TRACE-VSI: BENCHMARKING PROGRESSIVE SPATIAL REASONING ALONG EGOCENTRIC TRAVERSALS

Anonymous authors
Paper under double-blind review

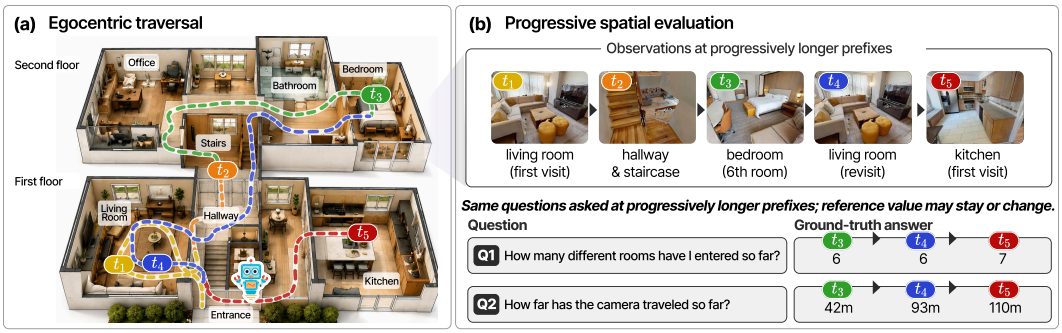


Figure 1: **Spatial judgments draw on experience accumulated along a home traversal.** (**a**) An egocentric camera follows a recorded path through a multi-floor home. (**b**) Two benchmark questions are evaluated at the same points in the video. Returning to the living room adds traveled distance without adding a distinct room. Entering the kitchen for the first time changes both reference answers. Questions are abbreviated for display, and the shown values are illustrative.

ABSTRACT

Understanding a home from egocentric video requires relating new views to previously seen places and accounting for movement between them. Answers to cumulative questions, such as how many distinct rooms have been entered, can therefore change as a traversal continues. End-of-video accuracy alone cannot reveal earlier mistakes or whether predictions track these changes. We introduce TRACE-VSI, a scan-grounded benchmark with 18 full-traversal question types about places, observed routes, and spatial configuration, and repeated questions about 5 cumulative quantities as observations accumulate. At each point, the model is queried independently using video from the start of the traversal up to that point. We score each answer for accuracy and compare predicted changes with changes in the correct answers using Spatial Change Accuracy (SCA). Across 30 systems, 22 exceed the answer-frequency baseline on place/time binding after the full traversal, but only 6 do so on observed-route questions. At earlier and later points, correct count changes often coexist with incorrect counts: Qwen3-VL-8B scores 55.3 SCA on Distinct Rooms but only 13.1% joint exact accuracy. Distance changes are also inaccurate: Qwen2.5-VL-72B has 23.6 m Path Length change MAE, compared with 9.0 m for a video-free prior. Larger frame budgets can improve change agreement without clear evidence of improved joint count correctness. These results reveal persistent count offsets and inaccurate distance increments, showing why accurate changes alone do not establish accurate spatial tracking.

1 INTRODUCTION

Spatial cognition relates landmarks, routes, and layouts encountered during movement, often through the idea of a cognitive map (Tolman, 1948; Siegel & White, 1975). Egocentric video poses a related challenge for multimodal models: questions about a home require connecting current views to previously visited places. Individual frames reveal only part of the environment, so recognizing what

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is visible must be combined with relating observations across place transitions and revisits. We ask whether models can track spatial state across a traversal: observed places, their relationships, and the route traveled so far. We do not assume a particular internal representation.

The quantities that describe this state can change as the traversal continues. In Figure 1, revisiting a room increases traveled distance but leaves the distinct-room count unchanged, whereas entering a new room changes both. Full-traversal accuracy cannot show whether a model answered correctly before these changes, and accuracy averaged across observation points hides how answers to the same question change. An unchanged overestimate can become correct as the true count grows. Conversely, a correct increment can preserve an error in the count. Evaluating spatial tracking therefore requires scoring each estimate at its observation point and checking how the estimates change together.

We introduce TRACE-VSI, a scan-grounded benchmark combining 18 full-traversal spatial QA types with progressive evaluation of 5 cumulative quantities. The full-traversal questions cover places, landmarks, route connectivity, and house-scale configuration. For the progressive evaluation, we ask the same question at earlier and later points. Each call uses the video from the traversal start up to that point, which we call a *prefix*. Calls are independent, and the correct answer, or *reference*, is recomputed at each endpoint. Models receive sampled RGB frames; reference answers come from scan annotations and recorded trajectories. In particular, route questions use connections actually traversed in the video, so their references describe the observed route through the home.

Our contributions are:

- We introduce TRACE-VSI, a scan-grounded benchmark with 18 full-traversal spatial QA types covering place and time, cumulative state, observed routes, displacement, distance, and area.
- We evaluate 5 cumulative quantities at independently queried observation prefixes, recomputing the reference at each prefix to assess both the estimates and their changes.
- Across general, spatially specialized, and streaming systems, we identify count offsets that survive correct increments, quantify errors in distance and area changes against simple priors, and examine how frame selection and budget alter these outcomes.

2 RELATED WORK

2.1 SPATIAL INTELLIGENCE

Spatial intelligence benchmarks vary in the observations they provide and the spatial relationships they ask models to recover. ScanQA and SQA3D evaluate question answering grounded in reconstructed 3D scenes (Azuma et al., 2022; Ma et al., 2023). Building on these settings, SceneVerse, EmbodiedScan, and MMScan expanded scene coverage and grounding annotations (Jia et al., 2024; Wang et al., 2024; Lyu et al., 2024). Video and multi-image evaluation extends spatial reasoning to information distributed across views (Yang et al., 2025a;b), while broader suites cover spatial scales, cognition, and reasoning (Sun et al., 2025; Xu et al., 2025; Wu et al., 2025b). Egocentric and embodied benchmarks further examine how observations support situated questions and recall (Majumdar et al., 2024; Li et al., 2026; Lei et al., 2026; Lang et al., 2026; Wu et al., 2026; Zhang et al., 2026). TRACE-VSI organizes these demands around a recorded traversal through a home. Its questions include place identity and binding, connections and route alternatives among observed places, and physical quantities derived from the same scan and trajectory.

Recent models have likewise introduced geometric supervision, spatial representations, and alignment objectives to improve spatial reasoning (Fan et al., 2025; Zheng et al., 2025a; Wu et al., 2025a; Cai et al., 2025; Liu et al., 2026). Other approaches acquire spatial evidence through active perception, adaptive view selection, or geometry-aware video training (Zheng et al., 2025b; Zhi et al., 2025; Han et al., 2024). For extended observation histories, explicit memory and retrieval mechanisms (Hu et al., 2025; Yang et al., 2025c; Anwar et al., 2024; Zheng et al., 2026; Xie et al., 2024; Ginting et al., 2025) and scene or spatiotemporal graph representations (Gorlo et al., 2025; Sohn et al., 2025; Saxena et al., 2024; Zemskova & Yudin, 2024; Werby et al., 2024) provide persistent structures for aggregating information across observations. These approaches motivate evaluating both the breadth of spatial questions a system can answer and the accuracy of its estimates as observations grow.

Our comparisons include general-purpose and spatially specialized models, alongside systems that explicitly retain or structure earlier observations.

2.2 Online Video Understanding

Online video understanding considers models that process visual observations causally as they arrive. VideoLLM-online enables language interaction with continuous video streams, while VideoLLM-MoD targets efficient processing of dense streaming inputs (Chen et al., 2024; Wu et al., 2024). StreamChat updates visual context during streaming generation, whereas VideoChat-Online maintains historical information through an online memory mechanism (Liu et al., 2024; Huang et al., 2025). TimeChat-Online and StreamingAssistant reduce computation for long video streams (Yao et al., 2025; Jin et al., 2025). These approaches manage the visual context available at each query time. Spatial evaluation must also specify what can be answered from that context and how the correct answer depends on the query time.

UCS-Bench asks timestamp-specific questions about user-relative location, movement, reachability, and quantities, including repeated queries at different timestamps (Wang et al., 2026). Stream3D-Bench covers camera motion, object relations, chronology, and scene measurement under memory, current-observation, and monitoring modes (Yu et al., 2026). These benchmarks already evaluate spatial questions at different observation times. TRACE-VSI combines this temporal perspective with house-scale observed-route questions and paired estimates of cumulative spatial quantities. The route questions include first hops, mandatory places, and transition costs; the progressive questions use references recomputed by the same construction rule at both prefixes. We score both estimates and their changes to detect correct increments that coexist with inaccurate spatial quantities.

3 Benchmark and Evaluation Protocol

We construct TRACE-VSI to evaluate streaming spatial intelligence from house-scale egocentric traversals. The benchmark combines questions about places, observed routes, and spatial configuration at the end of a traversal with repeated questions at multiple observation points.

3.1 Spatial Questions and Taxonomy

The 18 full-traversal question types are grouped by the spatial operation needed to answer them (Table 1). The groups cover place/time binding, cumulative spatial state, observed-route graphs, relative spatial displacement, motion distance, and area measurement/comparison. They describe answer requirements rather than task difficulty; the table gives an example and answer format for each type.

The progressive evaluation asks whether models relate newly observed views to places encountered earlier and account for the route traveled so far. We use 5 cumulative quantities to probe these requirements. Distinct Rooms counts rooms and hallways entered by the camera, requiring revisits to be distinguished from new places. Target Visits counts entries into a specified place, with successive visits separated by leaving and re-entering. Object-Containing Rooms counts distinct owner rooms of clearly observed target objects; seeing an object can contribute even when the camera has not entered its room.

Path Length measures the three-dimensional arc length of the camera route. Visited Floor Area sums the floor areas of visited rooms and hallways, so it measures the extent of visited places rather than visible floor pixels. A revisit can therefore increase distance and visits while leaving the distinct-room count and visited area unchanged. For each quantity, the same construction rule defines a reference at every eligible prefix from scan and trajectory records. Numerical answers then let us assess both each estimate and its change. Task and endpoint selection do not use model outputs.

3.2 Data Construction

We use a scalable automated construction pipeline to generate traversal videos and spatial QA references from scan annotations and trajectory records. The same construction rules apply to additional routes and observation prefixes, allowing references to be recomputed as the observed

Table 1: **Spatial question types in TRACE-VSI.** The 18 full-traversal types are grouped by the spatial operation needed to answer them. Each row gives an abridged example and answer format; the 5 progressive types are marked with †.

Question type	Illustrative question (abridged)	Answer format
Place/time binding		
Object-to-Room Binding	"In which place did the wardrobe appear?"	MCQ: 4 places
Same-Room Episodes	"The camera saw a beam early in the traversal and a fireplace later. Were they seen in the same room?"	Binary: Yes/No
First-Seen Order	"Which option orders refrigerator, oven, countertop item, flower vase, and book by first clear appearance?"	MCQ: 4 orders
Cumulative spatial state		
Distinct Rooms†	"Up to this point, how many distinct rooms or hallways has the camera entered?"	Scalar: integer
Object-Containing Rooms†	"Up to this point, in how many distinct rooms was a pillow clearly visible?"	Scalar: integer
Target Visits†	"How many separate times did the camera enter a space with a bench?"	Scalar: integer
Observed-route graph		
Direct Traversal	"Between which pair of places did the camera move directly?"	MCQ: 4 place pairs
Shortest-Route First Hop	"From a space with a stool to one with at least two clothes-hanger rods, which place comes first on an observed route with the fewest transitions?"	MCQ: 4 places
Mandatory Place	"Which place lies on every observed route from a space with a bathtub to a space with a fireplace?"	MCQ: 4 places
Destination-Cost Order	"From a space with a towel bar, which option orders four destinations from most to fewest transitions along their shortest observed routes?"	MCQ: 4 weak orders
Itinerary-Cost Order	"Which option orders four routes through the same three destinations by decreasing total transition count?"	MCQ: 4 rankings
Relative spatial displacement		
Floor Offset	"Compared with a space with at least two lights, how many floors higher or lower is a space with a cushion?"	Scalar: signed integer
Fixed-Axis Bearing	"Relative to the camera's initial viewing direction, what is the bearing from an earlier landmark viewpoint to a later one?"	MCQ: 8 directions
Motion distance		
Path Length†	"Across the video shown, how far did the camera travel along its recorded path, in metres?"	Scalar: real (m)
Shortest Observed Trip	"Among observed trips from a space with a magazine to a space with a telephone, what was the shortest traveled distance?"	Scalar: real (m)
Area measurement/comparison		
Visited Floor Area†	"What is the combined floor area of the distinct rooms and hallways entered so far?"	Scalar: real (m^2)
Room Area	"Approximately what is the floor area of a room containing a bed, two or more decorations, and a lamp?"	Scalar: real (m^2)
Room-Area Rank	"Which option orders four specified rooms from largest to smallest by floor area?"	MCQ: 4 rankings

route grows. The evaluated full-traversal subset contains 950 questions from 90 traversals of 25 scanned homes in HM3D-Semantics (Ramakrishnan et al., 2021). Median durations, with interquartile ranges in brackets, are 2.75 [2.11, 3.42] minutes for full traversals and 0.95 [0.47, 1.61] minutes for queried intermediate prefixes.

The source trajectories combine original RGB and pose streams supplied with HOV-SG (Werby et al., 2024) and additional routes generated in these scans. For generated routes, semantic annotations identify candidate places and floors, and visit points are selected on the navigable surface. Paths between visit points are checked against the scene navmesh. The selected routes include visits to multiple places and revisits, while route selection limits excessive reuse of the same path segments. Each route is converted to a timestamped camera-pose sequence; RGB and semantic frames are rendered at those poses, with the frame–pose correspondence retained.

Question construction is a separate step. Type-specific rules use scan annotations, geometry, camera poses, and recorded events to determine question targets and reference answers. Route-question references use place transitions recorded along a traversal, not the reachability graph used to plan its route. The observed graph links places when the camera traverses directly between them. Its edges define the available routes for connectivity questions and the transition costs used to compare route alternatives. A later formatting step produces the displayed question and options while preserving the selected question's operation and reference answer. Automated checks verify source provenance and consistency of these records. We also manually review a sample of the evaluation questions. Models receive sampled RGB frames, not the scan, semantic frames, poses, or planner graph used in construction.

3.3 Evaluation Protocol

Full-traversal questions use the complete recorded route. For the 5 cumulative types, we ask the same question at earlier and later prefixes and recompute its reference at each endpoint from the scan and trajectory records. Each prefix starts at the route's beginning and is queried independently. The later call does not receive the earlier answer. Questions, endpoints, and references are fixed before model evaluation.

We compare consecutive eligible endpoints whose references differ. For the three count types and Visited Floor Area, each interval contains exactly one annotated reference-changing event. Path Length pairs require movement but no semantic event because traveled distance changes continuously. These pairs measure how answers change with the available observations. The resulting cohort contains 953 pairs, with counts varying by the events eligible for each question type: 384 for Distinct Rooms, 185 for Object-Containing Rooms, 118 for Target Visits, 180 for Path Length, and 86 for Visited Floor Area. Table 9 reports the source-question and home counts; Appendix B gives the selection rules. These task-specific cohorts support comparisons between systems within each type, rather than a controlled ranking of task difficulty.

3.4 Evaluation Metrics

Endpoint scores. We score each answer against its endpoint reference, using exact accuracy for categorical types and the VSI-Bench mean relative accuracy (MRA) for numerical types:

$$\mathrm{MRA}(\hat{y}, y) = 100 \times \frac{1}{10} \sum_{\theta \in \{0.50, 0.55, \ldots, 0.95\}} \mathbf{1}\left[\frac{|\hat{y} - y|}{|y|} \leq 1 - \theta\right], \tag{1}$$

with exact matching when $y = 0$.

Change accuracy. To assess whether estimates change by the correct amount, let y_i^t and $\hat{y}_i^t$ be the reference and numerical prediction at endpoint $t \in \{\mathrm{early}, \mathrm{late}\}$. Define the signed error $e_i^t = \hat{y}_i^t - y_i^t$ and $\Delta y_i = y_i^{\mathrm{late}} - y_i^{\mathrm{early}}$ and $\Delta \hat{y}_i = \hat{y}_i^{\mathrm{late}} - \hat{y}_i^{\mathrm{early}}$. All progressive pairs used for the main change score have $\Delta y_i \neq 0$. We measure change agreement with Spatial Change Accuracy (SCA):

$$\mathrm{SCA}_i = 100 \max\left(0, 1 - \frac{|\Delta \hat{y}_i - \Delta y_i|}{|\Delta y_i|}\right). \tag{2}$$

An invalid numerical answer at either endpoint gives zero SCA. For each system and question type, we average pair scores within each source question, then average questions equally over the fixed eligible pairs. Early and late MRA use the same pairs and weights, with invalid answers scored as zero at their respective endpoints.

MRA and SCA normalize errors by different reference quantities, so their numerical difference has no common interpretation. All primary reference-changing count pairs have a $+1$ reference increment. SCA is therefore 100 for a predicted $+1$ increment and zero otherwise. An always-$+1$ rule achieves perfect SCA without estimating either count, which is why we also report endpoint and joint count correctness. For continuous quantities, SCA gives partial credit for smaller errors and clips errors at or above the reference-change magnitude to zero. We also report change error in meters or square meters. Main SCA evaluates reference-changing pairs, not whether answers remain unchanged when the reference is stable.

The signed errors obey

$$e_i^{\mathrm{late}} - e_i^{\mathrm{early}} = \Delta \hat{y}_i - \Delta y_i. \tag{3}$$

A correct increment leaves the signed error unchanged. If a room-count reference grows from 3 to 4, predictions of 4 and 5 preserve a one-room overestimate. Predictions of 4 and 4 miss the increment but become correct at the later endpoint.

Joint count correctness. We also measure whether both endpoint counts are exactly correct. For Distinct Rooms, Object-Containing Rooms, and Target Visits, let $c_i^t = \mathbf{1}[\hat{y}_i^t = y_i^t]$ indicate exact correctness at endpoint t, taking an invalid answer as incorrect. Let $\mathcal{P}_q$ be the prespecified eligible

pairs for source question q, and let $\mathcal{Q}$ be the source questions with at least one such pair. On this fixed denominator, joint count correctness is

$$p_{11} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{P}_q|} \sum_{i \in \mathcal{P}_q} \mathbf{1}[c_i^{\text{early}} = 1,\ c_i^{\text{late}} = 1]. \tag{4}$$

For count pairs with a $+1$ reference increment, SCA minus $100p_{11}$ equals the source-question-weighted percentage of pairs with a correct increment but incorrect counts at both endpoints. This identity requires the same pairs and weights for both scores.

Continuous numerical changes. To retain errors in meters or square meters, we also report change MAE. For each pair with valid numerical answers at both endpoints, the change error is $\Delta\text{Err}_i = |\Delta\hat{y}_i - \Delta y_i|$. Change MAE averages this error within source questions and then equally across questions, retaining the native unit of meters or square meters. The no-change rule has $\Delta\hat{y}_i = 0$; its change error is $|\Delta y_i|$. We compare its mean with a model's mean change error on exactly the same valid pairs and with the same source-question weights. The valid-pair subset is defined separately for each system and question type.

Post-hoc constant diagnostic. For each continuous type, we fit one constant to maximize source-question-averaged endpoint MRA on the reference-changing endpoints. Its in-sample score measures the endpoint credit obtainable with one fixed answer.

4 Experiments

4.1 Experimental Setup

We evaluate 30 primary systems across three groups: 14 general-purpose video MLLMs, including Qwen-VL and InternVL3; 14 spatially specialized video MLLMs, including SenseNova, Spatial-MLLM, and Cambrian; and 2 streaming video MLLMs, OASIS and Dispider. This comparison asks whether systems designed for spatial reasoning or streaming video make more accurate spatial judgments than general-purpose models.

For the primary evaluation, we sample RGB frames at 1 FPS from the traversal start through each queried endpoint. The progressive evaluation contains 1,411 queried prefixes from 386 source questions. The main progressive tables show representatives of these model groups; Appendix D reports the complete results.

For continuous quantities, we compare predicted changes with a video-free prior. For each held-out home, it predicts the median signed reference change from other homes of the same question type and cohort. This provides a control for how much change can be predicted without observing the video.

4.2 Full-Traversal Evaluation

4.2.1 Performance by Question Group

Models more often exceed their answer-frequency baseline on place/time binding than on observed-route graph questions. Among the 30 primary systems, 22 exceed the baseline on place/time binding, compared with 6 on observed-route graph questions (Table 2). The pattern extends to Fixed-Axis Bearing: only 5 exceed the 13.33% frequency baseline on the direction between landmark viewpoints relative to the initial camera orientation.

4.2.2 Differences across Systems

Model ordering also depends on the numerical task. Qwen2.5-VL-72B exceeds Qwen3-VL-32B on motion distance, at 39.85 versus 10.43 MRA, but the ordering reverses on area, at 22.57 versus 49.50.

Table 2: **Full-traversal spatial QA.** Primary systems use 1-FPS inputs. Cum., Mot., Place, and Route denote cumulative spatial state, motion distance, place/time binding, and observed-route graph. Floor and Bearing denote Floor Offset and Fixed-Axis Bearing. Avg. averages question types within each answer format; All mixes MRA and accuracy over all 18 types and is descriptive only. The best, second-best, and third-best scores in each format column are highlighted. *Direct-Me adaptation with a Qwen3-VL-8B readout.

Model	Numerical (*MRA*)					Choice (*Acc.*)					All
	Cum.	Floor	Mot.	Area	Avg.	Place	Route	Bearing	Area	Avg.	Avg.
Uniform random	–	–	–	–	–	33.33	25.00	12.50	25.00	26.25	–
Frequency-sampled	–	–	–	–	–	34.26	36.78	13.33	28.22	32.83	–
General-purpose video MLLMs											
Qwen2.5-VL-3B	33.16	19.21	23.35	27.45	27.54	46.84	21.45	7.02	26.67	28.14	27.87
Qwen2.5-VL-7B	33.19	33.68	34.76	36.20	34.40	52.03	37.57	7.02	46.67	39.76	37.38
Qwen2.5-VL-72B	40.43	**47.63**	**39.85**	22.57	**36.72**	53.97	37.37	**21.05**	70.00	43.98	40.75
Qwen3-VL-2B	43.77	2.89	2.71	8.89	19.67	52.17	26.01	10.53	30.00	32.71	26.92
Qwen3-VL-4B	48.76	35.53	5.21	29.18	31.32	57.08	31.18	10.53	43.33	38.10	35.09
Qwen3-VL-8B	48.34	30.00	5.62	30.00	30.78	56.07	29.88	15.79	50.00	38.34	34.98
Qwen3-VL-32B	39.79	19.47	10.43	**49.50**	32.34	**62.17**	41.81	10.53	70.00	47.61	40.82
InternVL3-1B	35.18	16.58	3.23	12.95	19.31	39.80	27.70	10.53	13.33	28.18	24.24
InternVL3-2B	49.52	0.00	5.83	27.86	26.99	36.13	22.79	19.30	23.33	26.50	26.72
InternVL3-8B	31.52	7.89	11.76	34.61	24.40	51.49	23.82	10.53	40.00	32.41	28.85
InternVL3-14B	**50.62**	26.05	4.38	19.02	28.09	48.98	30.41	5.26	66.67	37.09	33.09
InternVL3-38B	46.89	38.95	10.62	33.20	33.41	60.45	42.84	10.53	**76.67**	**48.27**	41.67
LLaVA-OneVision-7B	16.24	5.53	14.04	12.98	13.54	21.89	13.51	5.26	13.33	15.18	14.45
LLaVA-NeXT-Video-7B-Qwen2	19.95	2.63	22.28	12.80	16.58	27.12	17.70	7.02	26.67	20.36	18.68
Spatially specialized video MLLMs											
VG-LLM-4B	24.69	18.95	9.27	21.16	19.23	49.43	24.00	3.51	53.33	32.51	26.61
VG-LLM-8B	23.15	13.68	11.56	22.50	18.91	48.22	23.87	12.28	50.00	32.63	26.53
VLM-3R	15.58	3.95	1.25	10.16	9.19	21.76	13.09	5.26	23.33	15.93	12.93
Spatial-MLLM-v1.1	37.63	0.00	5.43	9.07	17.74	48.57	26.16	5.26	43.33	32.51	25.95
Cambrian-S-0.5B	11.67	0.26	0.42	2.39	5.11	9.51	8.96	1.75	6.67	8.17	6.81
Cambrian-S-1.5B	9.79	0.26	0.00	3.82	4.66	15.26	8.07	1.75	6.67	9.46	7.32
Cambrian-S-3B	8.36	5.53	0.79	10.48	6.64	12.81	9.13	3.51	6.67	9.43	8.19
Cambrian-S-7B	10.06	5.53	0.73	10.23	7.20	14.71	11.71	1.75	13.33	11.78	9.74
Cambrian-P-7B	38.42	3.95	3.85	40.25	25.93	51.59	23.96	14.04	**76.67**	36.52	31.81
SenseNova-SI-1.1 (InternVL3-2B)	39.71	24.47	5.62	16.02	23.36	45.60	25.06	7.02	36.67	30.58	27.37
SenseNova-SI-1.5 (InternVL3-8B)	35.88	15.00	5.52	14.77	20.40	49.99	33.87	5.26	50.00	37.46	29.88
SenseNova-SI-1.1 (Qwen2.5-VL-3B)	43.01	0.26	6.67	23.34	23.66	56.97	29.54	7.02	53.33	37.90	31.57
SenseNova-SI-1.1 (Qwen3-VL-8B)	47.70	19.21	11.38	39.36	32.98	61.33	37.57	14.04	53.33	43.92	39.06
SenseNova-SI-1.3 (Qwen3-VL-8B)	34.98	10.53	9.45	33.82	25.25	62.07	**47.46**	5.26	50.00	47.88	37.82
Streaming MLLMs											
Stream3D-VLM-4B (Qwen2.5-VL-3B)	7.88	2.11	1.04	0.23	3.54	31.76	29.86	10.53	53.33	30.85	18.71
Direct-Me* (Qwen3-VL-8B)	4.48	0.00	2.81	4.43	3.49	45.74	26.94	8.77	33.33	31.40	19.00
OASIS	37.43	0.00	5.73	17.55	19.86	13.33	0.00	0.00	0.00	4.00	11.05
Dispider	12.42	2.63	15.83	11.70	11.87	41.33	23.55	7.02	43.33	29.21	21.50

Across the 5 matched base–SenseNova comparisons, all SenseNova variants have higher choice accuracy, while 4 have lower numerical MRA. These comparisons concern complete systems within matched backbone families. Appendix C gives additional type-level results with 128 uniformly sampled frames.

4.3 Progressive Spatial Estimation

To examine estimates before the traversal ends, we compare answers to the same cumulative question at two independently queried prefixes (Table 3). Here, late denotes the second endpoint of a pair, not necessarily the end of the traversal.

4.3.1 Accuracy across Observation Prefixes

Accuracy as observations accumulate depends on the quantity being estimated. Qwen3-VL-8B's Distinct Rooms MRA increases from 41.4 to 52.3, whereas its Target Visits MRA decreases from 64.3 to 38.4. For Target Visits, exact accuracy also decreases from 58.7% at the earlier endpoint to 24.5% at the later endpoint; the decline is therefore present under both scoring rules. A longer prefix provides more observations but also requires accounting for an additional visit. The results describe accuracy on a growing spatial quantity, combining additional observations with a changed target value.

Table 3: **Endpoint and change accuracy on reference-changing pairs.** Early and Late report endpoint MRA, and SCA reports change accuracy, all on a 0–100 scale. Object Rooms abbreviates Object-containing Rooms. Scores use the same eligible pairs, averaged within source questions and then equally across questions; invalid responses score zero. Late is the second pair endpoint, not necessarily the traversal end. Primary systems use 1 FPS.

	Count questions									**Continuous quantities**					
	Distinct Rooms			Object Rooms			Target Visits			Path Length			Visited Floor Area		
System / input	Early	Late	SCA	Early	Late	SCA	Early	Late	SCA	Early	Late	SCA	Early	Late	SCA
Qwen2.5-VL-72B	45.5	52.6	45.5	57.7	54.3	47.4	56.6	20.2	17.1	59.8	51.1	32.7	3.2	6.4	12.1
Qwen3-VL-8B	41.4	52.3	55.3	46.3	53.7	40.3	64.3	38.4	22.5	0.2	0.6	7.8	61.0	53.7	21.8
InternVL3-38B	43.3	55.9	49.2	61.4	56.3	43.8	59.4	31.7	27.3	8.8	1.1	7.3	17.1	21.0	3.3
SenseNova-SI-1.3	23.1	29.8	34.9	33.1	42.1	32.0	47.7	45.7	15.9	2.8	2.5	2.4	9.6	21.0	9.1
Cambrian-S-7B	7.4	8.6	20.0	38.9	48.6	31.5	38.8	64.3	13.1	3.8	0.2	4.4	60.0	51.5	18.5
Cambrian-P-7B	3.3	5.4	10.1	17.5	41.6	15.9	34.1	61.4	5.9	0.1	0.0	2.0	70.7	71.8	12.7
OASIS	44.0	54.1	40.2	56.4	52.6	40.5	47.0	16.6	15.9	1.9	0.3	12.4	46.6	35.3	14.7
Dispider	1.7	1.2	8.1	1.3	1.5	0.0	7.3	3.8	4.2	16.4	16.9	11.2	5.6	8.8	6.7
Stream3D-VLM-4B	3.4	4.3	2.9	33.8	16.7	7.2	45.9	9.7	0.0	0.0	0.0	1.9	0.0	1.2	6.5
Direct-Me bridge	24.1	2.6	0.4	17.8	12.4	3.6	3.8	0.9	4.6	0.0	0.0	1.2	0.0	0.0	3.5

Endpoint accuracy does not establish whether the change between prefixes was estimated correctly. On Visited Floor Area, Cambrian-P-7B obtains 71.8 late MRA but only 12.7 SCA. An estimate can be close to the accumulated area while its change misses the amount added between endpoints. The post-hoc constant-answer diagnostic in Appendix F.4 quantifies endpoint MRA attainable with a fixed prediction under the observed reference distribution. Appendix D reports the complete 1-FPS results, including early, late, and joint exact-count correctness.

4.3.2 Count Offsets and Change-Magnitude Errors

Correct increments can coexist with count errors at both endpoints. On Distinct Rooms, Qwen3-VL-8B scores 55.3 SCA but only 13.1% joint count correctness (Tables 3–4). Because every reference increment is $+1$, the difference is a 42.2% weighted share of pairs with a correct increment but incorrect counts at both endpoints. The corresponding shares, computed from unrounded scores, are 30.3% for Qwen2.5-VL-72B and 34.5% for InternVL3-38B. On these pairs, the model overestimates or underestimates both counts by the same amount (Equation 3). Appendix F.5 reports additional correctness and repetition analyses on valid-response pairs.

Table 4: **Joint count correctness and change error.** DR/OCR/TV denote Distinct Rooms, Object-Containing Rooms, and Target Visits; PL/VFA denote Path Length and Visited Floor Area. Counts use fixed pairs. Change MAE (Δ) and no-change MAE (0Δ) use each system's valid pairs; the video-free prior uses all eligible pairs.

System	DR	OCR	TV	PL (m)		VFA (m^2)	
	$100p_{11}$ (%)			ΔMAE	0Δ	ΔMAE	0Δ
Qwen2.5-VL-72B	15.2	27.1	8.9	23.6	29.9	8.5	8.5
Qwen3-VL-8B	13.1	20.6	14.9	28.2	29.9	9.0	8.5
InternVL3-38B	14.6	28.2	15.3	28.6	29.9	14.9	8.5
SenseNova-SI-1.3	6.2	11.6	10.4	30.2	29.9	9.9	8.5
Cambrian-S-7B	0.3	15.4	8.0	29.0	29.9	9.8	8.5
Cambrian-P-7B	0.0	1.9	2.1	29.5	29.9	7.8	8.5
OASIS	14.5	23.9	7.8	26.8	30.0	11.1	8.5
Stream3D-VLM-4B	0.0	3.9	0.0	29.9	29.9	10.8	8.5
Direct-Me bridge	0.0	2.5	0.0	102.4	30.1	7.8	8.5
Video-free prior	–	–	–	9.0	29.9	4.6	8.5

For distance and area, the issue extends to estimating how much has been added. We compare change MAE with no-change and video-free predictions. On Path Length, Qwen2.5-VL-72B improves over no change (23.6 versus 29.9 m change MAE), while Qwen3-VL-8B remains closer to that baseline at 28.2 m. Both have larger errors than the video-free prior's 9.0 m. On Visited Floor Area, the prior has 4.6 m^2 change MAE versus 9.0 m^2 for Qwen3-VL-8B. Improving over no change does not ensure a more accurate change estimate than a prior computed without video.

4.4 Frame Selection and Budget

We test whether retaining more visual history reduces these errors. Our RIVER-inspired condition causally selects at most K RGB frames per prefix. It adapts the long–short-term retention principle of RIVER (Shi et al., 2026) to frame selection; QA calls remain independent.

Within a backbone, $K = 32$ versus $K = 64$ changes the budget of the same selector. The 1-FPS condition changes both frame count and selection policy; short prefixes can contain fewer than K sampled frames. Table 5 uses the same pairs as Table 3.

For Qwen3-VL-8B, increasing K from 32 to 64 raises Distinct Rooms SCA by +9.1 points (95% paired home-bootstrap interval [+4.4, +13.6]). Joint count correctness increases by +0.8 percentage points, with interval [-1.7, +4.1]. The evidence supports higher increment agreement, while the change in joint count correctness remains uncertain.

The outcome also differs across models and quantities. For Cambrian-S-7B, Distinct Rooms SCA increases from 15.7 to 20.9 while joint count correctness decreases from 0.8% to 0.3%; Cambrian-P-7B remains at 10.1 SCA and 0.0% joint count correctness. On Visited Floor Area, Qwen3-VL-8B's late MRA increases from 33.6 to 48.2 while SCA decreases from 16.6 to 13.7. A larger retained-frame budget thus does not consistently improve both endpoint and change scores in these conditions. Appendix D.1 reports the wider K sweep; Appendix C.1 reports the separate fixed-token full-traversal comparison.

Table 5: **Frame selection and budget on the pairs in Table 3.** Task cells report late MRA / SCA (0–100); DR both exact is joint Distinct Rooms correctness $100p_{11}$ (%). K bounds causally selected RGB frames.

Backbone	Input	Distinct Rooms	Object-containing Rooms	Target Visits	Path Length	Visited Floor Area	DR both exact
Qwen3-VL-8B	1 FPS	52.3/55.3	53.7/40.3	38.4/22.5	0.6/7.8	53.7/21.8	13.1
Qwen3-VL-8B	$K = 32$	55.9/43.1	51.0/37.4	38.9/26.8	0.0/5.2	33.6/16.6	12.3
Qwen3-VL-8B	$K = 64$	52.2/52.2	57.0/40.7	35.9/24.7	0.0/6.2	48.2/13.7	13.1
Cambrian-S-7B	1 FPS	8.6/20.0	48.6/31.5	64.3/13.1	0.2/4.4	51.5/18.5	0.3
Cambrian-S-7B	$K = 32$	10.4/15.7	51.8/26.1	61.9/10.1	0.2/2.4	46.4/14.5	0.8
Cambrian-S-7B	$K = 64$	7.3/20.9	49.4/32.0	65.5/14.0	0.2/5.1	49.5/11.8	0.3
Cambrian-P-7B	1 FPS	5.4/10.1	41.6/15.9	61.4/5.9	0.0/2.0	71.8/12.7	0.0
Cambrian-P-7B	$K = 32$	5.5/10.1	43.8/14.1	62.9/3.3	0.0/1.8	65.1/14.2	0.0
Cambrian-P-7B	$K = 64$	5.5/10.1	43.0/18.1	63.5/4.3	0.0/2.3	72.4/15.5	0.0

5 Discussion

Correct increments can coexist with the same signed count error at both independently queried endpoints. For distance and area, several systems improve over no change yet have larger change error than the video-free prior. Accurate spatial updates therefore require both the accumulated quantity and the amount added between observations.

Retaining more frames raises increment agreement in some comparisons without consistently raising joint count correctness. The additional observations can improve increment agreement while errors in the accumulated state remain.

6 Conclusion

TRACE-VSI combines scan-grounded full-traversal spatial QA with progressive evaluation across video prefixes. In full-traversal QA, more systems exceed the answer-frequency baseline on place/time binding than on observed-route graph or fixed-axis bearing questions. Across prefixes, correct increments can coexist with incorrect endpoint counts, while distance and area change errors exceed those of a video-free prior. Larger retained-frame budgets improve some increment judgments but do not consistently improve endpoint and change scores together.

The current progressive evaluation covers cumulative quantities in prerecorded indoor traversals. Future work can extend it to changing place and route relations and to active exploration.

REFERENCES

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. ReMEmbR: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. arXiv preprint arXiv:2409.13682, 2024.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418, 2024.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3R: Vision-language models augmented with instruction-aligned 3D reconstruction. arXiv preprint arXiv:2505.20279, 2025.

Muhammad Fadhil Ginting, Dong-Ki Kim, Xiangyun Meng, Andrzej Reinke, Bandi Jai Krishna, Navid Kayhani, Oriana Peltzer, David D. Fan, Amirreza Shaban, Sung-Kyun Kim, Mykel J. Kochenderfer, Ali-akbar Agha-mohammadi, and Shayegan Omidshafiei. Enter the mind palace: Reasoning and planning for long-term active embodied question answering. arXiv preprint arXiv:2507.12846, 2025.

Nicolas Gorlo, Lukas Schmid, and Luca Carlone. Describe anything anywhere at any moment. arXiv preprint arXiv:2512.00565, 2025.

Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. RoomTour3D: Geometry-aware video-instruction tuning for embodied navigation. arXiv preprint arXiv:2412.08591, 2024.

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

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: OVBench and VideoChat-Online. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3328–3338, 2025.

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. SceneVerse: Scaling 3D vision-language learning for grounded scene understanding. In *European Conference on Computer Vision (ECCV)*, 2024.

Xinqi Jin, Hanxun Yu, Bohan Yu, Kebin Liu, Jian Liu, Keda Tao, Yixuan Pei, Huan Wang, Fan Dang, Jiangchuan Liu, and Weiqiang Wang. StreamingAssistant: Efficient visual token pruning for accelerating online video understanding. arXiv preprint arXiv:2512.12560, 2025.

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

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. EgoSAT: A comprehensive benchmark of egocentric streaming interaction understanding. In *European Conference on Computer Vision (ECCV)*, 2026.

Chuhan Li, Ruilin Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. Learning situated awareness in the real world. arXiv preprint arXiv:2602.16682, 2026.

Xinying Lin, Xuyang Liu, Yiyu Wang, Teng Ma, and Wenqi Ren. V-CAST: Video curvature-aware spatio-temporal pruning for efficient video large language models. arXiv preprint arXiv:2603.27650, 2026.

Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M. Alvarez. StreamChat: Chatting with streaming video. arXiv preprint arXiv:2412.08646, 2024.

Zhaochen Liu, Limeng Qiao, Guanglu Wan, and Tingting Jiang. GeoAlign: Geometric feature realignment for MLLM spatial reasoning. arXiv preprint arXiv:2604.12630, 2026.

Ruiyuan Lyu, Jingli Lin, Tai Wang, Shuai Yang, Xiaohan Mao, Yilun Chen, Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, and Jiangmiao Pang. MMScan: A multi-modal 3D scene dataset with hierarchical grounded language annotations. arXiv preprint arXiv:2406.09401, 2024.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *International Conference on Learning Representations (ICLR)*, 2023.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498, 2024.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M. Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Saumya Saxena, Blake Buchanan, Chris Paxton, Peiqi Liu, Bingqing Chen, Narunas Vaskevicius, Luigi Palmieri, Jonathan Francis, and Oliver Kroemer. GraphEQA: Using 3D semantic scene graphs for real-time embodied question answering. arXiv preprint arXiv:2412.14480, 2024.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. FastVID: Dynamic density pruning for fast video large language models. In *Advances in Neural Information Processing Systems*, 2025.

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. RIVER: A real-time interaction benchmark for video LLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Alexander W. Siegel and Sheldon H. White. The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior*, 10:9–55, 1975. doi: 10.1016/S0065-2407(08)60007-5.

Tin Stribor Sohn, Maximilian Dillitzer, Jason J. Corso, and Eric Sax. R4: Retrieval-augmented reasoning for vision-language models in 4D spatio-temporal space. arXiv preprint arXiv:2512.15940, 2025.

Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, and Xiangyu Yue. SpaceVista: All-scale visual spatial reasoning from mm to km. arXiv preprint arXiv:2510.09606, 2025.

Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. doi: 10.1037/h0061626.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. EmbodiedScan: A holistic multi-modal 3D perception suite towards embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep it in mind: User centric continual spatial intelligence reasoning in egocentric video streams. In *International Conference on Machine Learning (ICML)*, 2026.

Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation. In *Proceedings of Robotics: Science and Systems*, 2024.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747, 2025a.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. SpatialScore: Towards comprehensive evaluation for spatial intelligence. arXiv preprint arXiv:2505.17012, 2025b.

Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. ST-Think: How multimodal large language models reason about 4D worlds from ego-centric videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5174–5183, 2026.

Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. VideoLLM-MoD: Efficient video-language streaming with mixture-of-depths vision computation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109922–109947, 2024.

Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-RAG: General non-parametric embodied memory for retrieval and generation. arXiv preprint arXiv:2409.18313, 2024.

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

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

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764, 2025b.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3D-Mem: 3D scene memory for embodied exploration and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. TimeChat-Online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025.

Hanxun Yu, Xuan Qu, Lei Ke, Boqiang Zhang, Yuxin Wang, Jianke Zhu, and Dong Yu. Stream3D-VLM: Online 3D spatial understanding with incremental geometry priors. arXiv preprint arXiv:2606.06891, 2026.

Tatiana Zemskova and Dmitry Yudin. 3DGraphLLM: Combining semantic graphs and large language models for 3D scene understanding. arXiv preprint arXiv:2412.18450, 2024.

Le Zhang, Jihan Yang, Soundarya Krishnan, Jimit Majmudar, Xiou Ge, Prasoon Puri, Prathamesh Saraf, Shruti Bhargava, Dhivya Piraviperumal, Yinan Ling, Cindy Pan, Hong Yu, Aishwarya Agrawal, and Bo-Hsiang Tseng. From where things are to what they are for: Benchmarking spatial-functional intelligence in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12052–12063, 2026.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3D world: Enhancing MLLMs with 3D vision geometry priors. arXiv preprint arXiv:2505.24625, 2025a.

Hongpei Zheng, Shijie Li, Yanran Li, and Hujun Yin. SpatialReasoner: Active perception for large-scale 3D scene understanding. arXiv preprint arXiv:2512.03284, 2025b.

Xinyi Zheng, Yunze Liu, Chi-Hao Wu, Fan Zhang, Hao Zheng, Wenqi Zhou, Walterio W. Mayol-Cuevas, and Junxiao Shen. SpatialMem: Metric-aligned long-horizon video memory for language grounding and QA. arXiv preprint arXiv:2601.14895, 2026.

Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. LSceneLLM: Enhancing large 3D scene understanding using adaptive visual preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3761–3771, 2025.

A Question Construction and Source Records

A.1 Construction rules

Table 6 summarizes how each question type is constructed and scored. Each type has a fixed target, response format, scoring rule, and eligible endpoint range.

Question type	Response	Key construction
Distinct Rooms	integer	Number of rooms and hallways the camera has entered by the endpoint
Object-Containing Rooms	integer	Number of distinct places where the target category is clearly visible
Target Visits	integer	Number of visits to a target place, separated by leaving and re-entering
Object-to-Room Binding	grounded MCQ	Place containing one visually distinguished object instance
Same-Room Episodes	binary	Whether two observed episodes occurred in the same place
First-Seen Order	ranking MCQ	Order in which a fixed set of object categories first clearly appeared
Floor Offset	signed integer	Floor difference between a fixed pair of places, derived from verified transitions between floors
Direct Traversal	grounded MCQ	Offered place pair that the camera traverses directly
Shortest-Route First Hop	grounded MCQ	First place after the source on a shortest route in the observed place-transition graph
Mandatory Place	grounded MCQ	Intermediate place shared by every simple route between fixed source and target places in the observed place-transition graph
Destination-Cost Order	ranking MCQ	Ordering, with possible ties, of destinations by the fewest observed transitions from one source
Itinerary-Cost Order	ranking MCQ	Ranking of ordered multi-goal routes by their total transition count
Fixed-Axis Bearing	8-way direction	Bearing between two anchors relative to a fixed viewing axis
Path Length	numerical	Three-dimensional arc length of the camera trajectory to the endpoint
Shortest Observed Trip	numerical	Minimum camera-path length among completed trips between two places
Visited Floor Area	numerical	Total floor area of rooms and hallways counted as visited
Room Area	numerical	Floor area of a fixed room under the benchmark's geometry convention
Room-Area Rank	ranking MCQ	Ranking of fixed rooms by floor area

Table 6: **Question-construction rules.** Numerical response formats include units, answer-format rules, and tolerances in the benchmark specification. Multiple-choice formats store canonical option values separately from their displayed positions.

A.2 Reference construction

Reference answers are derived from scan annotations, camera poses, and time-indexed events. Each question specifies its target entities and relevant source observations. Progressive pairs link the same question across two observation endpoints, each with its corresponding reference answer.

A.3 Inputs and inference

Each request contains one question and RGB frames sampled through one endpoint. Earlier and later requests are independent, and neither includes the response to the other. Each response is associated with its question and observation endpoint for paired evaluation. Table 7 distinguishes the two sampling schemes and their evaluation roles.

Condition	Frame selection	Evaluation role
Fixed budget (K frames)	The benchmark selects K chronological frames independently from each source prefix.	Full-traversal evaluation under a common frame budget and progressive input-condition comparisons; both include sampler and model behavior.
Fixed rate (1 fps)	Frames are supplied at a fixed temporal density of one frame per second, so the input grows with the prefix.	Primary full-traversal and progressive evaluations at fixed temporal density; the latter compares growing prefixes.

Table 7: **RGB frame selection and evaluation scope.** Each request draws from one source RGB prefix. The conditions differ in temporal sampling and the resulting frame count.

Prompt and decoding. Each question uses the same system instruction: *"You are given ordered RGB frames sampled from one complete egocentric traversal video. The frames are in chronological order, but adjacent frames may have non-uniform time gaps. Answer only from the visual evidence in*

Table 9: **Frozen reference-changing progressive cohort.** Counts are fixed before model scoring. The queried pool contains 1,411 distinct endpoints; 1,310 distinct endpoints participate in these 953 eligible pairs. Reused endpoints are counted within each pair for pairwise metrics.

Question type	Pairs	Source questions	Homes
Distinct Rooms	384	84	22
Object-containing Rooms	185	81	22
Target Visits	118	48	16
Path Length	180	90	25
Visited Floor Area	86	40	11

the traversal." The user message contains the question followed by an answer-format instruction. It requests one integer, signed integer, numerical value, Yes/No answer, or multiple-choice letter as required by the question type, without an explanation. All reported evaluations use greedy decoding (`do_sample=false`, one beam) with at most 32 new tokens. Full-traversal and progressive responses are scored with the recorded rule for each question type and evaluation.

Model inputs. Each model receives chronological RGB frames through its supported video or multi-image interface. Questions are evaluated independently without conversational history. Appendix F defines the scoring rules.

B Dataset Composition and Reference-Changing Pairs

Table 8 distinguishes full-traversal questions from the pre-pairing inventory of queried prefixes and the source questions from which those prefixes are derived. Table 9 separately reports the source questions that contribute at least one eligible consecutive-endpoint pair.

Table 8: **TRACE-VSI data statistics.** Full-traversal questions are evaluated once after the complete traversal. The intermediate-prefix row is a pre-pairing inventory: one source question can yield several queried prefixes, each with a reference that differs from its full-traversal reference. Durations summarize unique traversal videos or queried prefixes. MCQ chance gives row-weighted uniform and frequency-sampled baselines over 4- and 8-choice questions.

	Evaluated inputs	QA types	Scenes	Source traversals	Distinct source questions	Prefix inputs per source question	Med. dur. [IQR] (min)	MCQ rows	MCQ chance (uniform/freq.)
Full-traversal evaluation	950	18	25	90	950	–	2.75 [2.11, 3.42]	351	23.0% / 23.5%
Intermediate-prefix corpus	1,411	5	25	90	386	3.66	0.95 [0.47, 1.61]	0	–

Reference-changing pair selection. For each source question, eligible endpoints are ordered chronologically and paired consecutively. Retained pairs have different endpoint references. For the three count types and Visited Floor Area, the interval contains exactly one annotated reference-changing event, whose recorded post-event answer equals the later reference. Path Length has no semantic-event requirement. A source question can contribute more than one pair.

Figure 2 reports the 18 raw QA types and distinguishes the 1,411 intermediate-prefix inputs from their 386 source questions. Open numerical formats have no synthetic chance floor. For the 351 MCQ rows, Figure 3 therefore reports uniform chance, frequency-sampled accuracy, and a majority-position baseline separately by QA type. These quantities characterize the correct-option distribution of each subset.

The reference-changing inputs have a median duration of 0.95 [0.47, 1.61] minutes, where brackets give the interquartile range, and retain a median 34.5% of their paired full-traversal videos. Figure 4 reports the duration and retained-fraction distributions.

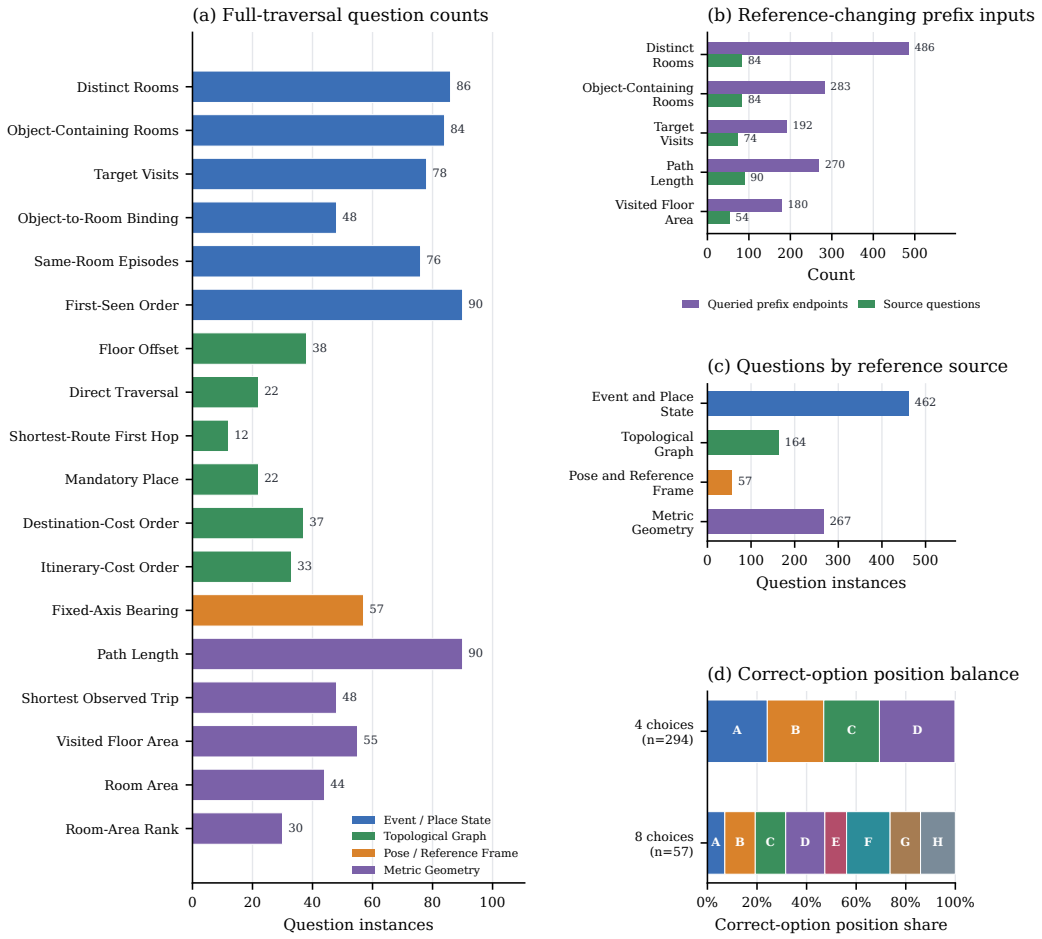


Figure 2: **TRACE-VSI evaluation data.** Panel (a) reports the full-traversal QA count for every raw type. Panel (b) inventories queried prefix endpoints and distinct source questions before consecutive-endpoint pairs are formed. Panel (c) groups full-traversal instances by the source of their reference answers. Panel (d) shows the correct-option position distribution for 4- and 8-choice questions. The reference structures in panel (c) are not supplied as model inputs.

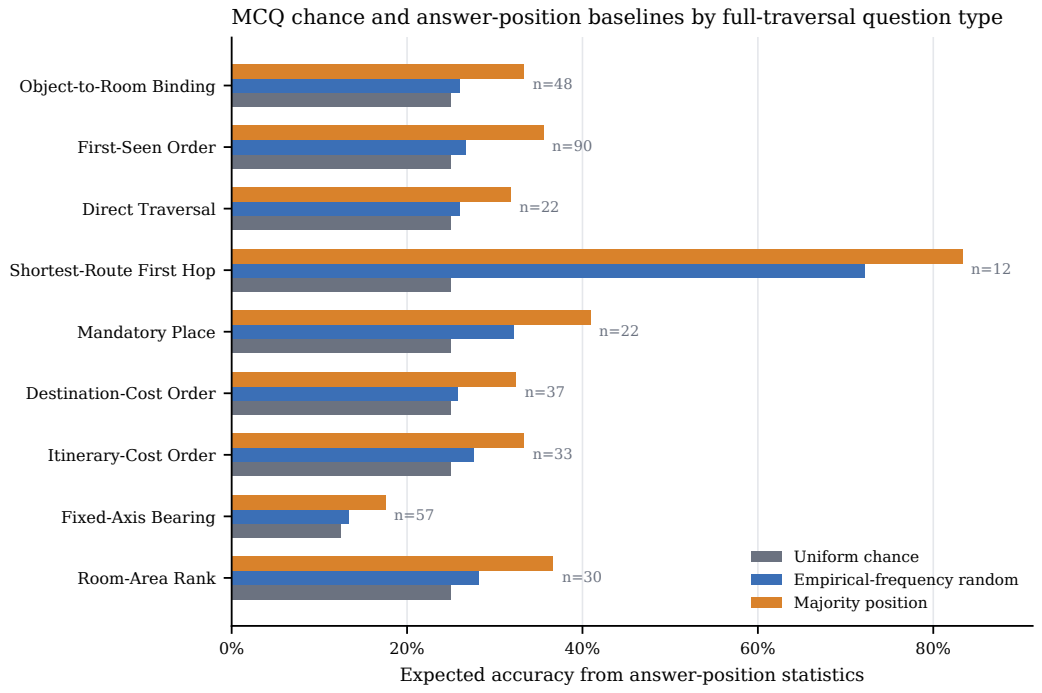


Figure 3: **Answer-position baselines for every MCQ type.** Uniform accuracy is $1/K$ for each row. Frequency-sampled accuracy is $\sum_j p_j^2$, where p_j is the type-specific empirical frequency of correct position j. The majority-position baseline always predicts the most frequent correct position for that type. The figure covers categorical answer formats.

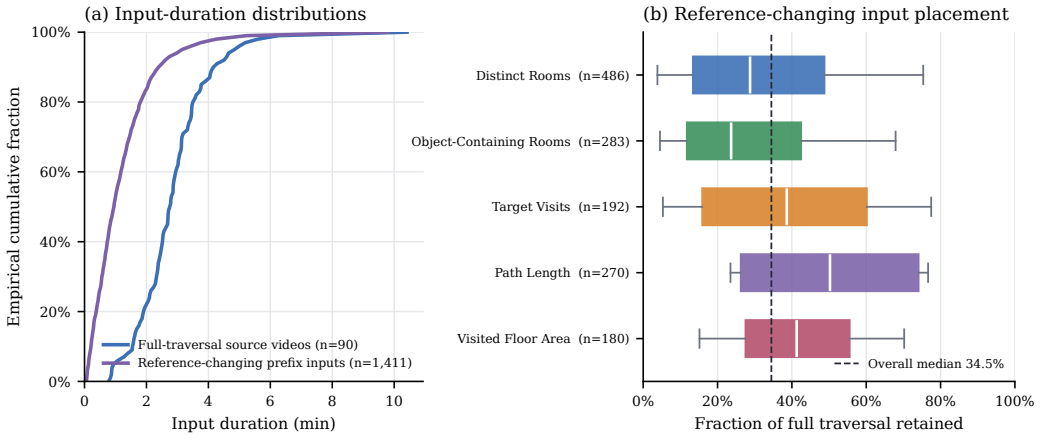


Figure 4: **Duration and placement of reference-changing inputs.** Panel (a) compares the empirical duration distributions of the 90 full-traversal videos and 1,411 independently evaluated prefixes. Panel (b) shows retained full-traversal duration by QA type. Boxes span the interquartile range, white lines mark medians, and whiskers span the 5th to 95th percentiles.

B.1 SAME-SEGMENT CONTROL COMPOSITION

The cross-system cohort contains 38 endpoint intervals drawn from 18 homes and 31 recorded trajectories. Each interval contributes Distinct Rooms and Object-Containing Rooms questions at the same two endpoints. The role counts are 12 stable/stable, 12 Distinct-Rooms-changed, 12 Object-Containing-Rooms-changed, and 2 changed/changed intervals. The primary mixed-role subset contains 24 intervals from 16 homes and 21 trajectories, with 12 intervals in each direction. Selection uses the reference records but not model responses.

Object-Containing Rooms counts distinct owner rooms in which the target object category has been clearly observed by the endpoint; it is not restricted to rooms entered by the camera. The stable/changed relation is obtained by replaying the fixed construction rule at both endpoints. The two question types and both endpoints share the same source-video interval and fixed-rate sampling schedule.

A balanced-cohort sensitivity retains the same 24 mixed-role intervals and adds 10 changed/changed intervals. It therefore contains 48 intervals from 18 homes and 34 trajectories, with 12 intervals in each of the four reference combinations. Qwen3-VL-8B is evaluated on this extension; its predictions on the shared mixed-role subset are identical to those in the cross-system cohort.

C ENDPOINT RESULTS BY RAW QA TYPE

Table 2 reports the primary full-traversal scores at one frame per second. This section provides type-level results under 128 uniformly sampled frames and comparisons across input conditions.

Under the 128-frame input condition, spatial specialization does not yield uniform gains across answer formats. Relative to their general-purpose counterparts, the SenseNova variants score higher on average choice accuracy in all 5 system comparisons but lower on numerical MRA in 4. Because the systems use different input interfaces and the two summaries cover different question families, this comparison does not isolate an effect of fine-tuning.

Figure 5 reports endpoint results with 128 uniformly sampled frames for all 18 raw question types. Each type retains its own accuracy or MRA score; the final column is an equal-weight mean across these different metrics.

Table 10 compares the equal-weight all-type average across four full-traversal sampling conditions for the 26 systems evaluated under all four conditions. The comparison changes frame selection and, for one-frame-per-second inputs, the delivered frame count.

C.1 TIMESTAMP COVERAGE AT A FIXED VISUAL-TOKEN BUDGET

Table 11 compares Qwen3-VL-8B input pipelines on 950 full-traversal questions. Uniform sampling passes 32 or 64 frames to the model. A VGGT-based selector examines 128 candidate timestamps and retains 32 or 64 frames that expand estimated spatial coverage (Wu et al., 2025a; Wang et al., 2025). V-CAST and FastVID encode 128 frames and compress their visual tokens (Lin et al., 2026; Shen et al., 2025). Matched budgets count visual tokens delivered to the language model, not the computation needed to select or encode frames. At 2,240 tokens, V-CAST and FastVID improve both room-count types over uniform sampling, but neither improves Target Visits; Path Length has zero MRA under every listed pipeline. These full-traversal results do not measure change across prefixes.

Table 12 varies the represented timestamps from 32 to 128 on a balanced 180-question subset while holding the tokens passed to the language model at 2,240.

The all-type average rises from 33.56 to 42.22. This trend is not uniform across answer formats: choice accuracy rises from 35.0 to 50.0, whereas numerical MRA varies between 30.25 and 32.50 after a starting value of 31.75. The mixed-score trend is therefore driven mainly by choice accuracy; the sweep does not show the same improvement for numerical estimation.

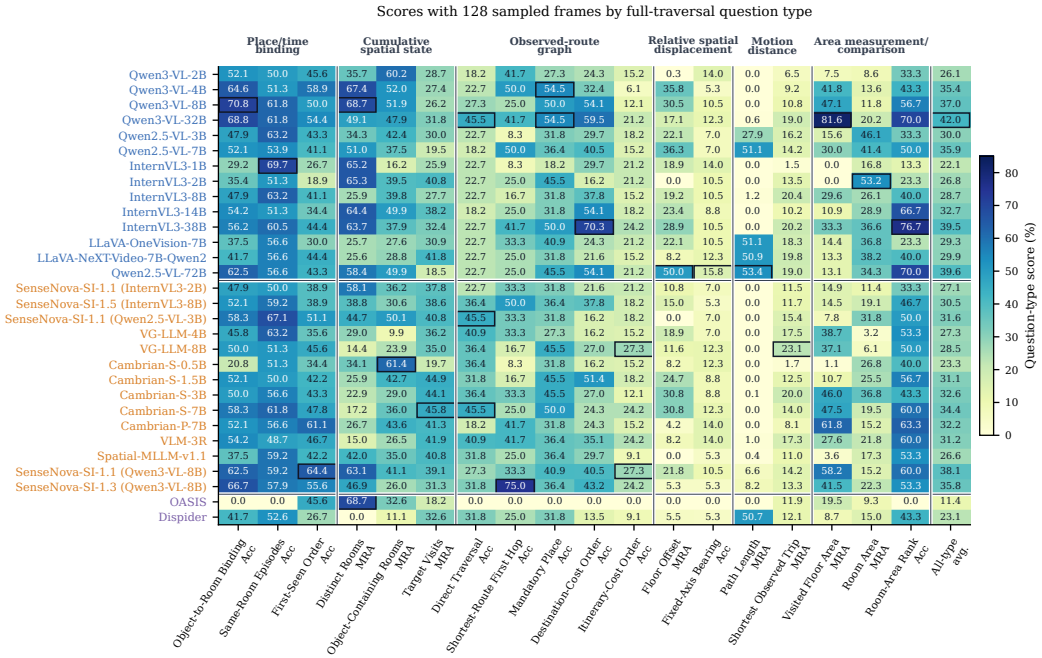


Figure 5: **Scores by question type with 128 uniformly sampled frames.** Cells report each question type's accuracy or MRA, with the equal-weight all-type average in the final column. Black outlines mark column-wise maxima. Each type retains its own metric, target distribution, and categorical baseline. Column bands follow the six question groups in Table 1; row bands group system classes.

C.2 Model size and frame budget

Figure 6 compares checkpoint variants within model families at three frame budgets. Figure 7 shows the 128-frame-minus-32-frame score separately for each raw question type and each of the 26 systems. Checkpoints can also differ in training data and architecture, so vertical comparisons at a fixed budget show within-family associations rather than controlled scaling effects. Frame-budget comparisons include the effect of resampling the input.

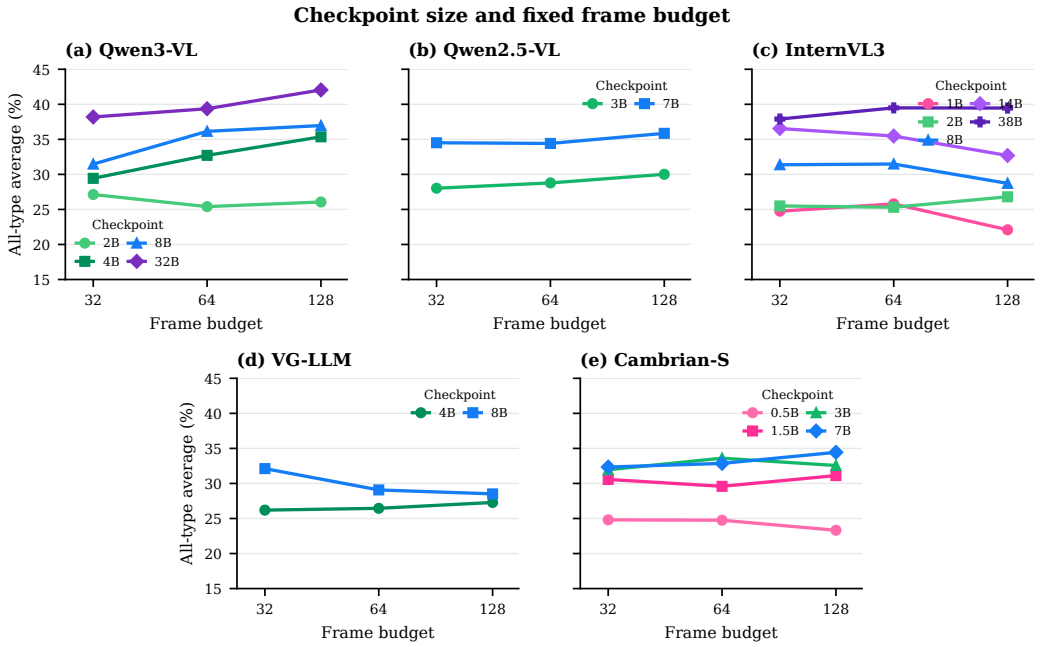


Figure 6: **Within-family checkpoint comparisons vary with the input condition.** Panels report the all-type average for checkpoint variants within the same model family. Lines connect the same checkpoint across frame budgets; vertical comparisons at a fixed budget show within-family checkpoint-size associations. Checkpoints also differ in training data, architecture, and other properties. The figure shows results at 32, 64, and 128 frames for the systems evaluated under all four input conditions.

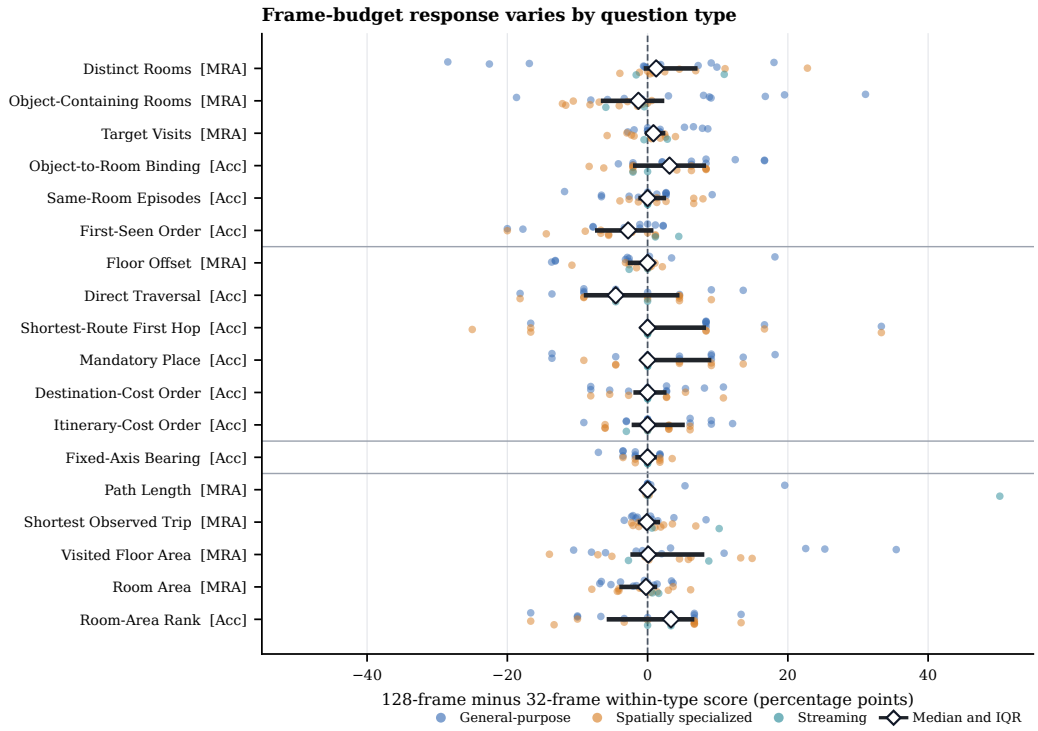


Figure 7: **Frame-budget changes vary across raw question types and systems.** Each row reports the question type's score with 128 frames minus its score with 32 frames for the 26 systems evaluated under all four reported conditions. Points are individual systems. Summary markers report the model-level distribution within the type. Positive values indicate a higher score with 128 frames. Because rows use either accuracy or MRA, these are within-type score changes rather than a common-unit ability difference.

Table 10: **Full-traversal scores across input conditions for 26 systems.** Each value is the unweighted mean of the 18 question-type scores on the same 950 questions. The first three conditions fix the frame count and sample uniformly over the traversal. The final condition samples one frame per second, so its frame count varies with traversal duration. All conditions retain the same question denominator.

Model	32 frames	64 frames	128 frames	1 frame/s
General-purpose video MLLMs				
Qwen3-VL-2B	27.13	25.40	26.06	26.92
Qwen3-VL-4B	29.43	32.71	35.36	35.09
Qwen3-VL-8B	31.47	36.14	36.96	34.98
Qwen3-VL-32B	38.19	39.37	42.05	40.82
Qwen2.5-VL-3B	28.03	28.78	30.01	27.87
Qwen2.5-VL-7B	34.51	34.42	35.85	37.38
InternVL3-1B	24.75	25.80	22.09	24.24
InternVL3-2B	25.50	25.31	26.80	26.72
InternVL3-8B	31.37	31.48	28.72	28.85
InternVL3-14B	36.55	35.47	32.69	33.09
InternVL3-38B	37.90	39.49	39.46	41.67
LLaVA-OneVision-7B	29.28	30.46	29.30	14.45
Spatially specialized video MLLMs				
SenseNova-SI-1.1 (InternVL3-2B)	28.05	28.37	27.14	27.37
SenseNova-SI-1.5 (InternVL3-8B)	32.27	30.51	30.51	29.88
SenseNova-SI-1.1 (Qwen2.5-VL-3B)	31.58	31.08	31.62	31.57
VG-LLM-4B	26.19	26.45	27.28	26.61
VG-LLM-8B	32.12	29.07	28.51	26.53
Cambrian-S-0.5B	24.80	24.75	23.32	6.81
Cambrian-S-1.5B	30.57	29.59	31.12	7.32
Cambrian-S-3B	31.96	33.60	32.56	8.19
Cambrian-S-7B	32.34	32.86	34.44	9.74
Cambrian-P-7B	32.42	32.68	32.18	31.81
VLM-3R	28.78	29.67	31.18	12.93
Spatial-MLLM-v1.1	25.74	27.70	26.65	25.95
Streaming video MLLMs				
OASIS	10.52	11.85	11.43	11.05
Dispider	20.09	23.19	23.14	21.50

Table 11: **Fixed-budget full-traversal results for five cumulative question types.** Qwen3-VL-8B scores are MRA (%). V-CAST (Lin et al., 2026) and FastVID (Shen et al., 2025) encode 128 frames before compression. Budgets count visual tokens passed to the language model. The dense row uses a larger budget and is shown for reference.

Input pipeline	Distinct Rooms	Object-containing Rooms	Target Visits	Path Length	Visited Floor Area
2,240 LM visual tokens					
Uniform (32 frames)	59.65	43.93	26.15	0.00	21.82
Geometry-selected (128→32)	59.77	42.62	30.00	0.00	29.27
V-CAST (128 frames)	71.05	53.81	24.49	0.00	28.91
FastVID (128 frames)	64.65	50.36	23.72	0.00	22.91
4,480 LM visual tokens					
Uniform (64 frames)	71.51	51.55	25.13	0.00	38.73
Geometry-selected (128→64)	70.00	47.62	26.15	0.00	42.91
V-CAST (128 frames)	74.07	54.17	24.49	0.00	42.00
Dense reference: 8,960 LM visual tokens					
Dense uniform (128 frames)	68.72	51.90	26.15	0.00	47.09

Table 12: **Timestamp sweep at a fixed downstream budget.** The balanced 180-question subset holds the tokens passed to Qwen3-VL-8B at 2,240 while varying the timestamps represented before V-CAST compression. Choice is exact accuracy, Num. is MRA, and Avg. is the equal-weight type-macro score.

Input condition	Time points	Pre-compression tokens	LM tokens	Choice	Num.	Avg.
Uniform	32	2,240	2,240	35.00	31.75	33.56
V-CAST	48	3,360	2,240	41.00	30.25	36.22
V-CAST	64	4,480	2,240	45.00	31.13	38.83
V-CAST	96	6,720	2,240	48.00	32.13	40.94
V-CAST	128	8,960	2,240	50.00	32.50	42.22

D PROGRESSIVE SCORE DETAILS

Table 13 gives endpoint MRA and SCA for all 30 primary 1-FPS systems. Scores use the fixed changed-reference pair set and source-question averaging as Tables 3–5.

Table 13: **Complete 1-FPS progressive endpoint and change scores.** Each cell gives early MRA / late MRA / SCA (0–100), source-question macro on the fixed reference-changing pairs. Invalid responses score zero. The main-paper subset appears in Table 3.

System / input	Distinct Rooms	Object-containing Rooms	Target Visits	Path Length	Visited Floor Area
Cambrian-P-7B	3.3/5.4/10.1	17.5/41.6/15.9	34.1/61.4/5.9	0.1/0.0/2.0	70.7/71.8/12.7
Cambrian-S-0.5B	18.0/18.5/9.6	0.0/0.0/0.0	35.4/27.4/4.3	2.9/0.1/5.4	0.9/0.8/2.6
Cambrian-S-1.5B	6.7/7.7/8.1	20.0/35.7/16.6	33.8/60.5/30.3	3.9/0.1/0.5	26.4/21.0/6.9
Cambrian-S-3B	10.3/12.0/21.8	13.2/27.7/30.3	29.4/56.0/8.2	9.8/1.4/12.0	29.6/27.7/4.0
Cambrian-S-7B	7.4/8.6/20.0	38.9/48.6/31.5	38.8/64.3/13.1	3.8/0.2/4.4	60.0/51.5/18.5
Dispider	1.7/1.2/8.1	1.3/1.5/0.0	7.3/3.8/4.2	16.4/16.9/11.2	5.6/8.8/6.7
InternVL3-14B	32.5/45.0/53.0	45.7/60.2/42.2	69.8/52.7/26.6	6.3/0.4/13.1	26.1/14.8/4.0
InternVL3-1B	24.7/38.8/32.6	37.3/30.6/18.6	43.4/42.2/10.8	0.1/0.4/11.5	0.0/0.0/0.6
InternVL3-2B	29.0/42.2/45.4	51.2/46.3/39.3	54.3/63.7/16.8	0.0/0.0/6.0	0.0/0.0/1.8
InternVL3-38B	43.3/55.9/49.2	61.4/56.3/43.8	59.4/31.7/27.3	8.8/1.1/7.3	17.1/21.0/3.3
InternVL3-8B	24.2/26.8/40.1	51.4/53.1/41.6	61.2/33.0/21.2	23.6/6.0/18.1	27.7/31.7/12.5
LLaVA-NeXT-Video-7B-Qwen2	11.9/12.7/15.3	20.3/37.5/32.6	52.4/57.6/15.8	8.5/29.9/0.4	0.8/3.8/0.0
LLaVA-OneVision-7B	5.7/5.1/17.7	25.3/31.0/31.7	43.1/49.5/9.5	7.2/26.1/5.3	2.2/6.3/0.0
OASIS	44.0/54.1/40.2	56.4/52.6/40.5	47.0/16.6/15.9	1.9/0.3/12.4	46.6/35.3/14.7
Qwen2.5-VL-3B	20.2/28.4/39.6	53.4/51.0/36.8	48.0/31.9/18.5	9.8/19.6/0.7	2.0/4.7/10.1
Qwen2.5-VL-72B	45.5/52.6/45.5	57.7/54.3/47.4	56.6/20.2/17.1	59.8/51.1/32.7	3.2/6.4/12.1
Qwen2.5-VL-7B	21.5/30.4/42.7	55.9/49.1/39.5	47.8/24.2/20.2	24.5/31.2/8.7	9.2/19.6/6.0
Qwen3-VL-2B	30.2/36.7/43.1	52.7/60.5/47.6	56.0/34.3/26.0	1.2/0.1/5.8	6.0/5.3/16.9
Qwen3-VL-32B	28.6/33.7/41.8	41.4/44.8/34.0	56.5/30.7/29.2	4.4/2.0/3.0	74.7/74.7/27.9
Qwen3-VL-4B	51.3/60.5/51.2	51.7/56.1/45.8	58.7/34.7/28.5	0.1/0.0/4.0	51.3/47.3/36.1
Qwen3-VL-8B	41.4/52.3/55.3	46.3/53.7/40.3	64.3/38.4/22.5	0.2/0.6/7.8	61.0/53.7/21.8
SenseNova-SI-1.1 (InternVL3-2B)	25.8/35.0/28.4	35.6/54.1/23.9	49.1/59.5/11.5	1.1/0.1/4.9	26.9/27.3/13.8
SenseNova-SI-1.1 (Qwen2.5-VL-3B)	14.5/20.8/19.5	26.9/52.0/20.6	36.9/62.3/0.5	5.8/0.4/0.9	3.5/7.9/9.5
SenseNova-SI-1.1 (Qwen3-VL-8B)	38.1/50.9/52.2	47.1/51.5/40.8	47.3/58.6/14.6	3.1/5.1/7.6	54.0/56.9/38.1
SenseNova-SI-1.3 (Qwen3-VL-8B)	23.1/29.8/34.9	33.1/42.1/32.0	47.7/45.7/15.9	2.8/2.5/2.4	9.6/21.0/9.1
SenseNova-SI-1.5 (InternVL3-8B)	20.4/27.5/23.5	43.9/52.1/29.0	39.1/59.9/3.8	0.2/0.0/1.1	46.3/31.0/9.1
Spatial-MLLM-v1.1	32.7/42.8/13.0	19.3/27.9/5.3	0.0/0.0/0.0	5.9/2.3/12.1	23.7/15.4/11.8
VG-LLM-4B	18.3/20.3/12.1	31.8/51.3/7.7	37.5/63.7/0.0	4.9/1.9/1.4	8.9/23.5/4.6
VG-LLM-8B	9.5/11.7/11.3	32.0/54.0/18.8	37.5/63.7/0.0	4.6/0.1/0.5	28.8/30.9/6.2
VLM-3R	9.4/11.5/20.7	14.3/32.3/21.6	36.4/62.2/0.0	15.9/2.0/6.1	7.8/10.5/13.8

Table 14 reports early, late, and joint exact-count correctness for every primary 1-FPS system on the same fixed reference-changing pair sets and source-question weights.

D.1 BOUNDED-FRAME SCORE DETAILS

Table 15 extends the within-backbone comparison to Cambrian $K = 16, 128$, using the same eligible pairs and references as the main progressive analysis.

Table 15: **Bounded-frame progressive score details.** Each cell is later-endpoint MRA / SCA (0–100) on the same fixed reference-changing pairs as Table 5. Qwen uses $K = 32, 64$; Cambrian-S/P also include $K = 16, 128$. Inputs are causally selected RGB frames, and QA calls are independent.

Backbone / input	Distinct Rooms	Object-containing Rooms	Target Visits	Path Length	Visited Floor Area
Cambrian-S / $K = 16$	21.0/17.1	44.6/25.0	62.8/6.7	0.1/0.6	24.8/15.0
/ $K = 32$	10.4/15.7	51.8/26.1	61.9/10.1	0.2/2.4	46.4/14.5
/ $K = 64$	7.3/20.9	49.4/32.0	65.5/14.0	0.2/5.1	49.5/11.8
/ $K = 128$	8.1/20.3	48.7/32.1	64.4/13.8	0.2/4.8	52.0/17.2
Cambrian-P / $K = 16$	10.3/9.5	47.1/12.2	61.7/3.1	0.0/1.5	48.3/15.8
/ $K = 32$	5.5/10.1	43.8/14.1	62.9/3.3	0.0/1.8	65.1/14.2
/ $K = 64$	5.5/10.1	43.0/18.1	63.5/4.3	0.0/2.3	72.4/15.5
/ $K = 128$	5.2/9.9	41.3/15.0	61.8/5.9	0.0/2.0	72.7/12.2
Qwen3-VL-8B / $K = 32$	55.9/43.1	51.0/37.4	38.9/26.8	0.0/5.2	33.6/16.6
/ $K = 64$	52.2/52.2	57.0/40.7	35.9/24.7	0.0/6.2	48.2/13.7

Table 14: **Fixed-denominator exact-count outcomes on reference-changing pairs.** Each task reports early exact accuracy, late exact accuracy, and joint count correctness $100p_{11}$ (%), macro-averaged by source question. Invalid answers count as failures. All rows use the primary 1-FPS input and the same eligible pairs within each task.

	Distinct Rooms			Object-containing Rooms			Target Visits		
System	Early	Late	Both	Early	Late	Both	Early	Late	Both
General-purpose video MLLMs									
Qwen2.5-VL-3B	5.9	9.1	3.0	46.0	36.5	21.3	43.9	23.3	12.6
Qwen2.5-VL-7B	7.0	9.7	3.3	47.4	33.8	21.0	43.0	13.8	6.8
Qwen2.5-VL-72B	30.1	29.4	15.2	50.3	38.7	27.1	51.9	11.7	8.9
Qwen3-VL-2B	16.3	16.3	8.2	45.3	43.7	25.6	50.9	24.0	16.3
Qwen3-VL-4B	32.3	34.1	17.8	43.2	39.4	23.1	52.9	22.3	16.9
Qwen3-VL-8B	23.5	26.1	13.1	38.7	39.7	20.6	58.7	24.5	14.9
Qwen3-VL-32B	14.1	13.4	4.9	34.5	30.2	12.4	51.7	18.0	13.8
InternVL3-1B	10.2	18.0	4.2	28.9	19.1	3.7	38.6	31.5	5.6
InternVL3-2B	11.4	18.0	5.1	43.2	32.1	20.6	48.9	49.7	16.8
InternVL3-8B	11.5	10.8	4.9	42.9	36.7	19.8	55.9	22.8	14.8
InternVL3-14B	15.7	19.5	8.8	38.3	43.6	18.3	64.2	38.6	24.9
InternVL3-38B	26.2	30.6	14.6	51.7	39.3	28.2	54.3	20.2	15.3
LLaVA-OneVision-7B	1.9	1.1	0.3	18.3	16.9	6.2	38.3	38.6	7.3
LLaVA-NeXT-Video-7B-Qwen2	4.6	3.6	0.7	14.4	25.9	6.9	46.8	43.5	13.2
Spatially specialized video MLLMs									
VG-LLM-4B	11.7	10.9	0.9	22.7	38.6	1.3	32.0	49.7	0.0
VG-LLM-8B	5.2	5.6	1.3	23.7	39.7	5.7	32.0	49.7	0.0
VLM-3R	4.3	4.5	0.5	8.8	23.3	2.3	31.0	48.7	0.0
Spatial-MLLM-v1.1	16.0	20.0	1.2	13.4	17.3	1.3	0.0	0.0	0.0
Cambrian-S-0.5B	9.9	8.8	0.8	0.0	0.0	0.0	32.5	19.6	1.6
Cambrian-S-1.5B	1.4	1.0	0.0	12.9	18.8	3.2	29.0	44.6	8.4
Cambrian-S-3B	2.9	2.6	0.2	7.1	19.0	3.3	23.4	41.1	0.0
Cambrian-S-7B	1.9	1.6	0.3	32.5	36.5	15.4	33.6	50.1	8.0
Cambrian-P-7B	0.5	0.7	0.0	9.6	28.2	1.9	28.5	48.0	2.1
SenseNova-SI-1.1 (InternVL3-2B)	12.7	17.2	4.2	26.4	38.5	8.2	43.5	45.6	10.4
SenseNova-SI-1.5 (InternVL3-8B)	10.3	11.7	2.3	34.6	35.6	11.9	33.7	46.6	2.8
SenseNova-SI-1.1 (Qwen2.5-VL-3B)	5.7	7.8	0.4	18.5	37.0	3.6	31.5	48.7	0.5
SenseNova-SI-1.1 (Qwen3-VL-8B)	20.5	25.0	11.6	37.4	36.4	19.4	41.2	45.0	10.9
SenseNova-SI-1.3 (Qwen3-VL-8B)	13.4	14.2	6.2	26.1	31.1	11.6	42.0	33.5	10.4
Streaming video MLLMs									
OASIS	27.7	29.5	14.5	48.4	39.0	23.9	43.6	10.9	7.8
Dispider	0.6	0.0	0.0	0.9	0.9	0.0	7.3	3.1	2.1

E SENSITIVITY OF THE REFERENCE-CHANGING ANALYSIS

The auxiliary response-repetition diagnostic (Appendix F.3) re-pairs later outputs only when the ordered reference values and question text are the same. The following comparisons test how its results depend on this grouping. Table 16 compares grouping by the reference values alone, by the reference values and question, and by those variables together with the home. Distinct Rooms has one question wording, so the first two groupings are identical. For Object-Containing Rooms and Target Visits, grouping only by the reference values permits re-pairing across different question wordings and yields larger observed-minus-expected agreement. Adding the home changes both the analytic expectation and the fraction of pairs in groups where re-pairing can change equality.

Adding home conditioning lowers the median excess to +3.6 percentage points for Distinct Rooms and +2.9 points for Object-Containing Rooms. The null is less informative for Target Visits: only a median 13% of its pairs can change equality under re-pairing. The small excess for this type therefore has little separation from pairings with the same conditioned output frequencies.

Table 16: **Conditioning changes the expected response-repetition rate on reference-changing pairs.** R is the observed response-repetition rate and E_{marg} subtracts the independent-marginal collision baseline. The conditioned columns instead subtract each null's analytic expectation. All percentages and percentage-point (pp) effects are medians across 30 systems. Conditioned excess is computed for each system before taking the across-system median. Positive counts and BH counts are out of 30. BH correction is performed separately for each null over the same defined 90 system–question-type comparisons. The final column is the median fraction of pairs in strata where re-pairing can change the equality statistic.

Type	Re-pairing strata	R (%)	Marginal E (pp)	Null expectation (%)	Median excess over null (pp)	Positive	BH $q < .05$	Pairs in non-fixed strata (%)
Distinct Rooms	Reference values	40.1	+22.8	28.7	+10.0	30/30	29/30	95
Distinct Rooms	+ exact question	40.1	+22.8	28.7	+10.0	30/30	29/30	95
Distinct Rooms	+ exact question + home	40.1	+22.8	36.5	+3.6	29/30	15/30	61
Object-Containing Rooms	Reference values	45.1	+18.1	28.6	+15.3	29/30	29/30	100
Object-Containing Rooms	+ exact question	45.1	+18.1	39.7	+5.1	29/30	18/30	67
Object-Containing Rooms	+ exact question + home	45.1	+18.1	41.8	+2.9	28/30	7/30	40
Target Visits	Reference values	77.1	+21.0	56.1	+18.6	26/30	26/30	99
Target Visits	+ exact question	77.1	+21.0	74.4	+1.5	19/30	5/30	13
Target Visits	+ exact question + home	77.1	+21.0	74.4	+1.3	19/30	2/30	13

Dependence among pairs from one source question. The eligible-pair set can contain multiple pairs derived from the same source question. Table 17 repeats the response-repetition analysis after selecting one pair per source question without consulting model outputs. We do not condition further on trajectory or source-question identity because doing so makes every re-pairing group a singleton. The one-pair analysis assesses sensitivity to repeated pairs from the same source question without making re-pairing impossible.

Table 17: **Sensitivity to selecting one pair per source question.** Across every draw, the across-system median excess remains positive for Distinct Rooms and Object-Containing Rooms. Within each of 200 seeded draws, one eligible pair is selected per source question before inspecting model outputs, and the same selection is used for every system. The question-conditioned re-pairing test and its defined 90-comparison Benjamini–Hochberg correction are then repeated. Brackets give empirical 2.5th–97.5th percentiles across draws.

Type	Full pairs → selected per draw	Median question-conditioned excess (pp) full → draw median [2.5th, 97.5th]	Draws with positive across-system median	BH-adjusted $q < .05$ counts (of 30) full → draw median [2.5th, 97.5th]
Distinct Rooms	384 → 84	+10.0 → +8.9 [+5.8–+12.2]	200/200	29 → 15 [5–23]
Object-Containing Rooms	185 → 81	+5.1 → +5.2 [+3.2–+7.0]	200/200	18 → 6 [1–13]
Target Visits	118 → 48	+1.5 → 0.0 [0.0–+1.0]	32/200	5 → 0 [0–0]

F SCORING AND STATISTICAL DETAILS

F.1 ENDPOINT AND PAIRWISE SCORING

Categorical endpoint types use exact accuracy. Numerical endpoint types use the MRA definition in Equation 1, with exact equality when $y = 0$. Every eligible endpoint is scored; an answer that does not satisfy the required format receives zero. The all-type average is the equal-weight mean of the 18 question-type scores, which mix MRA and exact accuracy.

For progressive endpoint scores, earlier and later responses are each scored against the reference at their own prefix. Early and late MRA use the same eligible pairs and source-question weights as SCA (Section 3.4). SCA and fixed-denominator joint count correctness retain invalid responses as failures. Response-repetition and native-unit change error require valid numerical answers at both endpoints because the compared values are otherwise undefined.

Primary count scores accept numerals; the auxiliary count diagnostics in Appendices F.3 and F.5 also accept English number words. The latter use valid-response pairs.

F.2 SAME-SEGMENT CORRECTNESS AND BALANCED-COHORT SENSITIVITY

Every system is scored on the same 24 mixed-role bundles. A question pair succeeds only when its answers match both endpoint references. Crossing the reference-stable and reference-changing pair outcomes gives four cases: both succeed, only the stable pair succeeds, only the changing pair succeeds, or neither succeeds. The control is separate from the main reference-changing SCA cohort.

The primary mixed-role comparison has a fixed denominator of 24 bundles per system. A question pair is correct at both endpoints only when both answers exactly match their references. Over all 38 intervals, the median equal macro of pair-level exact correctness across the two questions and two reference relations is 5.95%, and the largest observed value is 22.47%.

In the balanced Qwen3-VL-8B extension, the model is correct at both endpoints on 21 of 96 question pairs (21.9%; 95% home-cluster bootstrap interval, 13.3–30.0%). Table 18 reports its relation-specific and mixed-role outcomes.

Table 18: **Balanced-cohort sensitivity for Qwen3-VL-8B.** This extension contains 12 bundles for each of the four stable/changed reference combinations. Panel (a) reports the four endpoint-correctness transitions; each row contains 24 question pairs. Panel (b) retains the 24 mixed-role bundles and jointly classifies whether each question pair is correct at both endpoints. C and W denote exact correctness.

(a) Outcomes for each question and reference relation.

Question	Reference	C→C	C→W	W→C	W→W
Distinct Rooms	stable	25.0	16.7	8.3	50.0
	changed	20.8	16.7	12.5	50.0
Object-Containing Rooms	stable	29.2	12.5	0.0	58.3
	changed	12.5	20.8	16.7	50.0

(b) Joint outcomes in mixed-role bundles.

	Both C→C	Stable only	Changed only	Neither C→C
Mixed-role bundles	1 (4.2%)	4 (16.7%)	3 (12.5%)	16 (66.7%)

The interval resamples the 18 home identifiers with replacement, retains all bundles from each sampled home, and recomputes the equal macro over question type and reference relation. We use 2,000 bootstrap replicates; every replicate contains all four relations.

F.3 Response equality and conditioned re-pairing

Let $\mathcal{P}$ be a system's reference-changing pairs with valid numerical answers at both endpoints. The observed response-repetition rate is

$$R = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbf{1}[\hat{y}_i^{\text{early}} = \hat{y}_i^{\text{late}}]. \tag{5}$$

Let $p(v)$ be the frequency of value v across the early and late answers in $\mathcal{P}$. The marginal baseline is $C = \sum_v p(v)^2$, and the marginal excess is $E_{\text{marg}}^{\text{chg}} = 100(R - C)$ percentage points. This baseline preserves only the combined output frequencies.

The primary permutation null keeps the observed early outputs fixed and re-pairs late outputs only within groups sharing exact $(y^{\text{early}}, y^{\text{late}}, x)$, preserving both the reference values and question text. Singleton groups and groups whose later outputs are all identical therefore remain fixed. With $B = 2,000$ permutations, the one-sided Monte Carlo p-value is

$$p = \frac{1 + \sum_{b=1}^{B} \mathbf{1}[R_b \geq R]}{B + 1}. \tag{6}$$

Exchangeability here means that, after matching the question and ordered reference values, permuting the later outputs is assumed not to change their distribution. Re-pairing also breaks source-question identity, shared visual history, trajectory correspondence, and prefix-length matching; the null does not separately identify any of these factors. Let N_q be the expected agreement under this re-pairing distribution. The excess is $E_q^{\text{chg}} = 100(R - N_q)$ percentage points. Unlike C, N_q preserves the ordered reference values and question. The conditioned re-pairing distribution is the inferential null for the permutation p-value. It compares response association with conditional output frequencies but does not identify a causal process. Each system–question-type comparison requires at least 20 pairs with valid answers at both endpoints.

The sensitivity null further includes the home in each group and yields $N_{q,h}$ and $E_{q,h}^{\text{chg}} = 100(R - N_{q,h})$. We report the fraction of pairs in groups where re-pairing can change equality because finer conditioning can leave assignments fixed. Let f be this fraction. Because fixed groups contribute zero, the all-pair excess equals f times the excess among groups that can be re-paired. Since f depends on each system's output distribution, we report it for interpretation but retain the all-pair excess as the primary statistic. If $f = 0$, the null yields $E = 0$ and $p = 1$, and the test remains among the defined comparisons for multiple-testing correction. Home conditioning and one-pair-per-question sampling probe home composition and repeated-pair dependence, respectively. Neither conditions on shared trajectory content or prefix length, which would leave many groups too small or fixed for the present re-pairing test. Benjamini–Hochberg correction is applied separately within each null specification to the same defined 90 system–question-type tests (30 systems by three discrete or count-like types). We report one-sided p-values and BH q-values for these comparisons, interpreting $q < 0.05$ only within this defined family and under the exchangeability assumption. Because systems are evaluated on shared benchmark instances, counts of systems below this threshold are not a sample from a population of independent systems.

F.4 Path Length predicted-increase rate and change magnitude

The response and reference changes and absolute change error are defined in Section 3.4. All evaluated Path Length and Visited Floor Area reference changes are positive. We additionally report the predicted-increase rate, the fraction with $\Delta\hat{y}_i > 0$, and the unchanged-prediction rate, the fraction with $\Delta\hat{y}_i = 0$, over valid-response pairs. On pairs with valid numerical answers at both endpoints, Path Length estimates increase on 3.05–86.11% of evaluated pairs and remain identical on 2.1–92.1%. For Visited Floor Area, the increase rate spans 3.49–84.88%, and the post-hoc constant matches or exceeds 28 of the 30 endpoint MRA scores. These response patterns accompany the endpoint results in Table 3; an increase alone does not determine the error in meters or square meters. For example, Qwen2.5-VL-72B has the highest Path Length predicted-increase rate, 86.11%, with a mean change error of 23.64 m over all 180 eligible pairs. For Visited Floor Area, SenseNova-SI-1.1 (Qwen3-VL-8B) has an increase rate of 84.88% and a mean change error of 5.42 m^2 over all 86 eligible pairs. These examples report direction and magnitude on identical evaluated pairs within each system.

F.5 COUNT CORRECTNESS BY SYSTEM

Table 19 summarizes correctness and repeated answers on valid-response pairs under the answer normalization defined in Appendix F. Table 20 reports the four exact-correctness outcomes underlying its panel (a) for each system. Each source question contributes the average over its valid-response pairs. Denote these valid-response proportions by p_{ab}^{valid}. Within a system, early exact accuracy on that conditional set is $100(p_{11}^{\text{valid}} + p_{10}^{\text{valid}})$ and late exact accuracy is $100(p_{11}^{\text{valid}} + p_{01}^{\text{valid}})$. These identities hold within a system, not for separately computed across-system medians. The conditional p_{11}^{valid} is distinct from fixed-denominator p_{11} in Equation 4.

Table 19: **Count correctness across growing prefixes.** All values are percentages, summarized as separate medians over 30 systems. Panel (a) uses exact count correctness on reference-changing pairs with valid numerical answers at both endpoints after number-word normalization (Appendix F), averaged first within source questions. Panel (b) gives the pair-level repetition rate and the composition of repeated pairs. Early-only and late-only denote loss and recovery of exact correctness, respectively. Neither denotes error at both endpoints, irrespective of the predicted increment. Each system's outcome proportions sum to 100%; their separate medians need not.

(a) Exact correctness across the two endpoints

Question type	Both	Early only	Late only	Neither
Distinct Rooms	3.2	6.7	8.3	79.5
Object-Containing Rooms	11.8	18.9	18.5	44.3
Target Visits	8.9	32.1	29.4	33.2

(b) Correctness when the answer repeats

	Repeat rate	Among repeated pairs		
Question type		Early only	Late only	Neither
Distinct Rooms	40.1	7.5	15.9	76.7
Object-Containing Rooms	45.1	28.9	37.0	34.1
Target Visits	77.1	36.4	31.5	34.6

G ACCESS, LICENSING, AND RESPONSIBLE USE

Scan-grounded assets follow source licenses, including HM3D-Semantics (Ramakrishnan et al., 2021) and HOV-SG (Werby et al., 2024). We distribute metadata for rendered home-traversal videos by default and redistribute the videos only when source licenses permit. Released artifacts retain provenance and transformation metadata.

TRACE-VSI supports research on video-capable spatial reasoning and is not a certification for safety-critical navigation, surveillance, tracking, or human-equivalent spatial cognition. Users must follow source licenses, limit redistribution, and redact identifying content where appropriate.

Table 20: **Count correctness with normalized numerical answers.** C and W indicate correct and wrong answers, respectively; arrows run from the earlier to the later endpoint. Values are source-question-averaged percentages on pairs with valid numerical answers at both endpoints, as defined in Appendix F.5. Answers accept numerals and English number words, with scoring as defined in Appendix F. Each type's four outcomes sum to 100% within rounding. Systems are listed alphabetically, not ranked.

	Distinct Rooms				Object-Containing Rooms				Target Visits			
System	C→C	C→W	W→C	W→W	C→C	C→W	W→C	W→W	C→C	C→W	W→C	W→W
Cambrian-P-7B	0.0	0.5	0.8	98.7	1.9	7.7	26.3	64.1	2.1	26.5	45.9	25.6
Cambrian-S-0.5B	0.8	10.0	8.9	80.2	0.0	10.2	21.3	68.5	1.6	31.0	18.1	49.4
Cambrian-S-1.5B	0.0	1.4	1.0	97.5	3.2	9.7	15.6	71.5	8.4	20.6	36.2	34.9
Cambrian-S-3B	0.2	2.7	2.3	94.7	3.3	3.8	15.7	77.2	0.0	23.4	41.1	35.4
Cambrian-S-7B	0.3	1.6	1.3	96.8	15.4	17.1	21.1	46.4	8.0	25.6	42.1	24.3
Dispider	0.0	0.6	0.0	99.4	1.5	11.9	18.9	67.7	10.0	11.5	28.1	50.4
InternVL3-14B	8.8	6.9	10.7	73.7	18.3	20.0	25.3	36.4	24.9	39.3	13.7	22.1
InternVL3-1B	4.2	6.2	14.0	75.7	4.4	25.5	15.7	54.3	5.6	33.1	25.9	35.5
InternVL3-2B	5.1	6.3	12.9	75.7	20.6	22.6	11.5	45.3	16.8	32.0	32.9	18.3
InternVL3-38B	14.6	11.6	16.0	57.8	28.2	23.6	11.1	37.1	15.3	39.0	4.9	40.8
InternVL3-8B	4.9	6.6	6.0	82.5	19.8	23.1	17.0	40.1	14.8	41.0	8.0	36.1
LLaVA-NeXT-Video-7B-Qwen2	0.8	4.3	3.4	91.6	7.2	8.1	20.1	64.6	13.2	33.9	30.6	22.4
LLaVA-OneVision-7B	0.3	1.6	0.8	97.3	6.4	11.0	11.4	71.2	7.3	32.1	32.5	28.1
OASIS	14.5	13.2	15.0	57.3	23.9	24.5	15.1	36.5	8.9	35.4	3.3	52.4
Qwen2.5-VL-3B	3.0	2.8	6.1	88.0	21.3	24.7	15.2	38.8	12.6	31.3	10.8	45.3
Qwen2.5-VL-72B	15.2	14.9	14.2	55.7	27.1	23.3	11.6	38.1	8.9	43.0	2.8	45.3
Qwen2.5-VL-7B	3.3	3.7	6.4	86.7	21.0	26.4	12.9	39.7	6.8	36.2	6.9	50.0
Qwen3-VL-2B	8.2	8.0	8.1	75.7	25.6	19.7	18.1	36.6	16.3	34.6	7.7	41.4
Qwen3-VL-32B	4.9	9.3	8.6	77.3	12.4	22.0	17.8	47.7	13.8	37.9	4.2	44.1
Qwen3-VL-4B	17.8	14.5	16.3	51.3	23.1	20.1	16.3	40.5	16.9	36.0	5.4	41.7
Qwen3-VL-8B	13.1	10.4	13.0	63.5	20.6	18.1	19.1	42.2	14.9	43.8	9.6	31.6
SenseNova-SI-1.1 (InternVL3-2B)	4.2	8.5	13.0	74.2	8.2	18.2	30.2	43.3	10.4	33.1	35.1	21.4
SenseNova-SI-1.1 (Qwen2.5-VL-3B)	0.4	5.3	7.4	86.9	3.6	14.9	33.4	48.0	0.5	31.0	48.2	20.3
SenseNova-SI-1.1 (Qwen3-VL-8B)	11.6	8.9	13.4	66.1	19.4	18.0	17.0	45.6	10.9	30.3	34.1	24.7
SenseNova-SI-1.3 (Qwen3-VL-8B)	6.2	7.1	7.9	78.7	11.6	14.5	19.4	54.4	10.4	31.6	23.1	34.9
SenseNova-SI-1.5 (InternVL3-8B)	2.3	8.0	9.3	80.3	11.9	22.6	23.7	41.8	2.8	31.0	43.8	22.4
Spatial-MLLM-v1.1	1.7	16.7	20.1	61.5	4.4	21.3	35.7	38.6	1.0	31.0	49.7	18.3
VG-LLM-4B	0.9	10.9	10.1	78.2	1.3	21.4	37.2	40.0	0.0	32.0	49.7	18.3
VG-LLM-8B	1.3	3.9	4.3	90.5	5.7	18.0	34.1	42.3	0.0	32.0	49.7	18.3
VLM-3R	0.7	3.8	4.5	91.0	2.4	6.2	22.4	69.0	0.0	32.5	50.2	17.4