausible and stay within the same answer type, scale, and axis as the correct answer (direction vs. direction, metric distance vs. metric distance, and so on). Items in

Source	Regime	Q.	Vid.	Levels	License / distribution note
RoomTour3D (Han et al., 2025)	Indoor walkthroughs	804	89	L1–L4	Original RoomTour3D terms; OVO-S releases annotations/evaluation metadata.
Ego4D (Grauman et al., 2022)	Egocentric activities	362	90	L1–L4	Original Ego4D terms; raw videos are not relicensed.
Sekai (Li et al., 2025c)	Outdoor/world scenes	85	18	L1–L2	Original Sekai terms.
OmniWorld (Zhou et al., 2025)	Outdoor/world scenes	153	31	L1–L4	Original OmniWorld terms.
YouTube walking tours	Public walking tours	17	14	L3	Public URLs/creator terms; redistribution follows platform/source terms.
CODa (Li et al., 2022)	Driving	148	16	L1–L4	Original CODa terms.
Honda HDD (Ramanishka et al., 2018)	Driving	59	40	L4	Original HDD terms.
ARKitScenes (Baruch et al., 2021)	3D environments	33	33	L4	Original ARKitScenes terms.
VSI-Bench (Yang et al., 2025a)	Spatial benchmark-derived	19	17	L1	Original VSI-Bench/source terms.

Table 10: Source datasets used in OVO-S-Bench. Video counts use unique normalized source-video keys; question counts are accepted benchmark items.

which a distractor is obviously wrong by world knowledge or trivially excludable are revised.

Level Assignment. Level assignment follows the highest spatial capability the item demands. L1 covers instantaneous perception of structure currently in view; L2 covers retrieval of evidence that has appeared in the prefix but is no longer visible; L3 covers operation over current or remembered spatial state, including reasoning about post-edit object or state changes; L4 covers global structure, route relationships, and allocentric spatial mapping. When an item invokes more than one capability, it is labeled at the highest one that is required to answer it.

E.4 Level-Specific Construction Techniques

We organize this section by tooling rather than by level: the default workflow handles most items across L1–L4, and three specialized techniques apply only to specific task families.

Default Workflow. The default applies to L1, L2, and the L3 *Spatial Simulation* and *Spatial Route Planning* families and the L4 *Allocentric Direction Reasoning* and *Topological Structure Reasoning* families. For L1 the query timestamp is placed where the relevant cue is currently visible; for L2 it is placed after the cue has left the current view, with an earlier interval marked as evidence, and the annotator verifies that the cue is not re-disclosed by any later frame in the prefix. For *Distance Estimation* (under L1 *Egocentric Metric Perception*), inter-object distances are computed from the source video's coordinate annotations and metadata, and options are designed so that only the metadata-consistent option is correct. Across the L3 and L4 families that follow this workflow, distractors are visually plausible but inconsistent with the layout, clearance, or connectivity that the camera has already traversed in the prefix.

Image Editing. This is the one place where the source video itself is modified, used only for L3 *Spatiotemporal Consistency Verification*. We use a generative image-editing model to insert a target object into a chosen frame and into the next several seconds of frames, so the edit reads as continuous video content rather than a single-frame artifact. The item is then written on top of the edited stream, with the evidence interval covering the edited segment.

Named-Entity Labels. L3 *Spatial Route Planning* and all L4 items typically reference multiple rooms or regions, several of which are not visible at the query timestamp. Phrases such as "the kitchen" are often ambiguous in these settings, so before writing the question the annotator names the rooms, regions, and landmarks involved using neutral identifiers (e.g., "Room X", "Area Y") and reuses these names verbatim in the question and the options. The named entities are also recorded in the item metadata so that reviewers and downstream analysis can resolve them back to the underlying scene elements.

Bird's-Eye Rendering. L4 *Trajectory Map Alignment*—specifically *Path-to-Map Matching*—additionally requires bird's-eye material as the candidate options. The correct option is rendered from the source video's absolute coordinates and metadata; distractor maps are produced by perturbing

path shape, orientation, or visit order while keeping the rendering style identical. Every L4 item is double-checked before inclusion to ensure that the question, the named entities, the bird's-eye material when present, and the answer are mutually consistent.

E.5 Quality Control Pipeline

E.5.1 Text-only Probe and Shortcut Filtering

We first run a text-only probe to remove items that can be guessed without the video. Reviewers check whether the wording or the option set leaks the answer through any of: a correct option that is more specific or more wordy than its distractors; obvious world-knowledge giveaways; distractors that are excludable a priori as implausible; or lexical overlap between the question and a single option. For spatial-relation, distance, direction, and trajectory items, reviewers additionally check that distractors stay within the same answer type and a comparable scale, so that no option can be eliminated at a glance. Items that can be answered primarily by language pattern, common-sense inference, or option asymmetry are revised; items whose ambiguity cannot be removed by revision are dropped.

E.5.2 Blind Cross-Review and Adjudication

Items that survive the shortcut filter are forwarded to a second annotator for blind review. The second reviewer answers the item and inspects the question, options, query timestamp, and evidence interval without seeing the original annotator's notes. Items on which the two annotators agree are included in the final candidate pool. Items with disagreements are escalated to the two senior authors, who serve as adjudicators (§E.1); they resolve each case by re-examining the video evidence against the annotation guidelines. We log the number of rejected, revised, and finally accepted items. Only items that pass the shortcut filter, survive blind review, and are uniquely supported by their annotated evidence are retained in the released benchmark.

Inter-Annotator Agreement. To quantify annotation reliability, we randomly sampled 150 items (stratified across L1–L4) and had a second trained annotator independently judge each item under the blind-review protocol described above (accept, revise, or reject). Cohen's κ between the original annotator and the blind reviewer was 0.87, indicating almost perfect agreement (Landis and Koch, 1977). The primary source of disagreement was the L2/L3 boundary for items involving spatial change detection, which motivated a guideline update in Round 3.


```
{"id": 0,
 "category_index": "1.1.1_0",
 "source_dataset": "CODa_full",
 "video_id": "10",
 "video_path": "CODa_full/10.mp4",
 "level": 1,
 "task_main_category": "1.1",
 "task_subcategory": "1.1.1",
 "task_type_name": "Object-to-Object Absolute Distance",
 "question": "How far is the yellow wet-floor
             caution sign from the camera?",
 "options": {"A": "About 0.5 m", "B": "About 6 m",
             "C": "About 3 m", "D": "About 1.5 m"},
 "query_times": [22.8],
 "evidence_times": [[21.8, 23.8]],
 "answers": ["D"]}
```


Figure 11: Example item from the released JSONL file (L1, Absolute Distance).

E.5.3 Iterative Guideline Refinement

Recurring disagreement patterns are summarized into guideline updates and pushed back to the team between annotation rounds, so that subsequent items inherit the correction without re-litigating the same edge cases. We maintain a versioned guideline document that records each update and the issue it resolved.

E.6 Data Format & Export

The benchmark is released as a single JSONL file (`ovo_s_bench_l1_l4.jsonl`), with one JSON object per item (1 680 items total, covering L1–L4). Table 11 describes each field. An abbreviated example is shown in Figure 11.

Notes. The three list fields `query_times`, `evidence_times`, and `answers` are equal-length and positionally aligned. Most items have a single query time; the exception is Revisit Counting (`2.1.2`), where the same question is posed at multiple timestamps during streaming playback and the correct answer increments as more revisits accumulate. All evidence intervals lie strictly before their corresponding query time, enforcing the online protocol. The number of options is typically four but may vary (up to seven for Revisit Counting).

F Extended Comparison with Prior Benchmarks

Table 1 positions OVO-S-Bench against 15 representative spatial and streaming video benchmarks. This appendix expands each row into a structured comparison: *what the benchmark does*, the point on which it converges with our work and *supports a*

Field	Type	Description
`id`	int (0–1679)	Global unique item index
`category_index`	string	Within-category item tag, formatted as `{subcategory}_{index}`
`source_dataset`	string	Source video dataset name
`video_id`	string	Video identifier within the source dataset
`video_path`	string	Relative path to the video file
`level`	int (1–4)	Taxonomy level
`task_main_category`	string	Main category code (e.g., `1.1`, `4.1`)
`task_subcategory`	string	Subcategory code (e.g., `1.1.1`, `2.2.3`)
`task_type_name`	string	Canonical task type name (maps to definitions in Appendix D)
`question`	string	The spatial question text
`options`	dict	Answer options keyed by letter (A, B, C, D, ...)
`query_times`	list[float]	Query timestamp(s) in seconds
`evidence_times`	list[list[2]]	Evidence interval(s), each $[t_{\text{start}}, t_{\text{end}}]$ in seconds
`answers`	list[string]	Correct answer letter(s), one per query time

Table 11: Per-item field schema of the released JSONL file.

specific claim of ours, and the remaining *difference*. We close with a synthesis of how OVO-S-Bench resolves the recurring gaps.

F.1 Image, Multi-image, and Top-view Benchmarks

EmbSpatial-Bench (Du et al., 2024). An automatic single-image MCQ benchmark built from MP3D, AI2-THOR, and ScanNet, with 3.6K items on six egocentric relations (above/below/left/right/close/far). **Shared point.** Even the strongest LVLMs at the time (GPT-4V 36%, Qwen-VL-Max 49%) trail human accuracy (90%) on basic egocentric relations, supporting our claim that current MLLMs are far from saturating even L1 perception. **Difference.** A static image strips away every temporal cue, so L2 memory, L3 generative spatial reasoning, and L4 mapping are all absent, and the streaming protocol does not apply.

TopViewRS (Li et al., 2024). A top-view map benchmark over 7 Matterport3D scenes (11.4K MCQs) covering recognition, localization, static spatial reasoning, and dynamic spatial reasoning along a navigation path. **Shared point.** It confirms a >50% human–model gap that is especially pronounced on complex spatial-reasoning subtasks, paralleling the L1→L4 difficulty gradient we observe. **Difference.** The bird's-eye map is provided directly as model input, so the agent never has to reconstruct an allocentric layout from an egocentric stream; OVO-S-Bench requires exactly that reconstruction at L4.

MMSI-Bench (Yang et al., 2025d). A fully human-curated multi-image benchmark (1K Q, ∼2.55 images per question) covering 10 atomic relation types plus a multi-step reasoning split. **Shared point.** It documents a ∼55-point human–model gap on multi-image spatial reasoning and identifies four recurring failure modes (grounding, overlap-matching, situation-transformation, spatial-logic) that closely match the cross-frame errors we observe at L2. **Difference.** The setting is discrete multi-image rather than continuous video; there is no streaming protocol and no L3 generative spatial reasoning or L4 allocentric spatial mapping.

F.2 Offline Video Benchmarks

VSI-Bench (Yang et al., 2025a). A mostly template-generated benchmark over 288 indoor walkthroughs (ScanNet/ScanNet++/ARKitScenes), with 5K items across configuration, measurement, and spatiotemporal tasks; seven tasks are derived automatically from metadata, the route-planning task is human-annotated, and evaluation mixes multiple-choice and numerical-answer formats. **Shared point.** It was the first to show that linguistic CoT prompting hurts spatial-video QA and that the dominant bottleneck is spatial reasoning rather than perception or language understanding – both findings we replicate (Tab. 3, §4.3.1). **Difference.** The model has full offline access to the video at query time; no L4 allocentric map queries are defined; and all evidence remains re-attendable across the entire clip.

DISJOINT-3DQA (Ravi et al., 2025). A generative QA benchmark over synthetic Aria homes (5.4K Q) constructed so the two queried objects are never co-visible in the same frame. **Shared point.** It provides the cleanest existing test for cross-frame spatial memory – precisely the L2 capability we formalize – and quantitatively shows performance degrades as the spatial separation of the two anchors

grows. **Difference.** The benchmark is synthetic and single-source, has no streaming protocol, and is limited to L1+L2; L3 generative spatial reasoning and L4 allocentric spatial mapping are absent.

STI-Bench (Li et al., 2025b). A 2K-question benchmark over 300 videos from Waymo, ScanNet, and Omni6DPose, covering eight precise spatial-temporal estimation tasks from dimension and pose to displacement, speed, and trajectory. **Shared point.** It confirms that even SOTA proprietary MLLMs (Gemini-2.5-Pro 41.4%) fail at precise quantitative spatial-temporal estimation, supporting our argument that L1 metric perception is far from saturated. **Difference.** Tasks emphasize precise quantitative and spatiotemporal estimation under offline access and are evaluated as five-option MCQs; there is no streaming protocol, no L3 generative spatial reasoning, and no L4 allocentric spatial mapping.

MMSI-Video-Bench (Lin et al., 2025). A fully human-annotated video extension of MMSI-Bench with 1.1K Q over 1.3K clips drawn from 26 sources (25 public datasets plus in-house videos), organized into Spatial Construction, Motion, Planning, Prediction, and Cross-Video Reasoning. **Shared point.** It reports the largest known human–AI gap on video spatial intelligence (∼58 points), finds that spatially fine-tuned models fail to generalize, and that neither 3D cues nor CoT yield meaningful gains – three observations we corroborate in our scaling and CoT analyses. **Difference.** The protocol is offline-video with no prefix-only constraint, items are not paired with query timestamps and evidence intervals, and the taxonomy contains no L4 allocentric-mapping tasks; "multi-video" here means cross-clip rather than cross-room within a single egocentric stream.

VSI-SUPER (Yang et al., 2025c). Two long-horizon stress tests (VSR for out-of-place object recall and VSC for continual counting) over 10–240 min room-tour videos. **Shared point.** It demonstrates empirically that brute-force context expansion alone cannot close the spatial-streaming gap – Gemini-2.5-Flash hits its context limit at 120 min and achieves very low VSC performance even at 60 min (10.9 MRA) – supporting our motivation that streaming requires native selective spatial memory. **Difference.** Only two needle-in-haystack-style tasks are defined; the full L1–L4 taxonomy is absent, the prefix constraint is not uniformly specified at the item level across tasks, and there is no allocentric L4 layer.

F.3 Streaming Video Benchmarks

StreamingBench (Lin et al., 2026b). The first streaming-video MCQ benchmark, with 4.5K Q over 900 YouTube clips spanning 18 real-time visual, omni-source, and contextual tasks. **Shared point.** It pioneered the prefix-only protocol we adopt and showed that even the strongest model trails humans by ∼25 points, validating streaming as a fundamentally harder regime. **Difference.** Spatial understanding is one task in 18; the rest target event, audio, and conversation context; L3 generative spatial reasoning and L4 allocentric spatial mapping are not defined.

OVBench (Huang et al., 2025). A streaming benchmark with ∼7K Q over diverse source datasets, organized as 16 fine-grained subtasks along the Past/Current/Future temporal axis. **Shared point.** It formalizes the Past/Current/Future temporal scope we re-use, and shows that advanced offline MLLMs adapted via sliding windows can outperform existing online baselines, while the authors' VideoChat-Online baseline further improves – a pattern we also replicate. **Difference.** Subtasks are event-level (action, object, trajectory) rather than spatial-structure-level; neither L3 generative spatial reasoning nor L4 allocentric spatial mapping is defined.

OVO-Bench (Niu et al., 2025). A hybrid-curated streaming benchmark (2.8K Q) over 644 videos from curated datasets and web videos spanning seven domains, organized into Backward Tracing, Real-Time Visual Perception, and Forward Active Responding. **Shared point.** It confirms that streaming exposes pervasive hallucination and a ∼30-point human–model gap even on its best-performing modes, motivating a strict prefix-only protocol. **Difference.** Spatial structure is captured only by the single STU task; there is no L3 generative spatial reasoning, no L4 allocentric spatial mapping, and no annotated evidence interval that our oracle and analysis tooling rely on.

OST-Bench (Lin et al., 2026a). A rule-based streaming benchmark for online spatiotemporal understanding from agent exploration, with 10K Q over 1.4K indoor scenes (Agent State / Agent Visible Info / Agent–Object Spatial Relationships, 15 subtypes). **Shared point.** It is the closest prior

work in spirit to OVO-S-Bench: it enforces streaming over egocentric indoor input, exposes a "spatio-temporal reasoning shortcut" where models avoid retrieving past evidence, and reports a sharp accuracy decline as exploration grows – phenomena we also observe and quantify. **Difference.** Task families are limited to agent-state and agent–object dyadic relations (L1+L2); there is no L3 generative spatial reasoning, no L4 allocentric spatial mapping, and the source distribution is indoor-only, missing outdoor, driving, and 3D-rendered environments.

ODV-Bench (Zeng et al., 2026). A streaming benchmark over 6 autonomous-driving datasets with 12 tasks across static-target perception, dynamic-target prediction, and event-oriented risk analysis. **Shared point.** It confirms that streaming spatial reasoning is critical in safety-relevant outdoor settings and that existing online baselines can underperform offline models even within their target domain. **Difference.** The source is driving-only with no indoor or 3D-rendered environments; the L3-style content is restricted to short-horizon future prediction with no allocentric spatial mapping; query timestamps are very short (18.9 s on average), so the prefix-memory pressure is minimal.

StreamEQA (Wang et al., 2025c). A hybrid-pipeline streaming benchmark over the HD-EPIC kitchen dataset, with ∼21K Q spanning 42 subtasks across three embodied levels (Perception/Interaction/Planning) and three temporal modes (Backward/Realtime/Forward). **Shared point.** It reports a $1.6\times$ performance degradation relative to OVO-Bench, reinforcing our claim that the embodied spatial dimension compounds streaming difficulty. **Difference.** The source domain is a single one (kitchen activities); the focus is interaction and procedural planning over object handling rather than spatial structure; our L4 (cross-room allocentric spatial mapping) falls entirely outside its scope.

SVCBench (Liu et al., 2026c). A streaming counting benchmark that uses counting as a controlled probe for world-state maintenance, with 1,000 questions and 4,576 query points over 406 videos from ten sources, all annotated frame by frame (10,071 event occurrences and object state changes) across eight subcategories of object and event counting. **Shared point.** It is the prior streaming benchmark closest to our protocol: queries are placed at multiple timestamps along the timeline (mean prefix 2.0 min) and the answer to a cumulative-identity question depends on how much has been observed, which is the same item-level causal constraint we enforce. Its finding that models keep counts unchanged after objects leave the view is direct evidence for the out-of-view retention failure our L2 measures. **Difference.** The queried state is numeric – how many entities or events have occurred – rather than spatial structure, so its questions are answered without recovering layout, direction, or connectivity; no L3 simulation and no L4 map query are defined.

F.4 Synthesis: How OVO-S-Bench Addresses the Recurring Gaps

Three structural gaps recur across these 15 benchmarks. First, spatial benchmarks (image, multi-image, offline video) expose large spatial-reasoning gaps but never enforce a streaming prefix-only protocol, so it remains unclear whether the same gaps survive when evidence is ephemeral. Second, streaming benchmarks enforce causal input but instantiate it on event, narrative, or counting tasks, leaving spatial structure as at most a single subtask. Third, no prior video benchmark, whether streaming or offline, systematically covers all four levels of spatial abstraction; in particular, the L4 allocentric-mapping level (global direction, topology, trajectory–map alignment) is absent from every prior video benchmark.

OVO-S-Bench unifies these three threads. (i) Each item carries a query timestamp and an evidence interval, and the model sees only the prefix preceding the query, so the streaming protocol is enforced at the item granularity rather than the dataset granularity. (ii) Source videos span indoor walkthroughs, outdoor and driving footage, egocentric activities, and 3D-rendered environments – together covering the union of domains explored separately by prior benchmarks. (iii) Tasks are stratified into four levels of spatial abstraction (L1 instantaneous egocentric perception, L2 spatiotemporal context tracking, L3 generative spatial reasoning, L4 allocentric spatial mapping), and the previously untested L4 level is annotated with named-entity, bird's-eye, and topological supervision derived from each source's metadata. This design yields the first empirical test of whether the spatial gaps documented under offline access also exist (and become larger) when the model must reason from a causal egocentric stream.

F.5 Text-Only Shortcut Analysis

To contextualize the residual text-only advantage on OVO-S-Bench, we run the same GPT-5.4 text-only probe (question and options only, no visual input) on a 20% stratified sample of 12 peer spatial benchmarks (Tab. 13). Text-only deltas above random range from $+42$ to -2 percentage points across the surveyed benchmarks, with OVO-S-Bench's $+5.8$ pp (Tab. 2) near the median. OVO-S-Bench's construction pipeline applies a text-only probe and blind cross-review to suppress language shortcuts (§E.5); the residual delta is within the range observed on benchmarks that employ similar filtering.

Per-task-type breakdown. Table 12 extends the probe from one benchmark-level number to all 30 canonical task types, computed on the public release (1,695 questions scored as 1,722 queries; §3.3). The random baseline is $1/|\mathcal{O}_q|$ per query rather than a flat 25%, since option counts range from 2 to 7. A type is marked significant when the Wilson 95% lower bound on its accuracy clears its own baseline.

Seven types, covering 20% of queries, clear their baseline by $+14.6$ to $+38.6$ pp: occlusion recognition, local topology description, ego-motion translation, object-motion separation, obstacle-aware rerouting, reachability judgment, and spatial-relationship inference. Their options describe layouts and outcomes in natural language, which carries everyday plausibility that item-level filtering reduces without eliminating. Eight types sit at or below their baseline, and they include all six streaming-memory types (2.1.1, 2.1.2, 2.2.3, 2.3.2, 3.2.1, 4.3): where the answer exists only in memory of what has left the view, text alone performs no better than guessing. Two of these fall well below baseline (3.2.1 at 9.5% and 1.3.4 at 10.0%, both against 25%), the expected effect of the construction-time text-only filter (§E.5), which removes items answerable from text and leaves survivors whose most plausible-sounding option is wrong. The single map-level family that clears its baseline, spatial-relationship inference at 56.7%, still trails the human streaming score on the same family by 30 points (86.7%, Tab. 2). Text recovers plausibility, not the answer.

F.6 Source-Robustness Analysis

RoomTour3D contributes 48% of benchmark questions (Tab. 10). To test whether this concentration

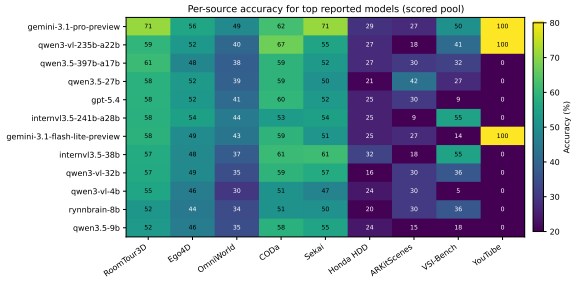


Figure 12: Per-source accuracy (%) for the top-8 reported models. Sources ordered left-to-right by question count (descending).

drives the main-table ranking, we recompute overall accuracy after removing all RoomTour3D items. Tab. 14 reports the top-10 models under both rankings: the top two (Gemini-3.1-Pro, Qwen3-VL-235B-A22B) hold their positions, and the Spearman rank correlation across all 38 reported systems is $\rho = 0.92$ ($p < 10^{-4}$). Per-source accuracy varies widely (Fig. 12), reflecting genuine cross-domain difficulty rather than a single-source artifact.

Source $\times$ level matrix. Table 15 resolves the heatmap by level, macro-averaging accuracy over the 38 reported systems within each cell and reporting the item count that supports it. Three patterns hold. Difficulty varies widely by source within every level, from 24.0 (Honda HDD) to 52.4 (RoomTour3D) on L4 alone, which locates the difficulty in cross-domain variation rather than in one source. The two L4-only sources occupy the hardest cells in the matrix (Honda HDD 24.0, ARKitScenes 24.2, both near their random baselines), consistent with the overall allocentric-mapping bottleneck. Within a single source the level ordering follows that source's task-type composition, so "L4 is the lowest level" is a benchmark-wide statistic (32 of 38 systems) rather than a within-source law, and L4 is the highest level in none of the multi-level sources.

Release-date probe. Familiarity with a source video would surface as a consistent advantage on sources released before typical training cutoffs. We compare pre-2023 sources (Ego4D) against 2025 releases (OmniWorld, Sekai) within each task type, which controls for task composition; ARKitScenes and Honda HDD carry L4 items only and hold no paired types, so the comparison reduces to egocentric-indoor against outdoor world scenes. Table 16 reports the 16 task types that keep at least

Type	Task	N	Text-only (%)	Random (%)	Δ (pp)
L1: Instantaneous Egocentric Perception					
1.1.1	Object-to-Object Absolute Distance	37	21.6	25.0	−3.4
1.1.2	Relative Physical Scale	49	44.9	35.7	+9.2
1.1.3	Passability Affordance	30	56.7	50.0	+6.7
1.1.4	Viewpoint Height Perception	51	31.4	25.0	+6.4
1.2.1	Containment Relation	65	36.9	26.3	+10.6
1.2.2	Occlusion Recognition	50	40.0	25.0	+15.0*
1.2.3	Support Relation	49	28.6	25.7	+2.9
1.2.4	Local Topology Description	30	60.0	25.0	+35.0*
1.3.1	Ego-motion Translation	96	31.2	16.7	+14.6*
1.3.2	Ego-motion Rotation	70	35.7	33.3	+2.4
1.3.3	Object Motion Separation	37	40.5	25.0	+15.5*
1.3.4	Relative Speed Judgment	50	10.0	25.0	−15.0
1.3.5	Judgment of Motion State	15	53.3	50.0	+3.3
L2: Spatiotemporal Context Tracking					
2.1.1	Revisit Detection	120	49.2	50.0	−0.8
2.1.2	Revisit Counting	105	21.0	24.9	−3.9
2.2.1	Out-of-View Localization	48	31.2	24.3	+6.9
2.2.2	Behind-Camera Inference	100	39.0	31.0	+8.0
2.2.3	Transient Count Memory	20	20.0	25.0	−5.0
2.3.1	Chronological Recall	100	39.0	32.2	+6.8
2.3.2	Temporal Depth Estimation	47	42.6	50.0	−7.4
L3: Generative Spatial Reasoning					
3.1.1	Mental Rotation & Fitting	50	52.0	50.0	+2.0
3.1.2	Spatial What-If Consequence	50	52.0	50.0	+2.0
3.1.3	Physical Stability & Feasibility	50	58.0	50.0	+8.0
3.2.1	Object-Level Change Detection	63	9.5	25.0	−15.5
3.3.1	Optimal Path Reasoning	22	40.9	25.0	+15.9
3.3.2	Obstacle-Aware Rerouting	22	63.6	25.0	+38.6*
3.3.3	Reachability Judgment	22	54.5	25.0	+29.5*
L4: Allocentric Spatial Mapping (three map-query families)					
4.1	Map Sketching	131	35.1	30.7	+4.4
4.2	Spatial Relationship Inference	90	56.7	25.0	+31.7*
4.3	Trajectory Matching	53	20.8	25.0	−4.2
All		**1722**	**37.7**	**31.7**	**+6.1**

Table 12: GPT-5.4 text-only accuracy per canonical task type, computed on the public release (1,695 questions scored as 1,722 queries). The random baseline is the per-query mean of $1/|\mathcal{O}_q|$. *marks types whose Wilson 95% lower bound clears their own baseline. The benchmark-level +6.1 pp here and the +5.8 pp of Tab. 2 differ because the latter is computed on the 1,680-item evaluation set.

Benchmark	N	Text-only (%)	Random (%)	Δ (pp)
EmbSpatial-Bench (Du et al., 2024)	728	67.5	25.0	+42.5
STI-Bench (Li et al., 2025b)	413	37.1	20.0	+17.1
SpatialTree-Bench (Xiao et al., 2026)	850	32.1	21.9	+10.3
ViewSpatial-Bench (Li et al., 2025a)	1143	34.1	25.9	+8.2
Mind-Cube (Yin et al., 2025)	4231	39.7	32.5	+7.2
OVO-S-Bench (ours)	**1680**	**37.1**	**31.3**	**+5.8**
OST-Bench (Lin et al., 2026a)	1112	44.2	40.5	+3.7
ERQA (Yin et al., 2025)	119	27.7	26.1	+1.6
SpatialScore-Hard (Wu et al., 2025)	68	30.9	30.6	+0.3
MMSI-Video-Bench (Lin et al., 2025)	222	22.1	23.2	-1.1
MMSI-Bench (Yang et al., 2025d)	190	25.8	27.0	-1.2
RoboBench (Luo et al., 2025b)	106	23.6	25.0	-1.4
VSI-Bench (Yang et al., 2025a)	498	27.1	29.3	-2.2

Table 13: GPT-5.4 text-only accuracy on spatial benchmarks (question and options only, no visual input). Δ is text-only minus random baseline. OVO-S-Bench row from Tab. 2; others from a 20% stratified sample.

Model	Paper (%)	R_{paper}	w/o RT3D (%)	$R_{w/o}$	ΔR
Gemini-3.1-Pro	59.2	1	54.4	1	0
Qwen3-VL-235B-A22B	53.6	2	49.7	2	0
InternVL-3.5-241B	50.9	6	48.8	3	-3
InternVL-3.5-38B	49.1	8	47.9	4	-4
GPT-5.4	50.9	5	47.8	5	0
Qwen3.5-27B	51.7	4	47.6	6	+2
Gemini-3.1-Flash-Lite	50.8	7	46.8	7	0
Qwen3.5-397B-A17B	52.1	3	46.0	8	+5
Qwen3-VL-32B	48.8	9	45.7	9	0
Qwen3.5-9B	45.8	12	43.6	10	+2

Table 14: Ranking stability after removing RoomTour3D (48% of items). R_{paper} and $R_{w/o}$ are rankings among 38 reported systems. Spearman $\rho = 0.92$ ($p < 10^{-4}$) across all 38 systems.

5 items in both groups. Per-type differences flip sign across the range −17 to +21 pp, with a residual mean of +5.0 pp that tracks the indoor–outdoor domain split rather than release year: the largest positive gap falls on relative speed judgment, where near-field egocentric motion is easier to read than outdoor walking and driving, and the largest negative gap falls on temporal depth estimation, where long outdoor traversals make elapsed time easier to estimate. Varying the minimum cell size from 3 to 10 items retains 19 to 11 types and moves the mean between +4.6 and +7.2 pp, leaving the

Source	L1		L2		L3		L4	
	N	%	N	%	N	%	N	%
RoomTour3D	174	43.6	308	48.7	163	55.1	159	52.4
Ego4D	166	43.1	113	47.4	74	31.1	9	38.0†
OmniWorld	111	32.1	22	39.1	8	33.9†	12	33.8
CODa	112	43.3	17	59.1	17	64.6	2	38.2†
Sekai	32	46.3	53	42.0	–	–	–	–
VSI-Bench	19	44.8	–	–	–	–	–	–
YouTube tours	–	–	–	–	17	54.9	–	–
Honda HDD	–	–	–	–	–	–	59	24.0
ARKitScenes	–	–	–	–	–	–	33	24.2
All	**614**		**513**		**279**		**274**	

Table 15: Source × level matrix over the 1,680-item evaluation set. % is accuracy macro-averaged over the 38 reported systems within the cell; N is the item count. †marks cells with fewer than 10 items. Row and column totals match Tab. 10 and Fig. 3d respectively.

Type	pre-2023		2025		Δ (pp)
	N	%	N	%	
1.1.2	12	49.0	10	35.0	+14.0
1.1.4	14	35.1	11	39.1	−4.0
1.2.1	18	43.9	14	34.9	+9.0
1.2.2	15	42.0	13	34.0	+8.0
1.2.3	13	35.1	9	41.1	−6.0
1.3.1	22	46.3	18	34.3	+12.0
1.3.2	19	42.5	16	35.5	+7.0
1.3.3	10	39.1	8	36.1	+3.0
1.3.4	16	51.1	12	30.1	+21.0
2.1.1	20	55.4	16	39.4	+16.0
2.1.2	16	53.3	12	48.3	+5.0
2.2.2	14	38.1	10	49.1	−11.0
2.3.1	15	42.8	12	40.8	+2.0
2.3.2	9	39.4	7	56.4	−17.0
3.2.1	18	39.1	6	31.1	+8.0
4.2	6	44.8	7	31.8	+13.0
Mean Δ over 16 types					+5.0

Table 16: Release-date probe. Accuracy macro-averaged over the 38 reported systems within each (task type, source group) cell, for the 16 types holding at least 5 items in both groups. pre-2023 is Ego4D; 2025 is OmniWorld and Sekai. Δ is pre-2023 minus 2025.

sign pattern unchanged. No configuration produces the uniform pre-2023 advantage that memorization would predict.

G Evaluation Methods Details

This appendix specifies the inference configuration used to produce every score in Section 4. We give the model inventory and provenance (§G.1), the decoding-parameter recipes (§G.2), the default frame-sampling policy (§G.3), the streaming-model ingestion protocol (§G.4), the prompt templates used at the model interface (§G.5), the verbatim policy-specific prompts used in the §B matrix (§G.6), the deterministic answer-extraction rule (§G.7), and the hardware and software stack (§G.8). The governing principle throughout is that every per-model choice (resolution, frame count, decoding temperature, reasoning-mode toggle) follows the recipe published by the model's authors; nothing is tuned on OVO-S.

G.1 Models Evaluated

The set of evaluated models is listed in Table 2 (the main results table, including closed-source proprietary MLLMs, open-source general video MLLMs, streaming video MLLMs, token-compression / memory-based methods, spatially fine-tuned MLLMs, and embodied foundation models). Open-source models are loaded from each author's official HuggingFace release; no weights, tokenizer, or vocabulary are modified, and inference is driven through the authors' released code path (vLLM for models whose card lists vLLM as a supported backend; the authors' provided inference scripts otherwise; see §G.8). Closed-source models are accessed through their providers' OpenAI-compatible endpoints. To allow re-evaluation against any future API revision, we record the access window: all closed-source runs reported in this paper were collected between 2026-02 and 2026-04 against the providers' `stable` (non-preview) endpoint tier where one exists, and against the latest preview tier otherwise (`gemini-3.1-pro-preview`, `gemini-3.1-flash-lite-preview`). The exact endpoint identifier is the model-name string shown in Table 2.

G.2 Decoding Parameters

Non-reasoning baseline. Every model whose row in Table 2 does not carry a thinking-mode toggle is run under the same decoding protocol: greedy decoding (T=0), no nucleus filtering, and a 1,024-token output cap. The cap is comfortably larger than the answer-letter response targeted by our prompt (§G.5) and accommodates models that emit a short justification before the letter.

Reasoning variants. For the reasoning-mode rows of Table 3, decoding follows the recipe published by each model's authors verbatim. Table 17 reproduces these recipes. Two design choices in the table merit note. First, the non-zero T=0.6 and top-p=0.95 used for InternVL-3.5-thinking are the authors' R1-style anti-repetition recipe documented on the model card; the same model card pre-

Family	T	top-p	Thinking toggle	Max tok
Qwen3-VL-thinking	0.0	–	`enable_thinking=true`	8,192
Qwen3.5-thinking	0.0	–	`enable_thinking=true`	32,768
InternVL-3.5-thinking	0.6	0.95	R1 system prompt + CoT user template	16,384
GLM-4.6V-Flash-thinking	0.0	–	`chat_template_kwargs.enable_thinking=true`	16,384
Grok-4.1-Fast-Reasoning	0.0	–	built-in (no user toggle)	4,096

Table 17: Reasoning-mode decoding recipes for the thinking-variant rows of Tab. 3. Each row reproduces the recipe documented on the corresponding model card. *Max tok* caps the thinking trace plus answer combined; truncation is tagged `T1a` (see §C.1).

scribes the chain-of-thought user template paired with the R1 system prompt, which is why InternVL-3.5-thinking is the only row that overrides the default prompt template (§G.5). Second, the per-family thinking-token cap (8K–32K) is the published thinking budget for each family; we observed truncation only on Qwen3.5-thinking when the trace exceeds 32K (591 traces across L1–L4, all tagged `T1a` per §C.1), which is recorded as a model-side failure rather than a configuration artifact.

G.3 Default Frame Sampling

For each query at timestamp t_q the default visual input is $N{=}128$ frames sampled uniformly across the prefix $[0, t_q]$. The choice of N is the largest uniform-frame budget that fits inside every evaluated model's published context window without truncation; Section B reports the sensitivity of overall and per-level accuracy to this choice. Frames are resized so the shorter side equals the spatial resolution prescribed by each model's training recipe (this falls in the 336–512 px range across the evaluation set), and the longer side is preserved up to the model's published per-frame visual-token cap where one exists (*e.g.* Qwen3-VL's `max_pixels` setting, Flash-VStream's 50,176-pixel cap). Frame extraction is cached: the decoded N-frame tensor is keyed by the SHA-1 of the video file and the sampling specification, so a re-run of the same model on the same annotation reads from cache and is bit-for-bit identical to the first run.

G.4 Streaming-Model Ingestion Protocol

Native protocol. Streaming-architecture models (StreamingVLM, StreamForest, Flash-VStream in the main results; HERMES, InfiniPot-V, FluxMem, StreamingTOM in the token-compression block) consume the video as a sequential stream and apply a model-side memory or KV-compression mechanism that is calibrated for a specific ingest rate. We therefore do *not* impose the uniform 128-frame default on these models for their main-result row in Table 2; instead the video is read from the start until t_q at the model's published streaming rate, and answers are generated using the resulting compressed state. Table 18 lists, for each streaming model, the ingest rate and the chunk or compression schedule we use. Every entry is the value provided by the model's authors (the model card, the released inference script, or the corresponding paper); no streaming hyperparameter was tuned on OVO-S, and the compression-side hyperparameters (*e.g.* StreamingTOM's contrastive-token retention thresholds) take the authors' released defaults.

Model	Ingest rate	Memory / compression unit
StreamingVLM	2 fps	continuous (linear-attention context)
StreamForest	1 fps	64-frame budget
Flash-VStream	2 fps	continuous, 50,176-pixel per-frame cap
HERMES	2 fps	16-frame chunk, KV pruned per chunk
InfiniPot-V	1 fps	KV-budget per chunk
FluxMem	1 fps	flux-memory replacement schedule
StreamingTOM	1 fps	32-frame encoder batch, contrastive-token retention

Table 18: Streaming ingest rate and compression schedule for each streaming-architecture model. All values from the corresponding paper, model card, or released inference script; no streaming hyperparameter is tuned on OVO-S.

G.5 Prompt Templates

Default template. With the exception of the InternVL-3.5-thinking row (see below), every model in every evaluation track receives the same prompt template. The question text and option strings are substituted into the placeholders verbatim:

```
You are evaluating a video understanding
task. Based on the video frames provided,
answer the following multiple choice
question.
Question: {question}
Options:
{options_text}
Instructions:
- Respond with ONLY the letter of your
answer (e.g., "A" or "B").
- Do not include any explanation or
additional text.
Your answer:
```

Chain-of-thought template (reasoning models with a CoT recipe). InternVL-3.5-thinking is the only main-text row whose model card prescribes an explicit chain-of-thought user template alongside the R1 system prompt; for that row we replace the *Instructions* block above with the following, while leaving the question and option blocks unchanged:

```
Instructions:
- Think step by step about the spatial
```

```
relationships and scene content shown in
the frames.
- After your reasoning, output your final
answer on a new line in the format:
Answer: X
```

All other thinking-mode rows (Qwen3-VL-thinking, Qwen3.5-thinking, GLM-4.6V-Flash-thinking, Grok-4.1-Fast-Reasoning) use the default template above; the reasoning trace is gated by the model-side thinking-mode flag listed in Table 17, not by the user prompt.

G.6 Policy-Specific Inputs

A practical clarification on the matrix in §B: the eight frame-sampling policies (*single@query*, *nearest-16f@4fps*, *uniform-32*, *uniform-64*, *uniform-128*, *uniform-256*, *log-decay-128*, *oracle-evidence*) differ *only* in the visual input fed to the model, not in the textual prompt. Every policy submits exactly the same multiple-choice prompt produced by the template above, with the same question and option strings. The *text-only* control baseline reported in the *Baseline & Controls* block of Table 2 is the same prompt submitted with an empty image list, and is evaluated under GPT-5.4 to bound how much of the benchmark is answerable from textual cues alone.

G.7 Answer Extraction

We adopt multiple-choice scoring throughout because it enables deterministic, reproducible evaluation without relying on LLM-as-judge for open-ended answers, which would itself introduce the spatial grounding errors this benchmark is designed to measure.

Model outputs are scored by a deterministic post-processor. For thinking-mode outputs we first remove any `<think>. . . </think>` span so that subsequent matching applies only to the post-thinking tail; outputs without a closing `</think>` after the published thinking budget are treated as truncated and tagged `T1a` per §C.1. The extractor then tries the following patterns in order on the (de-thinked) response, restricted to the letter range `A–G`, and returns the first match:

1. Within the last 300 characters: a tail `Answer: X` (including *final answer*, *final*).
2. Within the last 300 characters: a Cosmos-style `<answer> X` tag.
3. Within the last 300 characters: a bare single letter at the very end (*e.g.* "... therefore B").
4. Within the last 300 characters: a GLM-style `<|begin_of_box|> X` marker.
5. A single letter at the start of the stripped response.
6. Anywhere in the response: `answer / choice / option(s): X` with a true separator.
7. Anywhere in the response: a parenthesized or bracketed letter, *e.g.* "(B)" or "[C]".

Responses for which none of these patterns matches are scored as wrong; they are not skipped, and they contribute to the denominator of every per-task and overall accuracy in Section 4. The order above is the tail-first order that minimizes accidental matches against letters embedded in reasoning prose; the only step that scans the full response is step (vi), which requires the trigger word *answer*, *choice*, or *option(s)* followed by a real separator.

G.8 Hardware and Software

All open-source inference is run on NVIDIA H-series GPUs (H800 80 GB and H200 141 GB). Models whose authors release a vLLM-compatible inference path are served with vLLM $\geq$ 0.11 in offline batched mode; models without such a path (predominantly the streaming-architecture and token-compression entries) are served through the authors' released inference scripts at the dependency versions pinned in those repositories. All closed-source inference goes through the providers' OpenAI-compatible HTTPS endpoints. Video frame decoding uses `decord` as the primary backend with an OpenCV fallback for codecs unsupported by `decord`. The decoded N-frame tensor for each (model, query, sampling specification) tuple is cached on disk under a content-addressed key (the SHA-1 of the video file together with the sampling specification), so repeated evaluation against the same query reads from cache and produces a bit-for-bit identical visual input; the cache footprint for the full benchmark is on the order of 50 GB.

H Response-Type Decomposition

OVO-Bench (Niu et al., 2025) separates *when* a model should answer into three regimes: *Backward* (the answer lies in past observation), *Real-Time* (the answer lies in the current view), and *Forward* (the model must decide when to speak). OVO-S-Bench instantiates the first two with spatial content and states Forward as a non-goal (Section 1).

Level	Real-Time	Backward
L1	629	0
L2	0	540
L3	253	26
L4	0	274
Total	**882** (51%)	**840** (49%)

τ (s)	L3 RT	L3 BW	Overall RT (%)
0.5	238	41	50.3
1.0	248	31	50.9
2.0	**253**	**26**	**51.2**
3.0	257	22	51.5
5.0	263	16	51.8
10.0	271	8	52.3

Table 19: Response-type decomposition over the 1,722 scored queries of the public release (top) and its sensitivity to the L3 gap threshold τ (bottom). The row in bold is the setting reported in Section 4.1.

This appendix quantifies that scope so the claim is checkable rather than asserted.

Labeling rule. Let t_q be the query timestamp, e_q the end of the annotated evidence interval, and $g_q = t_q - e_q$ the evidence-to-query gap. Three of the four levels are determined by the level itself and the fourth by g_q:

- L1 is *Real-Time*: the evidence interval brackets the query and the answer is readable from the current view.
- L2 and L4 are *Backward*: the answer requires an observation that has left the view (L2) or an integration over the entire explored region (L4).
- L3 is *Real-Time* when $g_q \leq \tau$ and *Backward* otherwise, with $\tau = 2.0$ s.

L3 is the only level carrying both labels: its median evidence span equals L1's (2.0 s, §3.3) while a minority of its items reference earlier observation. The threshold matches that median, so evidence ending within 2 s of the query counts as still in view.

Counts and threshold sensitivity. Table 19 gives the decomposition and its sensitivity to τ. Sweeping τ from 0.5 to 10 s moves the overall split by 2.0 points (50.3% to 52.3% Real-Time) and reassigns 33 of 1,722 queries in total, because 85% of L3 gaps fall below 0.5 s. The decomposition is therefore a property of the taxonomy rather than of the threshold.

I Open-Ended Answering Pilot

Multiple-choice scoring keeps every reported number deterministic and reproducible, and it compresses partial spatial knowledge into a single letter. We measure both effects on a stratified sample of 200 items (50 per level), with GPT-5.4 answering in free form under the same frame input as the main protocol.

Scoring. An independent judge (Gemini-3.1-Pro) scores each free-form answer against the gold spatial fact on a three-point scale: 2 when the key relation, region, count, or layout matches; 1 when the relation or region is right but the specifics are imprecise; 0 when the answer contradicts the gold fact or is too vague to recover it. A second judge agrees with the first on the binary fully-correct decision for 96% of items. Prompts and both judges' scores are released with the code.

The two distortions run in opposite directions. Strict open-ended accuracy is 45.0 against 55.5 under multiple choice on the same items: options scaffold answers the model cannot produce unaided. In the other direction, 20% of free-form answers are partially correct, and among items failed under multiple choice, 37% still carry partially or fully correct spatial content in free form. One trace identifies that the pillows lie "side by side in rows" while missing their global triangular arrangement, which the letter-level score cannot distinguish from a model that has no layout at all. Awarding half credit for partial answers yields 55.0 against 55.5, so the two distortions cancel in aggregate while the item-level information they carry is recoverable only under graded scoring.

Level breakdown. Table 20 reports per-level counts. The ordering of multiple-choice accuracy across levels follows GPT-5.4's full-benchmark profile (Tab. 2), with L4 lowest. These 200 items score 55.5 under multiple choice against GPT-5.4's 50.9 over the full benchmark, so the sample is easier than the benchmark average and the pilot supports within-sample comparisons.

J Future Work

Scaling annotation with humans in the loop. Every item in this release is written by hand, which is what buys verified query timestamps, verified evidence intervals, and shortcut screening, and it is also what caps the scale at 1,680 items and 804

Level	Open-ended			Strict	Graded	MCQ
	full	part	none			
L1	26	9	15	52.0	61.0	60.0
L2	28	9	13	56.0	65.0	64.0
L3	21	12	17	42.0	54.0	56.0
L4	15	10	25	30.0	40.0	42.0
All	**90**	**40**	**70**	**45.0**	**55.0**	**55.5**

Table 20: Open-ended pilot on 200 stratified items (50 per level). *full*/*part*/*none* are judge scores 2/1/0. *Strict* counts only full credit; *Graded* awards 0.5 for partial; *MCQ* is the multiple-choice score on the same items.

person-hours. The route to a larger release keeps the human as verifier rather than author: a model proposes candidate question–option–answer tuples, and annotators run the same blind double review that governs the current pipeline (§E.5). Adapting an automated spatial-supervision pipeline such as Holi-Spatial (Gao et al., 2026) to emit streaming annotations, including a query timestamp and an evidence interval per item, would supply those candidates directly.

From diagnosis to remedy. The compression diagnostics locate the L4 failure before reasoning begins: evidence recall on L4 falls to 0.14–0.42 against 0.60–0.76 on L1 (Section 4.3.3), so an intervention has to change which spatial evidence survives, not how the surviving tokens are read. Two families of remedy follow directly. Explicit spatial memory keeps an allocentric structure updated as the stream arrives, through 3D-aware memory, test-time adaptation, world-model imagination, or structured maps. External geometric grounding supplies at inference what the representation lacks, through depth, open-vocabulary detection, or reconstruction tools invoked by an agent, as in Think3D (Zhang et al., 2026b); OVO-S-Bench measures such interventions per level, which is where a grounding remedy should show its effect. The open question our diagnostics sharpen is when to invoke such tools, since always-on tool use adds scaffolding cost on the levels a model already handles.

Evaluation protocol. Graded open-ended scoring (Appendix I), richer human baselines, and interactive embodied tasks each recover information the current protocol compresses away. We also plan a contamination-aware release protocol, with a lightweight public subset and a protected evaluation split if the benchmark is used for leaderboard evaluation.

K Additional Examples

The following pages present representative examples across all levels and task families of OVO-S-Bench. Each page pairs streaming spatial questions with visual evidence, query timestamps, and answers.

Level1.1 Egocentric Metric Estimation

RoomTour3D

Query Time: 770.0
Question: Based on the current frame, what is the approximate **distance** from the camera to the **black nightstand** next to the bed?
Options: A: 0.5 meters B: 6 meters C: 4 meters **D: 2 meters**



Query Time: 770.0

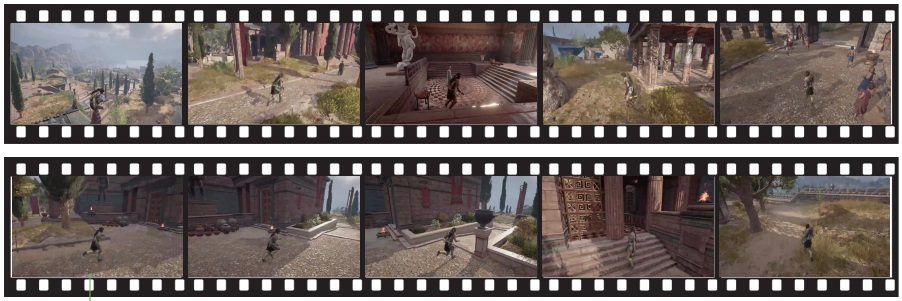
CODa_full

Query Time: 53.5
Question: In the current view, **which is taller**, the blue trash can or the metal trash can?
Options: **A: The blue trash can is taller.** B: The metal trash can is taller.



Query Time: 53.5

OmniWorld



Query Time: 165.667

Query Time: 165.667
Question: In the current frame, which is **closer to the camera**: the fire pit or the person?
Options: A: The fire pit is closer to the camera.
B: The person is closer to the camera.

Level1.2 Local Spatial Relationships

OmniWorld

Query Time: 212.5
Question: In the current view, what best describes the spatial relationship between the white car and the building on the left?
Options: **A: The car is in direct physical contact with the exterior of the building.**
B: The car is separated from the building by a row of parked bicycles.



Query Time: 212.5

CODa_full

Query Time: 40.8
Question: In the current frame, is the potted plant occluded? If so, what is it occluded by?
Options: A: The plant is partially blocked by the wooden column.
C: The plant is not occluded and is fully visible.



Query Time: 40.8

RoomTour3D

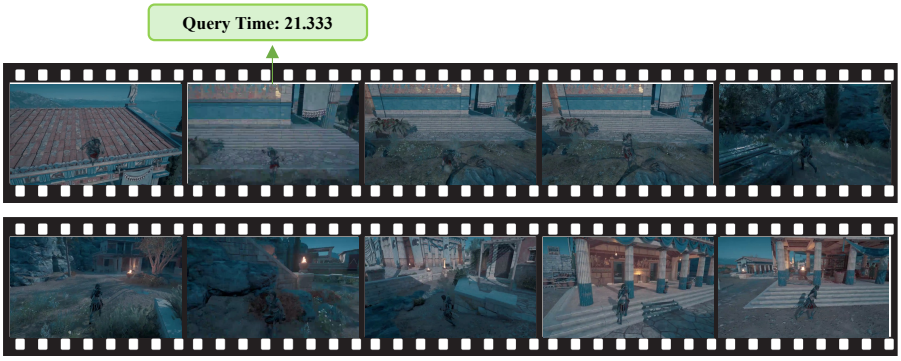
Query Time: 429.667



Query Time: 429.667
Question: In the current view, what object is directly supporting the empty vase?
Options: **A: the cabinet** B: cannot be determined from the video
C: the floor D: the desk

Level1.3 Dynamical Spatial Perception

OmniWorld



Query Time: 21.333
Question: Based on the visual flow, what is the camera's dominant translation **direction** in this segment?
Options: **A: Translating downward** B: Translating upward
C: Translating leftward D: Translating rightward
E: Translating forward F: Translating backward

Ego4D

Query Time: 52.4
Question: Based on the visual flow, is the camera currently yawing left, yawing right, or not rotating?
Options: **A: Yawing left** B: Yawing right C: Not rotating



CODa_full



Query Time: 6.5
Question: In this segment, what is the motion state of the person in dark clothing on the left side of the walkway relative to the surrounding scene?
Options: **A: The person in dark clothing on the left side of the walkway is moving independently relative to the surrounding scene.**

Level2.1 Visual Scene Re-identification

RoomTour3D

Query Time: 113.3
Question: At the current moment, is the camera showing the same place as at 73 seconds?
Options: A: Different from the referenced moment **B: Same as the referenced moment**

Reference Time: 73.0

Query Time: 113.3



OmniWorld

Query Time: 27.3



Query Time: 27.3
Question: By the time the video reaches the queried moment, how many times has this same hallway scene been shown in total?
Options: **A: 2** B: 4 C: 1 D: 3

CODa_full



Query Time: 55.3

Query Time: 55.3
Question: How many separate times does the camera wearer walk on an outdoor street with the herringbone brick pattern during the video?
Options: A: 3 **B: 2** C: 1 D: 4

Level2.2 Out-of-View State Maintenance

OmniWorld

Query Time: 178.5
Question: At the current moment, where is the giant humanoid statue located relative to the camera, outside the current view?
Options: **A: Out of view to the left** B: Out of view to the right
C: Out of view behind the camera D: Out of view in front of the camera



Query Time: 178.5

Ego4D

Query Time: 248.0
Question: Earlier in the video, a national flag was visible. Which country's flag was it?
Options: A: The flag of the United Kingdom B: The flag of France
C: The flag of the United States D: The flag of Canada



Query Time: 248.0

RoomTour3D

Query Time: 198.3
Question: As the camera moved past the staircase, how many beige balloons were attached near the stairs?
Options: **A: 4** B: 3 C: 6 D: 5

Query Time: 198.3



Level2.3 Chronological Event Sequencing

CODa_full

Query Time: 12.2
Question: In the video, was the indoor hall visited before or after the street/outdoor walkway with trees?
Options: **A: The hall was visited after the street.** B: The hall was visited before the street.



Query Time: 12.2

Ego4D

Query Time: 1133.0
Question: Thinking about the route shown in the video, what is the last indoor room that we visited before going outdoors?
Options: A: a toilet B: a dining room C: a kitchen **D: a living room**



Query Time: 1133.0

RoomTour3D

Query Time: 184.0
Question: Which option was shown for a longer duration in the video: the bathroom or the pool table room?
Options: **A: the bathroom** B: the pool table room

Query Time: 184.0



Level3.1 Spatial What-if Simulation

RoomTour3D

Query Time: 79.3
Question: After rotating the wall painting in the current view to a vertical orientation, can it pass through the doorway in front?
Options: A: Yes **B: No**



RoomTour3D

Query Time: 202.0
Question: In the current configuration, if the green pillow were translated only backward, would it make contact with the TV?
Options: A: Yes, it would make contact.
B: No, it would not make contact.



OmniWorld



Query Time: 99.6
Question: In the current view, is there enough space between the pillars in front for a bicycle to pass through?
Options: A: No **B: Yes**

Level3.2 Spatiol-Temporal Consistency Verification

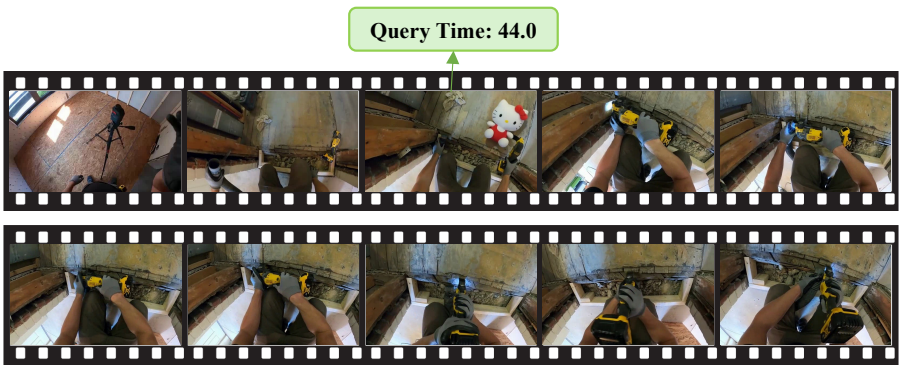
Ego4D

Query Time: 10.0
Question: Compared with the immediately preceding moment in the video, which object has newly appeared on the White mortar mixer in the current view?
Options: A: A red umbrella **B: A Homer Simpson**
C: A Hello Kitty D: A Golden Retriever

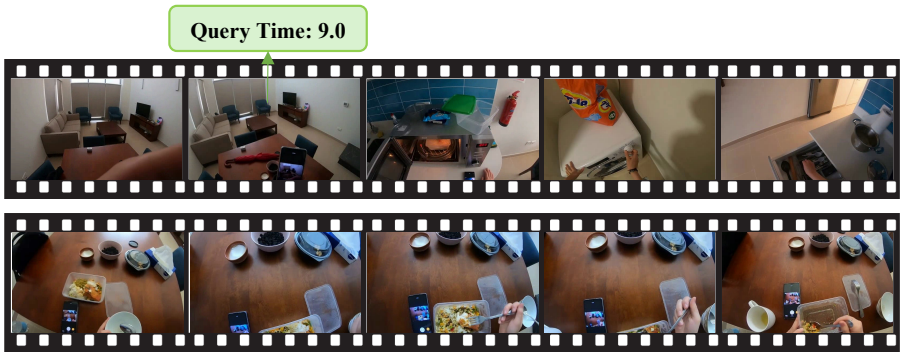


Ego4D

Query Time: 44.0
Question: Compared with the immediately preceding moment in the video, which object has newly appeared on the floor in the current view?
Options: A: A Golden Retriever B: A teddy bear
C: A yellow parrot **D: A Hello Kitty**



Ego4D



Query Time: 9.0
Question: Compared with the immediately preceding moment in the video, which object has newly appeared on the square table in the current view?
Options: A: A yellow parrot **B: A red umbrella**
C: A Hello Kitty D: A Homer Simpson

Level3.3 Navigation-Oriented Spatial Planning

RoomTour3D



Query Time: 392.0
Question: From Bathroom A to Billiard Table, which route is the shortest or most reasonable?
Options: A: Bathroom A -> Room A -> Corridor A -> Kitchen -> Billiard Table
B: Bathroom A -> Kitchen -> Room A -> Living room -> Billiard Table
C: Bathroom A -> Corridor A -> Room A -> Recreation Room -> Billiard Table
D: Bathroom A -> Room A -> Corridor A -> Living room -> Recreation Room -> Billiard Table

CODa_full

Query Time: 147.2
Question: If Path is blocked by an obstacle near Tree A, how should one detour from Avenue to Door?
Options: A: Detour via Avenue -> Door
B: Detour via Avenue -> Tree A -> Door
C: Detour via Avenue -> Tree A -> Path -> Door
D: No valid detour can be confirmed from the observed layout.



RoomTour3D



Query Time: 271.0
Question: Without passing through Path B, which option best describes whether Backyard is reachable from Garage based on the observed layout?
Options: .....
C: Reachable via Garage -> Backyard
D: Not reachable without passing through Path B.

Level4 Allocentric Spatial Graph

RoomTour3D

Query Time: 171.333
Question: Assuming the entrance faces due north, what is the order of the three rooms from east to west?
Options: A: Room B, Room C, Room A, **B: Room A, Room B, Room C,**
C: Room A, Room C, Room B, D: Room C, Room B, Room A



RoomTour3D



Query Time: 720.0
Question: Based on the full video, which of the following most accurately describes the topological relationship between Area A and Corridor A?
Options: A: There is no evidence of a connection between the two.,
B: Area A is an exterior area that leads into Corridor A.,
C: Area A must be accessed through Area D to reach Corridor A.,
D: Area A is an open functional space within Corridor A.

Honda



Query Time: 4144.2
Question: Based on the vehicle's driving status in the video, select the correct bird's-eye view trajectory map.
Options: A:  B:  **C:**  D: