MEMACT: STREAMING SPATIAL UNDERSTANDING WITH MEMORY AND ACTIVE EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial understanding is essential for embodied agents navigating and interacting with the physical world. As observations arrive incrementally, agents must reason in a streaming manner. Yet streaming understanding must also be closed-loop: task-relevant evidence may require both recalling past observations and actively acquiring new views. Existing streaming approaches largely emphasize processing incoming observations, leaving task-driven active exploration underexplored. We introduce MemAct, a framework for streaming spatial understanding with memory and action. A shared vision–language model incrementally writes structured historical memory, retrieves records and past frames on demand, and acquires additional evidence through movement and turning. Retrieved evidence and action observations inform subsequent decisions, enabling the agent to continue exploring or answer within a unified reasoning loop. We develop the MemAct suite, a data generation pipeline and collection across real and simulated environments, with approximately 1M training examples for spatial reasoning, memory writing, retrieval, and exploration. Its MemAct-bench comprises 20,000 main QA pairs evaluating current-observation understanding, historical evidence use, and active exploration, plus 2,000 separate Timing questions. MemAct combines supervised fine-tuning for memory and interaction with counterfactual action optimization using cumulative evidence gains, action costs, and terminal answer quality to improve exploration and stopping decisions. Experiments show that MemAct outperforms existing spatial understanding and streaming baselines and approaches frontier multimodal models on MemAct-bench while retaining streaming inference efficiency.

1 INTRODUCTION

Spatial understanding enables embodied agents to reason about objects, their locations, geometry, and relationships while navigating and interacting with an environment (Yang et al., 2024a;b; Huang et al., 2022; Jatavallabhula et al., 2023; Gu et al., 2023). In a live setting, observations arrive incrementally (Das et al., 2017; Gordon et al., 2017; Yokoyama et al., 2023): future views are unavailable, previously seen objects may leave the visual context, and some task-relevant details may never appear unless the agent changes viewpoint. A spatial question may therefore be answerable from the current view, require recalling an earlier observation, or require acquiring a new one. Streaming spatial understanding is consequently a closed-loop process in which the agent uses available evidence to decide whether to answer, retrieve history, or explore.

Recent streaming work advances spatial reasoning over incrementally arriving observations (Xu et al., 2025; Yu et al., 2026; Wang et al., 2025b). Yet these methods generally follow a predetermined camera stream: earlier evidence may leave the retained context, while needed views may never be acquired (Yang et al., 2024b; Hong et al., 2026). Active embodied QA systems choose new viewpoints, often using geometric maps or structured spatial memories (Jiang et al., 2025; Yang et al., 2024b; Ginting et al., 2025). Some also retain experience from before a question arrives. We study a complementary challenge: training a single streaming VLM to build retrievable memory as RGB observations arrive and, once queried, decide whether to retrieve earlier evidence, acquire a new view, or answer. Each result informs the next decision, without supplying depth maps or a reconstructed 3D map to the model at inference time. Figure 1 contrasts this memory-and-action loop with offline and passive streaming QA.

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

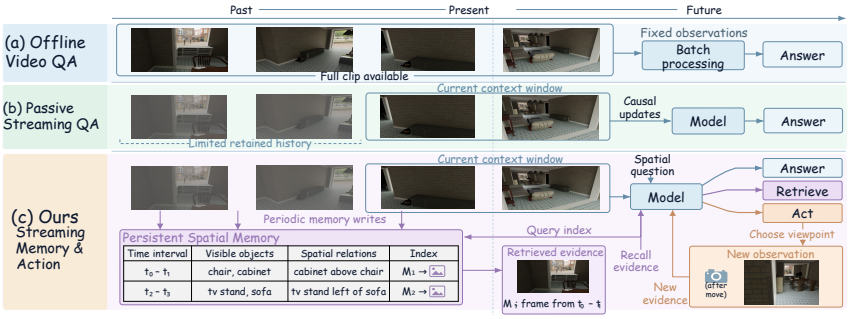


Figure 1: Offline, passive streaming, and closed-loop streaming spatial understanding. MemAct combines reasoning over incremental observations, historical memory retrieval, and active acquisition of missing evidence in a shared exploration-and-answering loop.

This setting poses three distinct challenges. First, a streaming VLM must preserve potentially useful evidence before knowing the question, even as earlier views leave its limited context. Second, once queried, it must determine whether the needed evidence is visible now, recoverable from history, or obtainable only through a new camera view. Learning this distinction requires causal supervision over what the agent has seen and what its actions reveal. Third, an informative view may require preparatory movements with little immediate gain, so training must credit their delayed value while teaching the model when to stop exploring and answer.

We introduce **MemAct**, a simple design that makes memory and exploration native decisions of a streaming VLM. The model processes incoming RGB observations incrementally and periodically generates compact, timestamped records linked to stored frames. When queried, it uses the same generation interface to answer, request past evidence, or select a discrete camera action. A a simple rule-based interface returns the requested records or frames, or the resulting new view, to the model for its next decision. This lets the VLM learn what to record, retrieve, and observe without a separate spatial planner, depth maps, or a reconstructed 3D scene map as model inputs. (Section 4)

For data construction, we introduce MemAct-data, a data suite that combines streaming spatial QA with supervision for memory writing, retrieval, and multi-step exploration. An automated pipeline constructs approximately one million training examples from real and simulated visual streams; constructed action supervision from changes in target visibility, enabling the model to recognize insufficient evidence and learn how to explore for more informative views. We also introduce MemAct-bench, with 20,000 questions testing reasoning from current, historical, and newly acquired observations, plus 2,000 separate questions testing when to answer. (Section 3)

For training, supervised fine-tuning establishes the shared memory, retrieval, exploration, and answering protocol. To credit preparatory actions with delayed benefits, bandit-based counterfactual optimization compares camera actions and immediate answering from the same causal state through multi-step continuations. Cumulative evidence gains, action costs, and terminal answer quality encourage useful exploration beyond immediate gains while learning when to stop; supervised QA and memory replay preserve existing capabilities (Section 4.4).

On MemAct-bench, MemAct-9B achieves 67.27% overall multiple-choice accuracy and 60.23 numerical mean relative accuracy, compared with 61.39% and 48.65 for StreamingVLM. Its aggregate scores are close to those of the evaluated GPT-6 Astra model, while it leads the separate response-timing evaluation. Ablations show that explicit memory substantially improves questions requiring historical evidence, and that active exploration and counterfactual optimization improve questions requiring new views. MemAct also remains competitive on established spatial reasoning benchmarks, including VSI-Bench, ReVSI, MMSI-Bench, and ViewSpatial. Together, these results support spatial understanding that can use the present, recall the past, and actively acquire missing evidence within a causal stream.

Our contributions are fourfold: 1)**Task formulation.** We extend streaming spatial understanding with task-driven active exploration and introduce MemAct, unifying historical evidence retrieval, new-view acquisition, and answering. 2) **Data construction.** We develop an efficient, automated pipeline for generating memory, retrieval, and exploration supervision, producing the MemAct suite with approximately 1M training examples from real and simulated environments. 3) **Training strategy.** We combine supervised interaction learning and cumulative-return counterfactual optimization

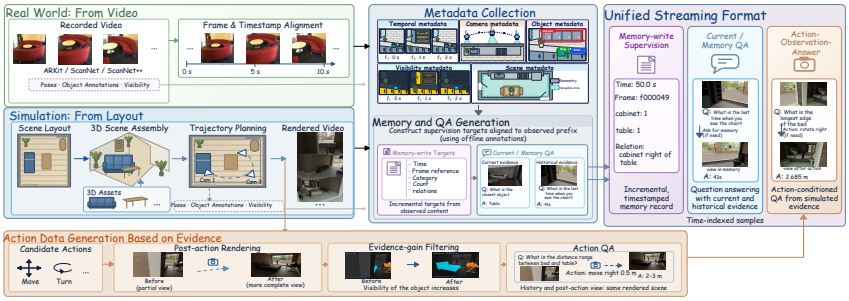


Figure 2: MemAct suite. Aligned temporal, camera, object, visibility, and scene metadata supports memory and spatial QA supervision from real and simulated observations. Dedicated rendering and geometric/evidence checks yield discrete multi-step camera-action data; all supervision forms causal streams.

in multi-stage training to improve exploration and stopping decisions. 4) **Evaluation.** We build MemAct-bench: 20,000 main QA pairs and 2,000 separate Timing questions evaluating memory, exploration, task performance, and streaming inference efficiency.

2 Related Work

Spatial Understanding. Recent work advances spatial understanding through dedicated benchmarks (Azuma et al., 2021; Ma et al., 2022; Fu et al., 2024) and spatially grounded training data (Chen et al., 2024; Cheng et al., 2024; Song et al., 2024). VSI-Bench (Yang et al., 2024a) evaluates spatial relations, measurement estimation, and spatiotemporal reasoning from indoor videos. SAT (Ray et al., 2024) synthesizes static and dynamic spatial QA in simulation, while Cambrian-S (Yang et al., 2025a) develops VSI-590K and spatial training, extending to continual sensing through latent frame prediction and surprise-driven memory management. These approaches strengthen spatial reasoning over supplied visual evidence. MemAct contributes an automated pipeline that constructs supervision for historical memory, retrieval, and multi-step exploration in streaming spatial QA.

Streaming Spatial Understanding. Streaming methods address causal processing and efficient retention of continuously arriving observations (Zhang et al., 2024; Yao et al., 2025; He et al., 2024). StreamingVLM (Xu et al., 2025) combines a compact KV cache with overlapping-chunk SFT to align training and streaming inference. Stream3D-VLM (Yu et al., 2026) integrates incremental geometry priors and geometry-adaptive voxel compression for online spatial understanding. StreamEQA (Wang et al., 2025b) evaluates perception, interaction, and planning across backward, real-time, and forward reasoning on recorded videos. StreamingBench (Lin et al., 2024) evaluates online understanding; VideoChat-Online (Huang et al., 2024) and LiveCC (Chen et al., 2025) develop memory-based processing and real-time commentary, respectively. MemAct additionally learns observation acquisition: the model retrieves historical evidence or executes camera actions, incorporates their outcomes, and decides whether to continue exploring or answer.

Active Perception. Active perception acquires task-relevant evidence through viewpoint selection and environmental interaction (Majumdar et al., 2024). EXPRESS-Bench (Jiang et al., 2025) evaluates the exploration and reasoning capability of VLMs. ESI-Bench (Hong et al., 2026) evaluates the perception–action loop through perception, locomotion, and manipulation, exposing failures in action selection and premature commitment despite insufficient evidence. MemoryEQA (Zhai et al., 2025) uses structured memory throughout exploration, while Mind Palace (Ginting et al., 2025) combines long-term recall with exploration and stopping. MemAct contributes a shared streaming VLM trained for memory writing, retrieval, and camera-action selection, with cumulative-return optimization of exploration and stopping using downstream evidence and answer quality.

3 Constructing the MemAct Suite

Our pipeline builds supervision for recording evidence, retrieving history, and acquiring new views through five stages: observation acquisition, metadata alignment, memory and QA generation, action rendering, and stream assembly (Figure 2). Each question is linked to its arrival time, supporting

observations, and memory or action supervision. Aligned metadata supports answer computation and camera-transition selection but remains separate from model inputs.

3.1 Observation Construction and Evidence Alignment

Simulated observations use 3D-FRONT (Fu et al., 2021) layouts and assets after removing conspicuously interpenetrating or implausibly scaled objects. Camera trajectories scan rooms in randomized visitation orders using mesh-based A* planning (Hart et al., 1968) and Bézier smoothing. Multiple trajectories diversify observation histories, while Cycles (Blender Foundation, 2026) rendering varies floor/ceiling textures and HDRI lighting. Trajectory construction is detailed in Appendix A.

Real sequences use ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021) videos, camera information, and 3D annotations. Both sources share the memory and QA pipeline; simulation additionally renders action-conditioned views with scene geometry, appearance, and imaging settings fixed within each interaction.

Unified Metadata and Temporal Evidence Alignment. We standardize categories, units, and coordinates across five metadata groups: **temporal** frame/time mappings, **camera** intrinsics and poses, **object** identities and geometry, **visibility** annotations and image-space measurements, and **scene** geometry and occupancy. Timestamps align projected objects and source visibility annotations with observations; field definitions appear in Appendix A.2.

Visibility indicators, projected-box occupancy, framing completeness, and visible-pixel fractions remain distinct. Initial simulated streams use source visibility and clipped box projections; action views additionally provide object masks. Projected occupancy measures image extent, not unoccluded surface area (Appendix A.2). At question time t, inputs contain only the causal prefix $O_{\leq t} = \{o_i : t_i \leq t\}$; action observations enter after their commands.

3.2 Memory and Spatial QA Supervision

Incremental memory targets. At 1 FPS, observations form five-second write windows. Memory targets contain timestamp-linked frame references, per-frame category counts, and optional image-plane relations. Visibility selects observed instances; projected positions on a 3×3 grid support unambiguous directional relations. Records describe observed content, not a complete scene graph; empty frames and final partial windows are retained.

Per-frame counts enumerate visible instances of each category; historical distinct-instance answers deduplicate sightings across the causal prefix using offline instance identities, which are excluded from model-visible memory (Appendix A).

Questions and evidence requirements. We select a time and template, bind eligible referents, compute the answer, and attach evidence. Object/camera metadata supplies distances, dimensions, and relative directions; temporal/visibility metadata supplies appearances, order, counts, and changes. Room-area questions require floor polygons and coverage checks. Candidates must satisfy visibility, referent-uniqueness, and answer-separation criteria.

Current QA uses the current observation or a specified recent window. **Memory QA** requires earlier appearances, previously observed instances not all currently visible, or historical relations; direct answering is allowed when this evidence remains in context. LLM rewriting varies wording while preserving referents, temporal constraints, and annotation-derived answers. Template calculations and examples appear in Appendix A.

Retrieval supervision. Historical evidence determines query categories, optional time ranges, and image requests. The deterministic router applies inference-time retrieval rules to constructed memory, yielding records and optional historical frames. Requests, responses, and answers form supervised traces; private instance-level checks exclude traces without sufficient evidence within budget without altering router outputs (Appendix C).

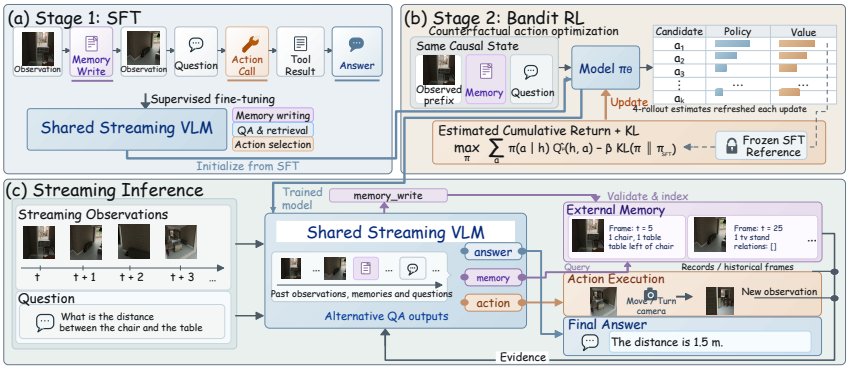


Figure 3: MemAct training and inference. A shared VLM learns incremental memory writing, answering, historical retrieval, and discrete-action view acquisition through SFT. Counterfactual optimization improves exploration and stopping using cumulative evidence gains, action costs, and terminal answer quality.

3.3 Multi-round Action Evidence Construction

Dedicated simulation rendering supplies action supervision. We bind identifiable targets and compute fixed answers: size is the longest physical bounding-box side, and distance is the Euclidean distance between 3D centers. Scene, targets, question, and answer remain unchanged, so exploration changes evidence rather than the task.

Discrete atomic actions. Five parameter-free commands translate forward, left, or right by 0.5 m in the current camera's horizontal frame, or rotate left/right by $45°$. Translation preserves orientation; rotation preserves position. The executor fixes magnitudes; supervision uses the model's commands. Collision and clearance checks precede rendering each admissible resulting state.

Evidence-based sequence selection. Joint target support is the minimum image-occupancy score across required objects, ensuring evidence for both targets in pairwise questions. Endpoints must satisfy occupancy and framing thresholds and exceed the best pre-question joint support by both relative and absolute margins. We distinguish acquiring historically missing adequate views from improving existing ones; measurement definitions and selection thresholds are in Appendix A.

We execute and render up to three admissible actions sequentially, retaining the earliest endpoint meeting these criteria. Intermediate actions may reposition the camera without immediate gain. Each demonstration stores the causal prefix, fixed question, command–observation pairs, and answer. Counterfactual optimization additionally uses rendered alternative outcomes; demonstration thresholds and RL rewards are separately defined in Appendices A and D.2.

Streaming Assembly and Quality Control. Across all task types, observations, memory writes, questions, commands, and tool results are assembled chronologically. Validation checks media, temporal/object references, answers, duplicates, command–observation alignment, budgets, and predecessor-state continuity. Missing evidence, failed checks, or excessive token length exclude a sample; construction metadata and private labels remain separate from public events (Appendix A).

The MemAct-data contains approximately 1M examples, all for training, while MemAct-bench contains 20,000 main QA pairs covering current understanding, historical evidence use, and active exploration, plus 2,000 separate Timing questions. Training-data composition and action-length distributions are in Appendix B, with construction thresholds in Appendix A.

4 Streaming Spatial Reasoning with Memory and Action

4.1 Problem Formulation and Overview

MemAct uses a shared VLM for memory and evidence-guided interaction. Figure 3 links inference to supervised interaction learning and counterfactual action optimization. At time t, the agent maintains causal context C_t in an incremental KV cache and external memory M_t. Following compact-cache streaming (Xu et al., 2025), the oldest visual tokens are preferentially evicted at the token

limit. A deterministic harness retains structured memory and historical images independently, enabling retrieval beyond the visual context.

For question q_t, history h_k contains the question, available observations, prior commands/results, and remaining budgets at round k. The model answers, retrieves, or acts; periodic memory writing is separate during observation. The harness executes commands without a learned router. Decision inputs exclude future observations, ground-truth geometry, private instance identities, and rewards.

4.2 Incremental Memory Writing and Retrieval

To preserve history beyond visual-cache eviction, MemAct writes increments ΔM_t every five seconds: per-frame categories, counts, frame references, and optional image-plane relations. The update $M_t = \mathrm{Update}(M_{t-5}, \Delta M_t)$ validates structure and temporal references without semantic correction; storage and retry rules appear in Appendix C.

Structured records locate historical evidence; retrieved frames recover visual detail absent from those records. A `use_memory` request specifies categories, an optional time range, and image requirements, returning matching records, relations, and optional frames to shared context. Each call returns at most two images, with four distinct historical images per question; filtering and pagination follow Appendix C.

4.3 Evidence Selection and Multi-round Action

Memory recovers past evidence; camera actions acquire views missing from retained observations. MemAct answers directly when evidence suffices, including historical QA; otherwise, it retrieves or selects one of five camera commands with executor-fixed magnitudes (Section 3.3).

A retrieval or camera command u_k and result z_k yield $h_{k+1} = \mathrm{Append}(h_k, u_k, z_k)$, informing subsequent decisions. Successive camera actions build on preceding states to improve evidence, with at most three attempts and an independent four-call retrieval budget per question. Answering terminates interaction; action-budget exhaustion requests a final answer. Action views remain in question context without automatic memory writes; the original stream resumes after exploration. Figure 4 illustrates current-view QA, memory-based last-observation QA, and three-action view improvement; Appendix E adds QA and memory-write examples.

4.4 Training Strategy

Stage I: Supervised Fine-Tuning. SFT learns writing, retrieval, answering, and single-/multi-step exploration from causal MemAct-data traces, including preparatory actions with delayed benefits, together with VSI-590K (Yang et al., 2025a). We minimize next-token cross-entropy over assistant tokens in memory writes, retrieval requests, camera commands, and answers, averaged within each sequence. Observations, prompts, and tool results supply context but are loss-masked; every decision uses only preceding events. Training settings and loss normalization appear in Appendix D.

Stage II: Counterfactual Optimization with Cumulative Returns. After SFT, a contextual-bandit objective uses multi-step candidate values to improve exploration and stopping. At history h, we evaluate the five camera commands and `answer_now` from independent copies with identical causal state and budget. Each camera attempt costs one unit before continuation under frozen policy snapshot μ. Values capture later evidence, answer quality, and preparatory-action benefits.

Evidence and answer rewards. View quality $E(x, q) \in [0, 1]$ combines normalized visible-pixel fractions, framing completeness that discounts cropping, and joint support limited by the weaker target in pairwise questions. The best quality over pre-question observations and preceding action outcomes is $\Phi(h, q) = \max_{x \in \mathcal{O}(h)} E(x, q)$.

Within an action rollout, let h_j denote the history after j camera attempts. Attempt a_j receives

$$r_j = \alpha[\Phi(h_{j+1}, q) - \Phi(h_j, q)]_+ - \lambda_a c(a_j) - \lambda_{\text{inv}} \mathbf{1}[\text{invalid}(a_j)]. \tag{1}$$

Revisiting previously acquired evidence earns no repeated gain. Invalid actions preserve the observation, receive no evidence gain, and consume an attempt. The terminal answer receives $R_{\text{QA}}(\hat{y}, y^\star)$,

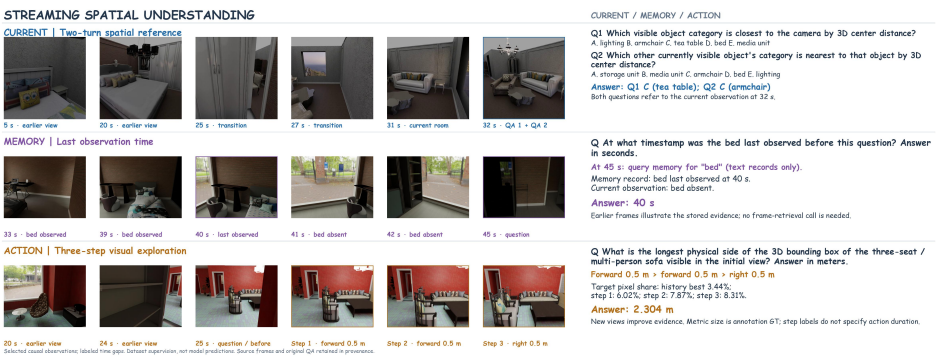


Figure 4: Current/Memory/Action QA: available observations, historical evidence, and newly acquired views.

using exact option correctness or numerical MRA. Evidence functions, coefficients, and reward ablations are detailed in Appendix D.2.

Cumulative values and policy updates. For a rollout of T camera attempts, decision j has return

$$G_j = \sum_{\ell=j}^{T-1} \gamma^{\ell-j} r_\ell + \gamma^{T-j} R_{\mathrm{QA}}(\hat{y}, y^\star). \tag{2}$$

Answer-now terminates without camera cost. We estimate candidate values from B continuations as $\widehat{Q}^\mu(h,a) = \frac{1}{B}\sum_{b=1}^{B} G^{(b)}(h,a)$ and train on initial and visited nonterminal histories. These rollouts use available causal context without initiating additional memory retrieval. Branch observations enter the context after their corresponding actions; branch validation is specified in Appendix A.

Summed conditional log probabilities of candidate decision-value tokens, without length normalization, define a categorical policy $\pi_\theta(a \mid h)$. We optimize

$$\mathcal{L}_{\mathrm{action}} = \mathbb{E}_h \left[- \sum_{a \in \mathcal{A}(h)} \pi_\theta(a \mid h) \widehat{Q}^\mu(h,a) + \beta D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h) \| \pi_{\mathrm{SFT}}(\cdot \mid h)) \right]. \tag{3}$$

KL regularization anchors the policy to the frozen SFT reference (Ouyang et al., 2022). We refresh μ before each update and detach estimated values within it. Supervised QA and memory replay interleave with action optimization to preserve answering, tool use, and memory writing. Appendix D.2 specifies sampling, loss normalization, and rollout settings.

5 Experiments

5.1 Experimental Setup

We compare closed-/open-source general-purpose VLMs, spatial and streaming models, and uniform-option, modal-answer-content, and modal-option-letter baselines. Models include GPT-6 Astra (OpenAI, 2026), Gemini-3.8-Flash (Google DeepMind, 2026), Qwen3.5 (Qwen Team, 2026a), Qwen3.8-27B (Qwen Team, 2026b), InternVL3.5 (Wang et al., 2025a), VLM-3R (Fan et al., 2025), Cambrian-S (Yang et al., 2025a), StreamingVLM (Xu et al., 2025), and Stream3D-VLM (Yu et al., 2026). We train MemAct on our dataset and VSI-590K (Yang et al., 2025a). At each decision, offline models receive 32 uniformly sampled frames from the available prefix; streaming models observe causally at 1 FPS. Multiple-choice and numerical questions use accuracy (ACC) and mean relative accuracy (MRA), respectively, following VSI-Bench (Yang et al., 2024a). Overall QA pools Current, Memory, and Action across domains by question count (13,134 multiple-choice; 6,866 numerical), scoring multi-turn questions per turn. Action is simulation-only, with all models using the same discrete interface and attempt budget and receiving action observations before answering.

Timing uses a separate 2,000-question set balanced across domains. Questions precede evidence; models observe at 1 FPS and wait or answer without memory/action tools. The first answer is final and on time within the inclusive five-second window from the first evidence-sufficient checkpoint. We report on-time rate (T), joint timing-and-answer accuracy (JA), and timing-gated MRA (JM), averaged over applicable questions (Appendix D.4).

Table 1: MemAct-bench results. ACC and MRA assess multiple-choice and numerical questions, respectively; on-time rate (T), joint timing-and-answer accuracy (JA), and timing-gated MRA (JM) assess response timing. Latency (seconds) spans image/context readiness to complete answer output (exclusions: Section 5.1).

Model / baseline	Current		Memory		Action		Overall QA		Timing	Latency
	ACC	MRA	ACC	MRA	ACC	MRA	ACC	MRA	T / JA / JM	s/QA
Closed-source models										
GPT-6 Astra	**68.16**	**53.04**	64.43	**68.68**	**59.57**	**49.82**	66.74	**61.15**	81.3 / 54.2 / 49.6	–
Gemini-3.8-Flash	59.26	46.29	56.41	59.55	52.20	43.42	58.17	53.14	67.1 / 38.9 / 36.0	–
Reference baselines										
Uniform options	23.22	–	22.71	–	25.00	–	23.11	–	–	–
Modal answer content	32.62	29.51	22.36	41.64	54.76	34.04	29.89	36.73	–	–
Modal option letter	26.10	–	25.68	–	27.68	–	26.00	–	–	–
Open-source models										
Qwen3.5-4B	36.31	22.96	46.63	43.98	31.50	21.44	39.50	34.23	30.9 / 12.2 / 10.5	1.558
Qwen3.5-9B	42.82	27.11	55.31	51.78	36.66	24.59	46.67	40.24	42.3 / 19.8 / 16.7	2.475
Qwen3.8-27B	44.47	28.51	57.24	53.72	39.65	27.23	48.44	42.10	46.4 / 22.4 / 19.5	5.423
InternVL3.5-8B	37.34	23.56	47.73	44.98	32.43	22.17	40.55	35.07	33.9 / 13.7 / 11.8	2.495
InternVL3.5-38B	39.74	25.03	51.66	48.78	34.88	24.08	43.45	37.86	38.8 / 16.8 / 14.6	5.446
Spatial models										
VLM-3R	43.52	27.71	56.63	52.69	36.78	24.96	47.55	40.98	44.0 / 21.0 / 17.5	2.622
Cambrian-S	42.54	26.56	54.20	51.38	35.71	24.30	46.11	39.80	43.4 / 20.1 / 16.8	2.446
Cambrian-S-FT	66.19	41.54	63.46	61.85	57.76	39.87	65.10	52.40	73.3 / 47.6 / 38.5	2.523
Streaming models										
StreamingVLM	63.24	38.06	58.85	58.54	46.03	33.66	61.39	48.65	63.6 / 39.1 / 31.3	**0.377**
Stream3D-VLM	60.42	36.78	55.88	56.46	43.76	32.32	58.54	46.92	61.0 / 35.7 / 29.2	0.410
MemAct-4B	62.01	37.50	64.13	63.80	51.54	37.68	62.43	51.87	84.9 / 53.1 / 44.0	0.383
MemAct-9B	67.22	52.28	**68.11**	67.72	58.28	48.74	**67.27**	60.23	**85.3 / 57.4 / 50.9**	0.420

Table 2: MemAct-9B Memory QA ablations: (1) content, (2) retention length, (3) write frequency.

A. Memory components

Setting	ACC	MRA
No explicit memory	54.52	52.24
Full memory	68.11	67.72
GT memory	70.77	71.06
w/o Frames	64.92	63.72
w/o Count	68.10	66.62
w/o Relations	67.91	67.72

B. Retention window

Window	ACC	MRA
10 s	58.52	57.45
20 s	60.91	60.12
30 s	63.20	62.48
40 s	65.04	64.47
50 s	66.50	65.95
Full history	68.11	67.72

C. Write interval

Interval	ACC	MRA
1 s	67.47	66.98
2 s	67.82	67.34
3 s	68.13	67.72
5 s	68.11	67.72
10 s	66.60	65.83

Latency spans image/context readiness to complete answer output, excluding frame acquisition, observation maintenance, memory operations, action execution/rendering, and network/queueing delays. Overall latency averages the two domains, each with 10,000 QA turns (Appendix D.3).

5.2 Main Results

All streaming models in Table 1 are also fine-tuned on our QA dataset; Cambrian-S is fine-tuned on our dataset's offline conversion as a spatial baseline (Cambrian-S-FT).

MemAct-9B achieves the highest Overall ACC (67.27) and leads all three QA categories in both ACC and MRA among open-source, spatial, and streaming models. It exceeds StreamingVLM by 5.88/11.58 percentage points in Overall ACC/MRA; MemAct-4B also improves both metrics, particularly on Memory and Action QA. GPT-6 Astra leads Current and Action QA and Overall MRA, but MemAct-9B achieves comparable aggregate performance: +0.53 points in Overall ACC and -0.92 in MRA. Its highest Memory ACC (68.11) exceeds Astra by 3.68 points. MemAct thus combines competitive spatial QA with incremental observation processing and, for MemAct-9B, answer-generation latency of 0.420 seconds per QA turn. MemAct-9B also leads all three Timing metrics (85.3 T, 57.4 JA, and 50.9 JM), combining timely responses with accurate answers.

5.3 Ablation Studies

Memory Ablations. Table 2 ablates memory content, retention, and write frequency. Full memory improves ACC/MRA over no explicit memory by 13.59/15.48 points; ground-truth memory adds another 2.66/3.34 points, indicating room for better memory construction. Historical RGB contributes most: removing it costs 3.19/4.00 ACC/MRA points. Counts mainly benefit numerical answers, while relations yield smaller ACC gains.

Table 3: MemAct-9B Action QA ablations: (1) RL algorithm, (2) reward design, and (3) action-step budget.

Setting	ACC (%) (Δ pp)	MRA (%) (Δ pp)	Evidence gain (points)	Effective gain rate (%)	Mean steps
A. Training strategy					
No action	47.45 (+0.00)	35.27 (+0.00)	0.00	0.00	0.00
SFT only	52.92 (+5.47)	41.87 (+6.60)	14.26	43.41	2.19
SFT + Bandit (full)	58.28 (+10.83)	48.74 (+13.47)	29.13	71.14	2.02
SFT + GRPO	57.92 (+10.47)	48.20 (+12.93)	28.34	69.78	2.08
B. Reward composition (cumulative)					
Pixel visibility	51.39 (+3.94)	39.63 (+4.36)	10.40	35.13	2.31
+ Full framing	52.55 (+5.10)	43.11 (+7.84)	17.79	50.90	2.20
+ Two-target evidence	55.81 (+8.36)	45.03 (+9.76)	21.69	58.07	2.14
+ Historical gain (full)	58.28 (+10.83)	48.74 (+13.47)	29.13	71.14	2.02
C. Action budget					
No action (0 steps)	47.45 (+0.00)	35.27 (+0.00)	0.00	0.00	0.00
At most 1 step	54.61 (+7.16)	43.26 (+7.99)	17.85	50.85	0.94
At most 2 steps	56.93 (+9.48)	46.73 (+11.46)	25.45	64.84	1.57
At most 3 steps (full)	58.28 (+10.83)	48.74 (+13.47)	29.13	71.14	2.02

Extending retention from 10 seconds to full history consistently raises ACC/MRA to 68.11/67.72, supporting evidence use beyond recent observations. More frequent writes do not consistently help: three-second intervals maximize ACC, but our five-second interval matches MRA with only 0.02 points lower ACC and fewer writes; ten-second intervals reduce both metrics.

Action Ablations. Table 3 compares training, rewards, and budgets using common evidence-gain metrics and mean camera-action attempts, including invalid attempts. Cumulative-return optimization adds 5.36/6.87 ACC/MRA points over SFT-only training and reduces mean attempts from 2.19 to 2.02. GRPO (Shao et al., 2024) achieves comparable performance. Reward ablations share terminal answer rewards, action costs, and cumulative returns. Framing adds 1.16/3.48 ACC/MRA points over pixel visibility alone; joint target evidence adds 3.26/1.92 points, supporting complete views and balanced evidence for relational questions. Replacing current-view gains with gains over the best historical view adds 2.47/3.71 points and raises the effective gain rate from 58.07% to 71.14%, encouraging acquisition of new evidence. Raising the action budget from one to three attempts adds 3.67/5.48 ACC/MRA points, supporting multi-step exploration.

5.4 Evaluation on Other Spatial Benchmarks

We evaluate MemAct-9B on VSI-Bench (Yang et al., 2024a), ReVSI (Zhang et al., 2026), MMSI-Bench (Yang et al., 2025b), and ViewSpatial (Li et al., 2025), comparing against publicly reported results (Li et al., 2026; Zhang et al., 2026; Dai et al., 2026; Wang et al., 2026; TaichuAI, 2026; SpatialAxiom, 2026) in Table 4. Despite operating as a streaming model, MemAct-9B remains competitive on these spatial benchmarks. It leads the listed models on ViewSpatial (52.21), outperforms both specialized spatial baselines on ReVSI and MMSI-Bench, and obtains 65.90 on VSI-Bench, behind Cambrian-S-7B's 67.50. These results suggest that our training supports spatial understanding beyond our benchmark, while retaining incremental observation processing.

Table 4: MemAct versus published results on spatial understanding benchmarks.

Model	VSI-Bench	ReVSI	MMSI-Bench	View Spatial
Closed-source models				
GPT-5.2	54.49	50.90	41.30	47.30
Gemini-3-Pro	55.10	60.90	45.20	50.40
Open-source models				
Qwen3.5-9B	57.90	51.60	36.50	48.20
Spatial models				
VLM-3R-7B	60.90	50.10	27.90	40.50
Cambrian-S-7B	67.50	49.10	27.10	41.30
Our method				
MemAct	65.90	55.30	37.80	52.21

6 Conclusion

MemAct combines structured memory writing, historical evidence retrieval, and active exploration in a streaming vision–language model for spatial understanding. The MemAct suite supports training and evaluation in real and simulated environments. SFT teaches memory and interaction behaviors; cumulative-return counterfactual optimization improves exploration and stopping decisions. Experiments show that historical evidence and newly acquired views improve spatial reasoning. Future work will extend discrete camera commands to more continuous, comprehensive action spaces and adapt MemAct to larger, more cluttered, and more dynamic scenes.

AI USE STATEMENT

AI tools assisted discussions of research ideas, code implementation, and manuscript preparation, including literature search and checking, language revision, figure editing, and experiment evaluation. Tools and models used included Codex, GPT-5.6-sol, GPT-6-Astra, and Gemini-3.8-Flash. Dataset construction also used LLM-assisted QA wording refinement while preserving grounded referents and geometry-derived answers. Separately, the benchmark evaluates language and vision–language models as experimental subjects. The authors directed these uses, reviewed and revised the assisted content, and take responsibility for the final paper and its scientific claims.

REPRODUCIBILITY STATEMENT

Data construction and quality control are described in Appendix A, with training-data composition and sampling in Appendix B. Appendix C specifies memory storage and retrieval, and Appendix D provides training settings, action-optimization objectives, and evaluation protocols, including the separate Timing evaluation. Qualitative examples are provided in Appendix E. We will provide supplementary files to support these descriptions and facilitate reproduction.

REFERENCES

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D Question Answering for Spatial Scene Understanding. *arXiv preprint arXiv:2112.10482*, 2021. doi: 10.48550/arXiv.2112.10482. URL https://arxiv.org/abs/2112.10482.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data. *arXiv preprint arXiv:2111.08897*, 2021. doi: 10.48550/arXiv.2111.08897. URL https://arxiv.org/abs/2111.08897.

Blender Foundation. Cycles. Official renderer documentation, 2026. URL https://www.cycles-renderer.org/. Accessed 2026-09-24.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. *arXiv preprint arXiv:2401.12168*, 2024. doi: 10.48550/arXiv.2401.12168. URL https://arxiv.org/abs/2401.12168.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. LiveCC: Learning Video LLM with Streaming Speech Transcription at Scale. *arXiv preprint arXiv:2504.16030*, 2025. doi: 10.48550/arXiv.2504.16030. URL https://arxiv.org/abs/2504.16030.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models. *arXiv preprint arXiv:2406.01584*, 2024. doi: 10.48550/arXiv.2406.01584. URL https://arxiv.org/abs/2406.01584.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. *arXiv preprint arXiv:1702.04405*, 2017. doi: 10.48550/arXiv.1702.04405. URL https://arxiv.org/abs/1702.04405.

Yalun Dai, Hao Li, Shulin Tian, Runmao Yao, Yuhao Dong, Fangzhou Hong, Zhaoxi Chen, Fangfu Liu, Baoliang Tian, Dingwen Zhang, Tao Wang, Kim-Hui Yap, and Ziwei Liu. S-Agent: Spatial Tool-Use Elicits Reasoning for Spatial Intelligence. *arXiv preprint arXiv:2606.20515v1*, 2026. doi: 10.48550/arXiv.2606.20515. URL https://arxiv.org/abs/2606.20515v1.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. *arXiv preprint arXiv:1711.11543*, 2017. doi: 10.48550/arXiv.1711.11543. URL https://arxiv.org/abs/1711.11543.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin 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. doi: 10.48550/arXiv.2505.20279. URL https://arxiv.org/abs/2505.20279.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D Furnished Rooms With layOuts and semaNTics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10933–10942, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Fu_3D-FRONT_3D_Furnished_Rooms_With_layOuts_and_semaNTics_ICCV_2021_paper.html.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive. *arXiv preprint arXiv:2404.12390*, 2024. doi: 10.48550/arXiv.2404.12390. URL https://arxiv.org/abs/2404.12390.

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. doi: 10.48550/arXiv.2507.12846. URL https://arxiv.org/abs/2507.12846.

Google DeepMind. Gemini 3.8 Flash Model Card. Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Accessed 2026-09-24.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: Visual Question Answering in Interactive Environments. *arXiv preprint arXiv:1712.03316*, 2017. doi: 10.48550/arXiv.1712.03316. URL https://arxiv.org/abs/1712.03316.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. *arXiv preprint arXiv:2309.16650*, 2023. doi: 10.48550/arXiv.2309.16650. URL https://arxiv.org/abs/2309.16650.

Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi: 10.1109/TSSC.1968.300136. URL https://doi.org/10.1109/TSSC.1968.300136.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. *arXiv preprint arXiv:2404.05726*, 2024. doi: 10.48550/arXiv.2404.05726. URL https://arxiv.org/abs/2404.05726.

Yining Hong, Jiageng Liu, Han Yin, Manling Li, Leonidas Guibas, Li Fei-Fei, Jiajun Wu, and Yejin Choi. ESI-Bench: Towards Embodied Spatial Intelligence that Closes the Perception-Action Loop. *arXiv preprint arXiv:2605.18746*, 2026. doi: 10.48550/arXiv.2605.18746. URL https://arxiv.org/abs/2605.18746.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. doi: 10.48550/arXiv.2106.09685. URL https://arxiv.org/abs/2106.09685.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual Language Maps for Robot Navigation. *arXiv preprint arXiv:2210.05714*, 2022. doi: 10.48550/arXiv.2210.05714. URL https://arxiv.org/abs/2210.05714.

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. *arXiv preprint arXiv:2501.00584*, 2024. doi: 10.48550/arXiv.2501.00584. URL https://arxiv.org/abs/2501.00584.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set Multimodal 3D Mapping. *arXiv preprint arXiv:2302.07241*, 2023. doi: 10.48550/arXiv.2302.07241. URL https://arxiv.org/abs/2302.07241.

Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 9091–9101. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00850. URL https://doi.org/10.1109/ICCV51701.2025.00850.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models. *arXiv preprint arXiv:2505.21500*, 2025. doi: 10.48550/arXiv.2505.21500. URL https://arxiv.org/abs/2505.21500.

Yunheng Li, Guohong Mu, Hao Li, Shengsheng Qian, Dingwen Zhang, Qibin Hou, and Ming-Ming Cheng. Annotations as Rollouts: Efficient and Scalable Reinforcement Learning for Video MLLMs. *arXiv preprint arXiv:2608.20492v1*, 2026. doi: 10.48550/arXiv.2608.20492. URL https://arxiv.org/abs/2608.20492v1.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding. *arXiv preprint arXiv:2411.03628*, 2024. doi: 10.48550/arXiv.2411.03628. URL https://arxiv.org/abs/2411.03628.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017. doi: 10.48550/arXiv.1711.05101. URL https://arxiv.org/abs/1711.05101.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated Question Answering in 3D Scenes. *arXiv preprint arXiv:2210.07474*, 2022. doi: 10.48550/arXiv.2210.07474. URL https://arxiv.org/abs/2210.07474.

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, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Majumdar_OpenEQA_Embodied_Question_Answering_in_the_Era_of_Foundation_Models_CVPR_2024_paper.html.

OpenAI. GPT-6 Astra System Card. System card, 2026. URL https://deploymentsafety.openai.com/gpt-6-astra. Accessed 2026-09-24.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human

feedback. *arXiv preprint arXiv:2203.02155*, 2022. doi: 10.48550/arXiv.2203.02155. URL https://arxiv.org/abs/2203.02155.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. Official release; citation recommended by the Qwen3.5 model cards, 2026a. URL https://qwen.ai/blog?id=qwen3.5. Accessed 2026-09-24.

Qwen Team. Qwen3.8-27B. Official model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed 2026-09-24.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models. *arXiv preprint arXiv:2412.07755*, 2024. doi: 10.48550/arXiv.2412.07755. URL https://arxiv.org/abs/2412.07755.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics. *arXiv preprint arXiv:2411.16537*, 2024. doi: 10.48550/arXiv.2411.16537. URL https://arxiv.org/abs/2411.16537.

SpatialAxiom. SpatialAxiom-9B: Official Model Card and Benchmark Results. Official model card, 2026. URL https://huggingface.co/SpatialAxiom/SpatialAxiom-9B. Accessed 2026-09-24.

TaichuAI. ZDTaichu5.0-9B: Official Model Card and Benchmark Results. Official model card, 2026. URL https://huggingface.co/TaichuAI/ZDTaichu5.0-9B. Accessed 2026-09-24.

Chongyu Wang, Ting Huang, Chunyu Sun, Xinyu Ning, Di Wang, and Hao Tang. Let Geometry GUIDE: Layer-wise Unrolling of Geometric Priors in Multimodal LLMs. *arXiv preprint arXiv:2604.05695v1*, 2026. doi: 10.48550/arXiv.2604.05695. URL https://arxiv.org/abs/2604.05695v1.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025a. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

Yifei Wang, Zhenkai Li, Tianwen Qian, Huanran Zheng, Zheng Wang, Yuqian Fu, and Xiaoling Wang. StreamEQA: Towards Streaming Video Understanding for Embodied Scenarios. *arXiv preprint arXiv:2512.04451*, 2025b. doi: 10.48550/arXiv.2512.04451. URL https://arxiv.org/abs/2512.04451.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Yao Lu, and Song Han. StreamingVLM: Real-Time Understanding for Infinite Video Streams. *arXiv preprint arXiv:2510.09608*, 2025. doi: 10.48550/arXiv.2510.09608. URL https://arxiv.org/abs/2510.09608.

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. *arXiv preprint arXiv:2412.14171*, 2024a. doi: 10.48550/arXiv.2412.14171. URL https://arxiv.org/abs/2412.14171.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards Spatial Supersensing in Video. *arXiv preprint arXiv:2511.04670*, 2025a. doi: 10.48550/arXiv.2511.04670. URL https://arxiv.org/abs/2511.04670.

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. doi: 10.48550/arXiv.2505.23764. URL https://arxiv.org/abs/2505.23764.

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. *arXiv preprint arXiv:2411.17735*, 2024b. doi: 10.48550/arXiv.2411.17735. URL https://arxiv.org/abs/2411.17735.

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. *arXiv preprint arXiv:2504.17343*, 2025. doi: 10.48550/arXiv.2504.17343. URL https://arxiv.org/abs/2504.17343.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. *arXiv preprint arXiv:2308.11417*, 2023. doi: 10.48550/arXiv.2308.11417. URL https://arxiv.org/abs/2308.11417.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. *arXiv preprint arXiv:2312.03275*, 2023. doi: 10.48550/arXiv.2312.03275. URL https://arxiv.org/abs/2312.03275.

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. doi: 10.48550/arXiv.2606.06891. URL https://arxiv.org/abs/2606.06891.

Mingliang Zhai, Zhi Gao, Yuwei Wu, and Yunde Jia. Memory-Centric Embodied Question Answering. *arXiv preprint arXiv:2505.13948*, 2025. doi: 10.48550/arXiv.2505.13948. URL https://arxiv.org/abs/2505.13948.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Memory-Based Real-Time Understanding for Long Video Streams. *arXiv preprint arXiv:2406.08085*, 2024. doi: 10.48550/arXiv.2406.08085. URL https://arxiv.org/abs/2406.08085.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. ReVSI: Rebuilding Visual Spatial Intelligence Evaluation for Accurate Assessment of VLM 3D Reasoning. *arXiv preprint arXiv:2604.24300*, 2026. doi: 10.48550/arXiv.2604.24300. URL https://arxiv.org/abs/2604.24300.

APPENDIX GUIDE

The linked guide below summarizes the supplementary material.

- **Dataset Construction Details** (Appendix A): observation alignment, evidence measurements, memory and QA targets, action trajectories, and question templates.
- **Training Data Composition and Sampling** (Appendix B): episode counts, training-data sources and composition, and action-length statistics.
- **Memory Router Details** (Appendix C): memory writes, category/time queries, pagination, and image retrieval.
- **Training and Evaluation Details** (Appendix D): optimization, action rewards and rollouts, QA scoring, and timing evaluation.
- **Qualitative Examples** (Appendix E): current-observation, historical-evidence, active-view, and memory-write examples.

A DATASET CONSTRUCTION DETAILS

This section details how observations are converted into supervision for memory writing, spatial question answering, and camera actions. We first establish a shared temporal representation, then describe target construction and filtering, and finally specify the action space and multi-step trajectories. The question catalog provides examples and construction rules for individual templates.

A.1 OBSERVATION PREPARATION AND ALIGNMENT

Simulation episodes use existing 3D-FRONT (Fu et al., 2021) layouts and assets. We remove objects with conspicuous interpenetration or erroneous scale while preserving the remaining scene structure. Scanning trajectories connect geometric centers and selected centers of maximal inscribed circles in randomized order, with the camera oriented toward the room interior. A* (Hart et al., 1968) plans paths on a mesh-derived occupancy map over the camera's relevant height range, and Bézier curves smooth the resulting paths. Multiple trajectories provide different visitation orders and observation histories for each layout. Floor and ceiling textures and randomly selected HDRI environments vary appearance, and Cycles (Blender Foundation, 2026) renders the RGB observations and auxiliary visibility annotations.

Real observations originate from ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021) videos. Frames are aligned with camera poses and source 3D annotations through physical timestamps. Coordinate transformations and projection associate annotated objects with individual observations. Both branches provide time-ordered RGB frames, category labels, instance correspondences, geometry, and per-frame visibility. We sample observations at 1 Hz and express geometric distances in meters. Per-frame visibility follows the source annotations. Pixel-level measurements are used where rendered masks are available; projected bounding-box area is separately identified when used as an evidence proxy.

For a question at time t, historical evidence is selected from the observation prefix $O_{\leq t}$. An action-generated observation becomes available only after the corresponding command is executed. Offline geometry and instance identifiers determine supervision, whereas the model receives RGB observations, timestamps, retained context, and returned tool content. Each example records its question time, evidence-frame references, answer, and construction provenance.

A.2 METADATA AND EVIDENCE MEASUREMENTS

Temporal metadata maps frame references to timestamps. Camera metadata contains intrinsics and poses; object metadata supplies persistent identities, categories, 3D centers, dimensions, and available meshes or boxes. Scene metadata supplies geometry, occupancy, and floor polygons. Visibility metadata links these records to individual observations. Distances use meters, image fractions are dimensionless, and each evidence measurement records its provenance.

Visibility and image occupancy. Let I be the image rectangle, $N = WH$ its pixel count, $b_o(x)$ the source visibility indicator, and $n_o(x)$ the object's visible mask-pixel count. Pixel visibility is $p_o(x) = n_o(x)/N$. For observations using projected-box evidence, let $R_o(x)$ be the image-plane bounding rectangle of the projected world-axis-aligned 3D box. The occupancy proxy is

$$u_o(x) = b_o(x)\frac{\text{area}(R_o(x) \cap I)}{WH}. \tag{4}$$

The indicator gates the proxy, but box area does not measure the unoccluded object surface. Missing source visibility is distinct from an explicit negative observation.

Framing completeness. Transform the corners of the mesh-local bounding box to the world and camera frames, project them using camera intrinsics, and form their convex hull $H_o(x)$. Framing completeness is

$$f_o(x) = \frac{\text{area}(H_o(x) \cap I)}{\text{area}(H_o(x))}. \tag{5}$$

A valid ratio requires positive projected area and a box entirely beyond the 0.01 m near plane. Near-plane intersections are retained only as diagnostics, not valid framing scores; fully behind-camera and degenerate projections are unavailable. Framing describes image-boundary clipping, not occlusion by other objects.

Occlusion support. An isolated-object render uses the same camera and target geometry while hiding other scene objects. If its target mask contains $n_o^{\text{iso}}(x) > 0$ pixels, define

$$v_o^{\text{occ}}(x) = \min\{1, n_o(x)/n_o^{\text{iso}}(x)\}. \tag{6}$$

The ratio is unavailable when the isolated mask is missing or empty, or the unbounded ratio exceeds 1.01. Values up to 1.01 are capped at one to accommodate small rasterization differences. This measures visible support relative to the isolated in-frame silhouette; framing is checked separately. Required unavailable measurements fail the corresponding quality check rather than being replaced with zeros.

A.3 Incremental Memory Targets

Memory targets are generated over consecutive five-second windows. Each target contains one observation record per sampled frame, including frames with no visible objects. A record specifies the frame reference and visible category counts. Optional relations describe image-plane left, right, above, or below relationships, with subjects and objects localized on a 3×3 image grid. The final incomplete window is also retained.

Let $V_{f,c}$ denote the set of visible instances of category c in frame f. The per-frame count is $|V_{f,c}|$, whereas the historical distinct-instance count is

$$N_c(t) = \left| \bigcup_{f:\, t_f \le t} V_{f,c} \right|. \tag{7}$$

The latter is computed using offline instance correspondences. For example, observing the same two chairs in three frames yields per-frame counts of two and a historical count of two. Memory targets preserve frame references so that detailed appearance and instance evidence can be recovered from historical images.

A.4 Spatial QA Construction

Each QA example is constructed by selecting a causal observation prefix, binding its referents, computing an answer from aligned annotations, checking evidence and ambiguity conditions, and instantiating a question template. Multiple-choice questions use distinct semantic alternatives and retain the mapping between option letters and answers. Numeric questions specify their units. All QA text undergoes one LLM polishing pass to diversify phrasing while preserving referents, temporal constraints, units, option meanings, and annotation-derived answers.

Current QA. Current questions use the latest observation or an explicitly specified recent window, with a default duration of five seconds. Camera-to-object and object-to-object distances use 3D centers. Trajectory length sums consecutive camera translations, endpoint displacement measures the distance between the initial and final positions, and orientation change measures the net endpoint rotation. Referents must be uniquely identifiable under the question wording. Nearest-camera queries require a minimum 0.15 m separation between the two closest candidates; designated-object nearest-neighbor queries require 0.40 m. Appearance and disappearance questions require a unique category change between the specified observations.

Memory QA. Memory questions require evidence preceding the current window. First and last appearances are computed over the full causal prefix, historical counts use instance-set unions, and historical relationships are evaluated at the referenced frame. Appearance-order questions require successive first-observation times to differ by at least two seconds. Historical-neighbor questions bind a uniquely observed target that is absent from the current view and select neighbors among objects co-visible at the specified historical time. These evidence conditions define the question family; the model can answer directly when its retained context already contains the required information.

Chained questions and geometric examples. For linked questions, the first turn establishes an object or category and subsequent turns retain that referent. These linked QA turns are distinct from multiple camera actions taken to answer a single question. Each turn is associated with its own arrival time and evidence prefix. Height questions compare vertical extent rather than elevation above the floor, with a 0.30 m margin between the two tallest candidates. Room-area questions use the room's floor polygon, including floor beneath furniture, and require at least eight observations, 90% boundary-angular coverage, and a viewing-direction span of at least 150°.

A construction example. Consider a nearest-object question with visible object-center distances of 1.0, 1.4, and 2.1 m. The first object supplies the answer because its 0.4 m lead exceeds the 0.15 m ambiguity margin. Replacing the second distance with 1.08 m rejects the example. For a first-appearance question, the answer is the earliest visible frame in the complete prefix, and the example is assigned to Memory QA when that evidence precedes the current window.

A.5 Selecting Question Times and Object Referents

Construction operates on aligned frame records rather than RGB alone. Each record links a timestamp and image to its camera pose, visible-instance set, and category labels. Object records supply persistent instance identifiers, 3D centers, and the geometric fields required by each template; image-plane relations additionally use projected positions, and room-area questions use floor polygons. These annotations select candidates and compute labels offline. Public questions identify objects through category names, explicit nearest-instance qualifiers, or image regions, without exposing instance identifiers or geometric answers.

Candidate times and evidence windows. We enumerate candidate times on the sampled observation sequence and apply template-specific temporal conditions before constructing a question. Single-view questions use the question frame; motion and visibility-change questions require the specified earlier endpoint. Nearest-neighbor, fixed-reference relation, and height-comparison templates enumerate interior frames. Last-seen, historical coappearance, historical direction, and appearance-order templates are proposed every five sampled frames starting at index ten. Memory evidence must belong to the available prefix and be covered by completed memory writes. Room-area candidates are proposed when a continuous within-room observation run meets the coverage conditions. A frame that does not satisfy a template's conditions produces no example for that template.

Object eligibility and combinations. For nearest-neighbor, fixed-reference relation, and height-comparison templates, candidate objects are visible at the question frame and its two adjacent frames, and each selected category has exactly one visible instance at the question frame. The adjacent-frame test selects stable observations offline; the following frame is not included in the question's model input or used to compute its geometric answer. Category-conflict checks remove semantically overlapping options. We then enumerate reference–target pairs for relations

and nearest-neighbor questions, or four distinct-category objects for height comparison. Missing geometric fields remove the affected combination. Other templates retain their own bindings: a nearest-instance qualifier can resolve multiple objects of one category, whereas historical single-instance questions require a unique previously observed instance.

Answer computation and acceptance. For a bound combination, we first compute the geometric or temporal quantity, then check the template's uniqueness and evidence conditions. A designated-object nearest-neighbor question requires a 0.40 m gap between its two closest eligible neighbors. Height comparison uses world-vertical bounding-box extent and requires a 0.30 m gap between the two tallest objects. Fixed-reference left/right questions use the 3×3 image grid: the correct target lies in the reference object's row and on the requested side; distractors come from the opposite side or categories absent from the scene. The category alternatives are filtered before option letters are assigned. These conditions apply to the corresponding templates, rather than serving as universal thresholds for all QA.

Historical selection and memory verification. The last-seen template selects a category absent at the question frame whose latest occurrence is at least five frame indices earlier. That historical frame also supplies candidates for coappearance and relative-direction questions, with its timestamp stated in the question. Coappearance distractors are absent from the referenced frame. Direction questions require unique category instances and a same-row or same-column grid relation. First-appearance-order questions select three categories whose first sightings occur at least two seconds apart, with the latest first sighting at least five frame indices before the question. We replay the corresponding teacher memory writes and retrieval requests, reconstruct the answer from returned records or relations, and require agreement with the annotation-derived label. Direction examples additionally require the referenced image to be retrievable within the image budget.

Room evidence and retained records. Room-area construction uses simulation episodes with a valid floor polygon and unambiguous camera-to-room assignment. Leaving the assigned room resets the observation run. Each polygon edge contributes eight boundary samples; coverage counts samples falling within the horizontal viewing angle of at least one observation in the run. We require eight observations, 90% angular boundary coverage, and a 150° heading span. This is a geometric coverage proxy rather than a wall-pixel visibility test. Each retained QA records its template, question time, evidence-frame references, bound objects where applicable, answer, and source-scene provenance. Missing evidence images and failed template checks are excluded before serialization.

Worked current and memory examples. Consider an illustrative current nearest-neighbor example at $t = 12$ s. A cabinet, bed, chair, and lamp have unique category referents and pass the adjacent-frame visibility check. Their center distances from the cabinet are 1.0, 1.6, and 2.2 m. The bed is retained as the answer because the nearest-neighbor margin is 0.6 m; the question refers to the cabinet, and its evidence is the frame at 12 s. If the chair were 1.2 m away, the combination would fail the 0.40 m margin. For an illustrative memory example at $t = 20$ s, a lamp was last visible at 8 s and is absent afterward. Its last-seen question has answer 8 s, with the frame at 8 s as positive evidence and the intervening prefix establishing that no later sighting occurred. The teacher write covering that frame is replayed, and a lamp-category query must recover the same latest timestamp. A later lamp sighting would change the label rather than preserve the earlier answer.

A.6 Action-conditioned Data

Action data are constructed through dedicated rendering of one- to three-step discrete camera-action sequences. Each example fixes the starting observation, question, target objects, and geometry-derived answer. Every action is executed from the preceding camera state and produces a newly rendered observation under unchanged scene geometry, appearance, and imaging settings. Evidence and geometric checks select demonstration sequences; alternative action outcomes are rendered and recorded separately for counterfactual optimization.

Start times and target binding. At candidate question times in simulated observation streams, we select objects with available geometry, a resolvable initial-view referent, and an evidence-area score

of at least0.0002. Referents use a unique category or a unique category within a 3×3 image region. Up to sixteen objects are retained in descending score order. Each object supplies a size candidate, and each unordered pair supplies center-distance and distance-range candidates. Size is the longest physical bounding-box side, using source physical dimensions when available or transformed mesh dimensions otherwise; pair distance is the Euclidean distance between fixed 3D centers. The same bound objects and answer are retained throughout the path.

Evidence measurements and endpoint selection. Action construction uses mask-derived visible-pixel fractions for both pre-question observations and newly rendered action views. Corresponding target masks are required to evaluate the historical baseline, ensuring that evidence gains compare the same measurement. Let $p_j(x)$, $f_j(x)$, and $v_j^{\mathrm{occ}}(x)$ denote visible-pixel fraction, framing completeness, and occlusion support, respectively, as defined in Appendix A.2. For the required objects $\mathcal{J}_q$, define

$$P(x,q) = \min_{j \in \mathcal{J}_q} p_j(x), \qquad B_t(q) = \max_{x \in O_{\le t}} P(x,q). \tag{8}$$

The minimum requires joint visual support for pairwise questions, while the historical maximum prevents improvement from being measured only against a weak starting view. An endpoint x is retained only if

$$p_j(x) \ge 0.005, \quad f_j(x) \ge 0.80, \quad v_j^{\mathrm{occ}}(x) \ge 0.70 \qquad \forall j \in \mathcal{J}_q, \tag{9}$$

$$P(x,q) \ge 1.5 B_t(q), \qquad P(x,q) - B_t(q) \ge 0.002. \tag{10}$$

Required unavailable measurements fail the quality check. These conditions define a single selection profile for all rendered demonstrations.

A historical view is adequate when every required object satisfies the same visibility, framing, and occlusion thresholds in that view. We retain separate labels for examples lacking an adequate historical view and examples improving an already adequate view. The historical baseline includes all pre-question observations, regardless of whether they satisfy the adequacy thresholds.

Discrete action interface and multi-step interaction. The action interface consists of five commands: forward, leftward, and rightward translation by 0.5 m, and left and right rotation by 45°. Translation follows the current camera's horizontal coordinate frame and preserves orientation; rotation preserves position. The model selects a command, and the runtime assigns its fixed magnitude. Each returned observation is associated with the resulting camera state. Interactions allow at most three action attempts, each continuing from its predecessor state and producing a specifically rendered observation. Demonstrations terminate at the earliest endpoint satisfying the unified quality and gain criteria above. Intermediate views need not satisfy the endpoint thresholds or improve evidence monotonically, allowing preparatory actions.

Geometric checks and serialization. For each segment, a mesh-based check rejects a surface intersection along the straight camera-center connection. Positions sampled at intervals no greater than 0.025 m require at least 0.125 m surface clearance and support within 2 m below. The serialized example contains the causal prefix, unchanged question, ordered command–image pairs, and final answer; future trajectory images enter the model context only after their corresponding commands. Duplicate candidates sharing the source trajectory, bound target or pair, endpoint, and template are removed. Distance-range and proximity candidates share the same geometric quantity. Identical serialized sequences are deduplicated across these template labels, and the final release retains the distance-range label.

Rendered alternatives for counterfactual optimization. Selected action sequences provide SFT demonstrations. Counterfactual optimization additionally requires alternative outcomes for the five discrete commands at each evaluated state, rendered under the same scene and imaging settings. Demonstration endpoint thresholds do not filter these alternatives: valid actions with little or negative evidence improvement remain available for comparison, and physically invalid actions are recorded separately. Branch validation distinguishes missing rendered outcomes (coverage errors) from physically invalid actions, which preserve the observation and consume an attempt. RL rewards use the mask-derived measurements defined in Appendix D.2.

Worked three-step example. Consider a size question about a uniquely identified cabinet whose longest physical side is 1.2 m. Suppose its best historical visible-pixel fraction is 0.001, and three successive rendered views yield 0.002, 0.004, and 0.007. The first two fail the 0.005 visibility threshold. The third achieves a gain ratio of 7 and an absolute gain of 0.006; with framing completeness 0.85 and occlusion support 0.80, it satisfies all endpoint conditions. If all transitions pass geometric checks, we retain the causal prefix, fixed question, three discrete command–image pairs, and answer 1.2 m. A failed transition rejects this sequence; if an earlier view satisfies every condition, the demonstration ends there instead.

A.7 Assembly and Quality Control

We assemble observations, memory targets, questions, and tool outcomes into causally ordered training sequences. The training release contains all retained QA examples; the main test and Timing sets are separate evaluation collections. Validation checks timestamps, frame references, referent consistency, answer computations, duplicate semantic questions, tool-image correspondence, and interaction budgets. Examples with missing required evidence, failed constraints, or excessive sequence length are excluded. Public model inputs and private annotation fields are stored separately.

A.8 Question Catalog and Construction Examples

Tables 5, 6, and 7 list the Current, Memory, and Action QA templates, respectively, with representative questions, semantic answers, and construction rules. Current templates use present or recent evidence; Memory templates require historical observations; Action templates specify questions supported by newly acquired views. For multiple-choice examples, we display the semantic answer rather than its option letter. Linked templates show the opening question and specify the subsequent turn in the construction rule.

Table 5: Current QA templates and examples.

Template	Example and construction
approached object category	**Question:** What is the category whose object had the largest decrease in camera-to-3D-center distance among objects visible both now and 5 seconds ago? Choose from the listed options. **Answer:** storage unit. **Construction:** Compare objects visible at both endpoints. Require a distance decrease of at least 0.20 m and a 0.10 m lead over the runner-up.
became visible category	**Question:** Select the option that identifies the category that became visible between 5 seconds ago and now. **Answer:** plant. **Construction:** Take the difference between endpoint category sets and require one newly visible category.
camera motion bucket	**Question:** What is the range of the camera's straight-line endpoint displacement over the last 5 seconds? Choose from the listed options. **Answer:** almost stationary (at most 0.2 meters). **Construction:** Discretize straight-line camera endpoint displacement, excluding ambiguous interval boundaries.
camera path length m	**Question:** Estimate the total length of the camera's trajectory over the last 5 seconds. Answer in meters. **Answer:** 0.0. **Construction:** Sum Euclidean translations between consecutive camera positions in the specified window.
camera rotation deg	**Question:** Provide your numerical estimate of the net difference between the camera's starting and ending orientations over the last 5 seconds. Answer in degrees. **Answer:** 120.0. **Construction:** Compute net rotation between endpoint camera orientations, expressed in degrees.

Template	Example and construction
camera to object distance bucket	**Question:** What is the camera-to-3D-center distance range for the closest visible sofa? Choose from the listed options. **Answer:** very near (at most 1.25 meters). **Construction:** Bind the nearest visible instance of the named category and discretize its center distance to the camera.
camera to object distance m	**Question:** Give a numerical value for the camera-to-3D-center distance to the closest visible sofa. Answer in meters. **Answer:** 1.06. **Construction:** Bind the nearest visible instance of the named category and compute its center distance to the camera.
closer than before	**Question:** Select the option that identifies the change in camera-to-3D-center distance to the table over the last 5 seconds, considering the object visible at both times. **Answer:** much closer (decrease of at least 0.75 meters). **Construction:** Uniquely match the category referent across the two observations and classify its camera-distance change.
closest category then distance	**Question:** Choose the answer corresponding to the category of the currently visible object closest to the camera by 3D center distance. **Answer:** sofa. **Construction:** First identify the nearest visible category; then query the distance interval of the same object. Count both QA turns.
closest category then relation	**Question:** Determine the category of the currently visible object closest to the camera by 3D center distance and select the matching choice. **Answer:** sofa. **Construction:** First bind the nearest visible object; then identify its nearest neighbor among currently co-visible objects.
closest neighbor category	**Question:** Which option describes the listed category whose currently visible object has its 3D center closest to the cabinet? **Answer:** bed. **Construction:** Require visibility at the question frame and its adjacent frames for offline stability selection, and unique category referents; the nearest two candidates must differ by at least 0.40 m.
closest object category	**Question:** Choose the answer corresponding to the category of the currently visible object closest to the camera by 3D center distance. **Answer:** sofa. **Construction:** Rank currently visible objects by camera-to-center distance; require a nearest-candidate margin of 0.15 m.
disappeared category	**Question:** From the options below, indicate the category that disappeared from view between 5 seconds ago and now. **Answer:** sofa. **Construction:** Take the reverse endpoint category-set difference and require one disappeared category.
fixed reference relation	**Question:** From the options below, indicate the listed category whose currently visible object is to the left of the desk in the image at the time of this question. **Answer:** double bed. **Construction:** Use projected image positions with unique referents; select a same-row left/right relation and place distractors on the opposite side.
multi candidate height	**Question:** Which option describes the listed category whose currently visible object has the greatest vertical height (not elevation above the floor)? **Answer:** wardrobe. **Construction:** Compare world-vertical bounding-box extents of four distinct-category candidates; require a 0.30 m gap between the tallest two.

Template	Example and construction
object category then later visibility	**Question:** Determine the category of the currently visible object closest to the camera by 3D center distance and select the matching choice. **Answer:** media unit. **Construction:** Bind a category/object in the first turn and ask about its visibility only after the later observation arrives.
object relative to camera	**Question:** Which option describes the direction of the closest visible bed relative to the camera now? **Answer:** far to the right. **Construction:** Transform the bound object center into the current camera frame and assign the corresponding direction category.
room floor area	**Question:** Calculate the total floor-footprint area of the room just surveyed that the camera is currently inside, including floor beneath furniture from the observations. Answer in square meters. **Answer:** 8.43. **Construction:** Compute the valid room floor-polygon area. Require at least eight frames, 90 percent boundary-angular coverage, and a 150-degree heading span.

Table 6: Memory QA templates and examples.

Template	Example and construction
first appearance order	**Question:** Identify the first-observation order of the categories armchair, dining table, storage unit since this episode began by selecting an option. **Answer:** storage unit $\rightarrow$ dining table $\rightarrow$ armchair. **Construction:** Sort three category-first-observation times, separated by at least two seconds and preceding the current window.
first seen category then timestamp	**Question:** Determine the category seen earliest among the listed choices so far in this episode and select the matching choice. **Answer:** lighting. **Construction:** Identify the earliest category among the options, then query its original first-observation timestamp.
first seen object category	**Question:** Among the choices, which one gives the category seen earliest among the listed choices so far in this episode? **Answer:** lighting. **Construction:** Compare first-observation times of the listed categories; require a unique earliest option before the current window.
first seen timestamp s	**Question:** Estimate the first-seen timestamp of the lighting category, measured from the start of this episode. Answer in seconds. **Answer:** 0.0. **Construction:** Return the earliest observed timestamp for the category over the complete causal prefix.
historical coappearance category	**Question:** Choose the answer corresponding to the listed category also visible alongside the lighting at timestamp 3 seconds. **Answer:** cabinet. **Construction:** Use the explicitly referenced historical frame; the correct category is co-visible with the anchor and distractors are absent.
historical relative direction	**Question:** From the options below, indicate the image-plane position of the storage unit relative to the table in the observation at timestamp 44 seconds. **Answer:** left. **Construction:** Use uniquely bound objects in the historical frame and their same-row or same-column image-plane relationship.

Template	Example and construction
last seen timestamp	**Question:** Report the last timestamp at which the lighting category was observed before this question as a number. Answer in seconds. **Answer:** 44.0. **Construction:** Take the last occurrence before the question; require current absence and an occurrence preceding the pending write window.
seen count by category	**Question:** Report the number of distinct table objects observed so far in this episode as a number. Answer with an integer count (objects). **Answer:** 4. **Construction:** Union visible instance IDs over the prefix; require historical write coverage and that the current view does not contain all counted instances.
single seen object then closest category	**Question:** Choose the answer corresponding to the category nearest to that lighting by 3D center distance, among the other objects visible at timestamp 0 seconds alongside the lighting seen earlier and absent now. **Answer:** table. **Construction:** Bind the only previously observed instance of the target category, absent now; find its nearest co-visible historical neighbor.
single seen object then closest distance	**Question:** What is the 3D center distance from that lighting to its nearest neighbor, among the other objects visible at timestamp 0 seconds alongside the lighting seen earlier and absent now? Answer in meters. **Answer:** 0.776. **Construction:** Use the same historical target binding and return its minimum center distance to another object visible in the referenced frame.

Table 7: Action QA templates and examples.

Template	Example and construction
object size m	**Question:** Estimate the longest physical side length of the 3D bounding box of the drawer chest / corner cabinet visible in the initial view. Answer in meters. **Answer:** 1.185. **Construction:** Use source physical dimensions or transformed mesh bounding-box dimensions to obtain the longest physical side in meters; hold the bound object fixed across views.
object to object distance bucket	**Question:** Identify the 3D center-to-center distance range between the sofa in the middle left of the initial view and the table in the lower left of the initial view by selecting an option. **Answer:** near (1 to 1.75 meters). **Construction:** Map the bound pair's center distance to an unambiguous interval; require evidence for both objects.
object to object distance m	**Question:** Report the 3D center-to-center distance between the sofa in the lower left of the initial view and the chair visible in the initial view as a number. Answer in meters. **Answer:** 3.041. **Construction:** Compute the Euclidean distance between the bound 3D centers; retain the same pair throughout the action path.

B Training Data Composition and Sampling

We report source composition, supervision counts, and action-path statistics separately. A QA turn is one question–answer pair; a linked two-turn example contributes two QA turns. Memory writes count completed supervision windows, while camera-action counts refer to executed transitions. Unique questions, demonstration paths, and alternative transitions have separate denominators. Demonstration length counts executed camera actions.

Table 8: Episode composition. The simulation row covers Current and Memory QA; real-world rows count original recordings and unique retained segments.

Source	Layouts / recordings	Episodes / segments
3D-FRONT	4,823	14,534
ScanNet	1,412	1,511
ScanNet++	796	1,457
ARKitScenes	4,681	5,764
Real-world total	6,889	8,732

Table 9: Training QA composition by family. Counts denote answer turns.

Family	Training QA turns
Current	302,661
Memory	402,662
Action	294,677
Total	1,000,000

Sampling is organized hierarchically by QA family, semantic template, and source scene. Within each family, we balance templates over eligible examples and preserve scene diversity. Linked question groups are selected together, and memory-writing targets remain attached to their causal windows. Dataset composition and training exposure are reported separately: sampling an example multiple times does not increase the number of distinct QA examples.

The training release contains 1,000,000 QA answer turns: 302,661 Current QA, 402,662 Memory QA, and 294,677 Action QA. All of these turns are training data. MemAct-bench uses a separate 20,000-turn main QA test set and an additional 2,000-question Timing set. Training and evaluation sets are disjoint at the scene level. The 1,000,000 training QA turns are serialized into 342,868 streams covering 11,724 source-qualified scene identifiers. Retained sequences satisfy the 180,000-token budget; the longest contains 178,705 tokens. Table 9 breaks down training answer turns by QA family, while Table 10 separates their sources and shows that Action QA is simulation-only. Table 11 reports demonstration lengths; one-action paths are most common, with two- and three-action paths providing multi-step supervision. Table 12 reports training-set fractions, rather than repeated exposure during optimization, and Table 13 gives the finer template-level counts, including linked answer turns.

B.1 Episode Composition and Temporal Segmentation

For 3D-FRONT, an episode is an observation sequence associated with a layout, camera trajectory, and appearance configuration. Current and Memory QA use 14,534 simulated episodes from 4,823 layouts. The Action subset contributes 12 additional layouts outside this set, giving 4,835 simulated layouts across the complete release. Action observations are not counted as independent episodes in Table 8.

Long real-world recordings are divided into shorter temporal segments. The final release uses 8,732 unique segments from 6,889 original recordings. A recording is counted when at least one segment is retained, without requiring its full duration to be used. All retained segments reported here belong to the training collection. These recording identities are distinct from episode IDs and serialized training streams. Together, 4,835 simulated layout IDs and 6,889 real recording IDs account for the 11,724 source-qualified scene identifiers reported above; recording identifiers do not establish that every recording depicts a different physical environment.

Table 10: Training QA composition by source.

Source	Current	Memory	Action
3D-FRONT	135,671	218,983	294,677
ScanNet	31,304	30,431	0
ScanNet++	19,160	29,201	0
ARKitScenes	116,526	124,047	0

Table 11: Training action demonstration lengths after selection.

Actions	Training QA turns
1	216,418
2	46,858
3	31,401

Table 13: Template counts in the training release. Linked templates count all retained answer turns.

Template	Training QA turns
Current QA	
approached object category	18,733
became visible category	18,733
camera motion bucket	18,733
camera path length m	18,733
camera rotation deg	18,733
camera to object distance bucket	18,733
camera to object distance m	18,733
closer than before	18,733
closest category then distance	18,732
closest category then relation	18,732
closest neighbor category	18,739
closest object category	18,732
disappeared category	18,732
fixed reference relation	1,618
multi candidate height	1,316
object category then later visibility	18,732
object relative to camera	18,732
room floor area	18,732
Memory QA	
first appearance order	61,034
first seen category then timestamp	36,870
first seen object category	18,435
first seen timestamp s	18,435
historical coappearance category	60,843
historical relative direction	26,743
last seen timestamp	60,847
seen count by category	24,399
single seen object then closest category	47,528
single seen object then closest distance	47,528
Action QA	
object size m	128,469
object to object distance bucket	70,293
object to object distance m	95,915

Table 12: Training-set composition by QA family and source. Each fraction uses the 1,000,000 training QA turns as its denominator.

Group	QA turns	Fraction
Current	302,661	30.27%
Memory	402,662	40.27%
Action	294,677	29.47%
ARKitScenes	240,573	24.06%
ScanNet	61,735	6.17%
ScanNet++	48,361	4.84%
3D-FRONT	649,331	64.93%

C Memory Router Details

The memory router maintains model-written records and retrieves historical evidence through deterministic rules. Its responsibilities are structural validation, persistent storage, category/time filtering, and budgeted image return. The shared model supplies the semantic content of memory writes and formulates retrieval requests.

C.1 Writing and Storage

A memory write uses the fields `type`, `observations`, and `relations`. Observation entries contain a frame reference and a list of category–count pairs. The router requires contiguous half-open windows $[t_s, t_e)$, exactly one record per supplied frame, normalized lowercase category strings, positive integer counts, and valid relation endpoints. Empty object lists represent frames without visible objects. A valid write is stored with its frame/time mapping and a payload receipt. Repeating an identical write returns `already_stored`; a different payload for the same window is rejected as a conflicting retry. Semantic predictions are stored as generated by the model.

C.2 Queries, Pagination, and Images

Each question allows at most four memory-retrieval calls. A retrieval request specifies categories, an optional time range, and whether images are required. Category matching returns the union of matching records, with an additional co-visibility indicator for records containing all requested categories; an empty category list selects all categories. Results are ordered by time and paginated at eight records per page. A continuation cursor is bound to the original query and its record snapshot.

When images are requested, the router selects one representative per contiguous matching run, preferring co-visible records and the timestamp nearest their midpoint. Additional candidates are ordered by the summed counts of requested categories. Each call returns at most two images, with at most four distinct historical images per question. Returned images remain in context and are not resent. New questions reset pagination and image budgets; textual frame references remain available after the image budget is exhausted. These limits concern retrieved history, not action observations or the entire context.

Relations contain a frame, a predicate, and subject/object endpoints specified by category and image-grid region. Query time ranges are half-open. Category filtering selects frame records, and the response includes all relations attached to frames in the returned text page or image list. It does not independently filter relation endpoints by query category.

Illustrative protocol. The following payload illustrates a single-frame write at $t = 1$ s. A complete write contains one observation entry for every supplied frame in its window. The chair query returns this frame's category counts and the chair–table relation, together with the image if budget permits.

```
{"type":"memory_write",
 "observations":[{"frame":"f000001","objects":[
   {"category":"chair","count":1},
```

```
    {"category":"table","count":1}]}],
  "relations":[{"frame":"f000001",
    "subject":{"category":"chair","region":"middle_left"},
    "predicate":"left_of_in_image",
    "object":{"category":"table","region":"middle_right"}}]}
{"type":"use_memory","categories":["chair"],
  "time_range":[0,5],"with_frames":true}
```


D Training and Evaluation Details

Training first learns the shared output protocol through supervised fine-tuning and then optimizes action decisions using cumulative continuation returns. Both stages use the same vision–language model and structured outputs. This section specifies optimization, causal rollout construction, and answer scoring.

D.1 Shared Configuration and Supervised Fine-Tuning

We use Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026a) with LoRA (Hu et al., 2021) applied to the query and value projections, using rank eight, scaling factor sixteen, and zero dropout. Base-model parameters, including the visual encoder, are frozen. We use our dataset together with VSI-590K (Yang et al., 2025a) for SFT. Both stages use AdamW (Loshchilov & Hutter, 2017) with a learning rate of 10^{-5} and gradient clipping at 1.0. Training uses eight GPUs, one training item per GPU per microbatch, and eight gradient-accumulation steps, giving 64 items per update. SFT items are serialized sequences; action-optimization items are decision histories whose candidate rollouts are evaluated internally. Each serialized training sequence is limited to 180,000 tokens, including text, visual tokens, and supervised assistant outputs. Sequences exceeding this budget are discarded without truncation or splitting.

SFT sequences preserve the order of observations, memory triggers, questions, and tool results. Only assistant outputs contribute to the token loss: memory records, retrieval requests, camera commands, and answers. Prompt and tool-result tokens provide context. The loss is averaged over supervised assistant tokens within each sequence and then over the effective batch. Linked questions retain their original referents and causal context.

D.2 Cumulative-return Action Optimization

Action optimization evaluates camera and stopping decisions using evidence increments and terminal answer quality. Construction thresholds select examples; the following scores define training rewards and common evidence metrics.

View-level evidence. For the one or two required objects $\mathcal{O}_q$, use the visible-pixel fraction $p_o(x)$ and framing completeness $f_o(x)$ defined in Appendix A.2. Define

$$v_o(x) = \min\{p_o(x)/0.10, 1\}, \qquad s_o(x) = v_o(x) f_o(x), \tag{11}$$

$$E_{\text{pixel}}(x, q) = \frac{1}{|\mathcal{O}_q|} \sum_{o \in \mathcal{O}_q} v_o(x), \tag{12}$$

$$E_{\text{frame}}(x, q) = \frac{1}{|\mathcal{O}_q|} \sum_{o \in \mathcal{O}_q} s_o(x), \tag{13}$$

$$E_{\text{joint}}(x, q) = \min_{o \in \mathcal{O}_q} s_o(x). \tag{14}$$

Normalization saturates at ten percent image occupancy. Framing discounts cropped targets, and joint evidence is limited by the less-supported object. For single-object questions, framing and joint scores coincide. These reward inputs use mask-derived pixel fractions, not projected-box occupancy.

Historical gain and cumulative returns. Let k index camera-action decisions from zero and T denote the total number of attempted camera actions before answering; write $[z]_+ = \max(z, 0)$.

The full configuration maintains $\Phi(h, q) = \max_{x \in \mathcal{O}(h)} E_{\text{joint}}(x, q)$ over pre-question observations and preceding action results, with increment $d_k = \Phi(h_{k+1}, q) - \Phi(h_k, q)$. The first three ablation configurations instead use $d_k = [E(x_{k+1}, q) - E(x_k, q)]_+$ with their respective view score. Local increments can reward recovery of previously acquired evidence; historical gain removes this repeated reward.

$$r_k = 0.10 d_k - 0.03 - 0.20\mathbf{1}[\text{invalid}(a_k)], \qquad G_k = \sum_{j=k}^{T-1} r_j + R_{\text{QA}}. \tag{15}$$

We use unit action costs, discount $\gamma = 1$, and at most three camera-action attempts. Invalid actions preserve the observation, receive no evidence gain, and consume an attempt. Answer-now terminates without camera cost; budget exhaustion requests a final answer. R_{QA} scores the model's generated answer by exact option correctness or numerical MRA, as defined in Appendix D.3. In the full configuration, evidence increments telescope to $0.10[\Phi(h_T, q) - \Phi(h_0, q)]$, while action costs favor shorter paths with equal final evidence and answer quality.

Counterfactual candidate values. On rendered branch states, candidates are the five fixed-magnitude commands and `answer_now`. All candidates start from an independent copy of the same history and remaining total budget. We average $B = 4$ continuations per candidate:

$$\widehat{Q}^{\mu}(h, a) = \frac{1}{4} \sum_{b=1}^{4} G^{(b)}(h, a). \tag{16}$$

A frozen snapshot of the current policy is refreshed before each update. Subsequent camera decisions are sampled from its categorical policy at temperature one; terminal answers use greedy decoding. Training includes initial histories and visited nonterminal histories. No new memory retrieval is introduced inside these rollouts.

Candidate scores sum conditional log probabilities of decision-value tokens and are normalized across candidates without length normalization. The action loss is

$$\mathcal{L}_{\text{action}} = \mathbb{E}_h \left[-\sum_a \pi_\theta(a \mid h) \widehat{Q}^{\mu}(h, a) + 0.05 D_{\text{KL}}(\pi_\theta(\cdot \mid h) \| \pi_{\text{SFT}}(\cdot \mid h)) \right]. \tag{17}$$

Values are detached during updates and retain their original scale. The loss sums over all candidate actions using the four-rollout value estimates. The SFT reference remains frozen. Private scores and alternative outcomes never enter the decision prefix.

Supervised replay. Training items are sampled with probabilities $\rho_{\text{action}} = 0.50$, $\rho_{\text{QA}} = 0.25$, and $\rho_{\text{mem}} = 0.25$. An action item contributes the counterfactual loss for a decision history; QA and memory items contribute assistant-token-averaged SFT losses for answers/tool calls and incremental writes, respectively. Averaging item losses gives the mixture objective

$$\mathcal{L}_{\text{mix}} = \mathbb{E}_{u \sim \rho} \mathbb{E}_{z \sim \mathcal{D}_u} [\ell_u(z)]. \tag{18}$$

The sampling probabilities set item frequencies, and sampled item losses are averaged with equal weights. Rollout branches are evaluated within each action item.

Ablations and common metrics. Reward variants progressively add framing, joint aggregation, and historical gain to pixel visibility, retaining the same cumulative-return definition. Memory-component and reward variants are retrained under their respective configurations. GT memory replaces model-written records with the annotation-derived memory targets described in Appendix A. Action-budget comparisons use the full policy with limits of one, two, or three attempts. All reward variants are evaluated with the same full-history joint score:

$$\Delta_i = \Phi(h_{T_i}, q_i) - \Phi(h_{0_i}, q_i), \tag{19}$$

$$\text{EvidenceGain} = \frac{100}{N} \sum_i \Delta_i, \tag{20}$$

$$\text{EffectiveGainRate} = \frac{100}{N} \sum_i \mathbf{1}[\Delta_i > 0.10]. \tag{21}$$

The initial history includes all observations available before acting. Both metrics include all questions, including zero-action answers. Mean steps counts attempted camera actions, including invalid attempts.

D.3 Streaming Inputs and Main QA

All streaming models process observations chronologically at 1 FPS. Offline models receive up to 32 uniformly sampled frames from the available prefix, using all available frames when fewer are present. Future frames are withheld. On the Action subset, comparison models can also issue actions under the same action interface and attempt budget, receiving the corresponding observations before answering. Multiple-choice answers are scored by exact option agreement. For nonzero references $y \neq 0$, numeric answers use mean relative accuracy (Yang et al., 2024a) over thresholds $\mathcal{T} = \{0.05, 0.10, \ldots, 0.50\}$:

$$\mathrm{MRA}(\hat{y}, y) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathbf{1}\left[\frac{|\hat{y} - y|}{|y|} \leq \tau\right]. \tag{22}$$

For zero-valued references, we use exact zero agreement. We report results by QA family, and separate one-, two-, and three-action budgets when evaluating active reasoning. All compared models share the same question times, referents, and available observation prefixes, represented according to their streaming or offline input protocols.

Offline QA prompts. Offline QA prompts present chronologically ordered frames with timestamps, followed by the question and its labeled options or numerical unit. When submitting a final answer, models return exactly one JSON object, {`"type":"answer","answer":value`}. For multiple-choice questions, `value` must be a single uppercase option letter rather than the option text. For numerical questions, it must be a JSON number in the unit specified by the question; object counts must be integers. Linked questions include preceding questions and the model's own predictions as context.

Main QA aggregation and latency. The main evaluation contains 10,000 QA turns from each domain. Overall ACC pools the 13,134 multiple-choice turns, and Overall MRA pools the 6,866 numerical turns, weighting each task and domain by its applicable question count. Action contributes simulated questions only; no zero score is imputed for the real domain. Linked questions are scored per answer turn. Latency starts when the required images and context are ready and ends when the full answer has been produced. It excludes frame acquisition, ongoing observation maintenance, memory-tool operations, action execution or rendering, and network or queueing delays. The overall mean uses all 20,000 main QA turns, equivalent to averaging the two domain means with equal weights.

D.4 Timing Evaluation

Timing contains 1,000 simulated and 1,000 real-world questions presented before their evidence appears. Frames arrive chronologically at 1 FPS, retaining observed history and withholding future frames. At each checkpoint the model outputs `wait` or `answer`; the first answer is final. Memory and action tools are disabled.

The first answerable time t_i^* is derived from visibility and evidence references: the unique earliest option-category appearance, the queried category's first appearance, the earliest prefix leaving one consistent appearance-order option, or the arrival of the referenced historical frame for coappearance and direction questions. Tied first appearances are excluded, and derived answers must match ground truth. The model does not receive t_i^*; no example is answerable at the initial frame.

For first-answer time t_i^a, let $z_i = \mathbf{1}[t_i^* \leq t_i^a \leq t_i^* + 5\,\mathrm{s}]$, or zero if unanswered. The inclusive window uses video time, not inference latency. An answer at the deadline is allowed; waiting at that

checkpoint is unanswered. We report

$$T = \frac{100}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} z_i, \tag{23}$$

$$JA = \frac{100}{|\mathcal{Q}_c|} \sum_{i \in \mathcal{Q}_c} z_i \mathbf{1}[\hat{y}_i = y_i], \tag{24}$$

$$JM = \frac{100}{|\mathcal{Q}_n|} \sum_{i \in \mathcal{Q}_n} z_i m_i. \tag{25}$$

Here $m_i \in [0, 1]$ is numerical MRA; $|\mathcal{Q}| = 2{,}000$, $|\mathcal{Q}_c| = 1{,}787$, and $|\mathcal{Q}_n| = 213$. Denominators include all applicable questions, with early, late, and unanswered cases contributing zero. Timing is separate from Overall QA.

E Qualitative Examples

We provide visual examples of the evidence used by each QA family and of the incremental memory records that support historical retrieval.

Current observations and recent changes. Figure 5 illustrates approaching objects, newly visible categories, and camera displacement using observations five seconds apart. Figure 6 distinguishes trajectory length, camera rotation, and camera-to-object distance, while Figure 7 combines metric distance, changes in distance, and a linked category–distance query. Figure 8 shows closest-object identification, a follow-up relation query, and disappearance from view. Figure 9 illustrates tracking a selected referent across turns, camera-relative direction, and height comparison; Figure 10 shows nearest-neighbor and fixed-reference spatial relations. Together, these examples distinguish geometric reasoning in the current view from short-window temporal comparisons.

Historical evidence. Figure 11 illustrates retrieving the first observed category and its timestamp, and Figure 12 contrasts first-appearance lookup with counting distinct previously observed instances. Figure 13 uses an earlier observation to recover the nearest category or distance for a referent no longer visible. Figure 14 shows historical coappearance and image-plane direction; Figure 15 uses timestamps to recover the last observation and first-appearance order. Figure 16 provides further distinct-instance counting and historical nearest-neighbor examples, illustrating why the current view alone may be insufficient.

Acquiring additional views. Figure 17 shows turns that expose a target for size or proximity questions, and Figure 18 illustrates categorical and numerical inter-object distance queries after a turn. Figure 19 shows a turn and a lateral movement that provide additional views for inter-object distance questions. Figure 20 illustrates multi-step translations that increase target-pixel evidence relative to the best historical view before size estimation. These examples visualize the evidence acquired by actions; metric answers are derived from the scene annotations.

Incremental memory writes. Figure 21 shows five-second write windows in which objects leave view or new categories and counts appear. Figure 22 illustrates changing categories and a frame-linked spatial relation. Both figures retain individual frame entries within each write window: per-frame visible counts are not cumulative distinct-instance counts.

CURRENT

Questions grounded in current observations and short temporal context

approached object category



5 s | temporal evidence



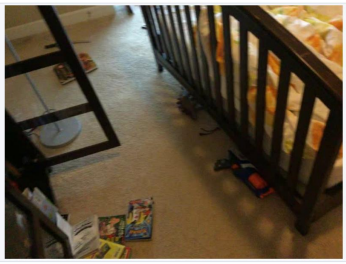
10 s | question / current

Q Among objects visible both now and 5 seconds ago, which category's object had the greatest camera-to-center distance decrease?

A. door B. cutting board C. power strip D. microwave E. pan

Answer: D - microwave

became visible category



50 s | temporal evidence



55 s | question / current

Q Comparing now with 5 seconds ago, which category became visible?

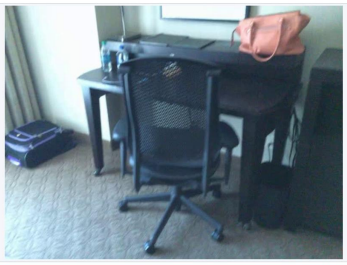
A. window B. armchair C. bed D. door E. cabinet

Answer: A - window

camera motion bucket



16 s | temporal evidence



21 s | question / current

Q What was the camera's straight-line endpoint displacement range over the last 5 seconds?

A. almost stationary (at most 0.2 meters) B. small (0.35 to 0.75 meters) C. moderate (1 to 1.5 meters) D. large (at least 2 meters)

Answer: B - small (0.35 to 0.75 meters)

Figure 5: Current QA sample 1

CURRENT

Questions grounded in current observations and short temporal context

camera path length m



47 s | temporal evidence



52 s | question / current

Q Estimate the total camera trajectory length over the last 5 seconds, in meters.

Answer: 1.266 meters

camera rotation deg



2 s | temporal evidence



7 s | question / current

Q Estimate the net endpoint camera orientation difference over the last 5 seconds, in degrees.

Answer: 48.624 degrees

camera to object distance bucket



51 s | question / current

Q What is the camera-to-center distance range for the closest visible trash can?

A. very near (at most 1.25 meters) B. near (1.5 to 2.5 meters) C. mid-range (3 to 4.5 meters) D. far (at least 5 meters)

Answer: B - near (1.5 to 2.5 meters)

Figure 6: Current QA sample 2

CURRENT

Questions grounded in current observations and short temporal context

camera to object distance m

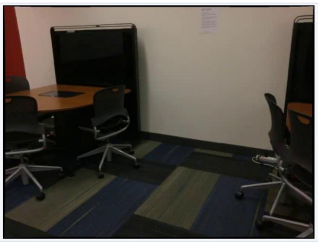


33 s | question / current

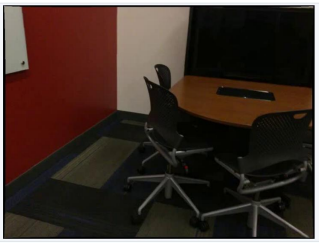
Q Estimate the camera-to-center distance to the closest visible bed, in meters.

Answer: 3.625 meters

closer than before



5 s | temporal evidence



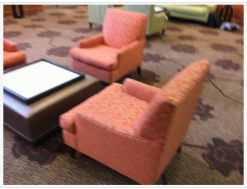
10 s | question / current

Q How did the camera-to-center distance to the table visible at both times change over the last 5 seconds?

A. much closer (decrease of at least 0.75 meters) B. slightly closer (decrease of 0.2 to 0.5 meters) C. slightly farther (increase of 0.2 to 0.5 meters) D. much farther (increase of at least 0.75 meters)

Answer: A - much closer (decrease of at least 0.75 meters)

closest category then distance



47 s | question / current

Q1 Which visible object category is closest to the camera by 3D center distance?

A. chair B. sofa C. table D. picture E. monitor

Answer: A - chair

Q2 What is that selected object's camera-to-center distance range now?

A. very near (at most 1.25 meters) B. near (1.5 to 2.5 meters) C. mid-range (3 to 4.5 meters) D. far (at least 5 meters)

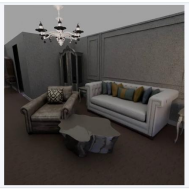
Answer: B - near (1.5 to 2.5 meters)

Figure 7: Current QA sample 3

CURRENT

Questions grounded in current observations and short temporal context

closest category then relation



32 s | question / current

Q1 Which visible object category is closest to the camera by 3D center distance?

A. lighting B. armchair C. tea table D. bed E. media unit

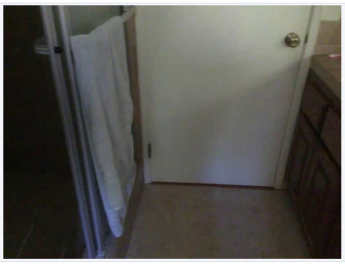
Answer: C - tea table

Q2 Which other currently visible object's category is nearest to that object by 3D center distance?

A. storage unit B. media unit C. armchair D. bed E. lighting

Answer: C - armchair

closest object category



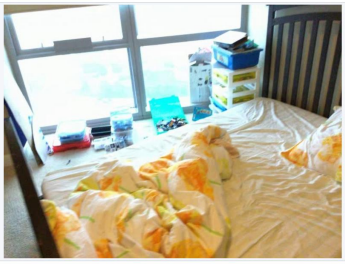
31 s | question / current

Q Which visible object category is closest to the camera by 3D center distance?

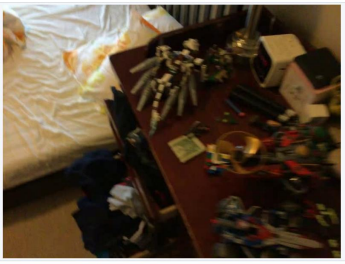
A. armchair B. door C. toilet D. art E. sink

Answer: B - door

disappeared category



53 s | temporal evidence



58 s | question / current

Q Comparing now with 5 seconds ago, which category disappeared from view?

A. door B. cabinet C. window D. armchair E. bed

Answer: C - window

Figure 8: Current QA sample 4

CURRENT

Questions grounded in current observations and short temporal context

object category then later visibility



2 s | question / current



7 s | question / current

Q1 Which visible object category is closest to the camera by 3D center distance?

A. art B. door C. chair D. basin E. barstool

Answer: C - chair

Q2 Is that previously selected object visible now?

A. yes B. no

Answer: B - no

object relative to camera



30 s | question / current

Q Where is the closest visible chair relative to the camera now?

A. far to the left B. slightly to the left C. in front D. slightly to the right E. far to the right

Answer: C - in front

multi candidate height



30 s | question / current

Q Among the choices, which one gives the listed category whose object visible at the time of this question has the greatest vertical height (not elevation above the floor)?

A. bed B. storage unit C. table D. media unit

Answer: B - storage unit

Figure 9: Current QA sample 5

CURRENT

Questions grounded in current observations and short temporal context

closest neighbor category



14 s | question / current

Q Choose the answer corresponding to the listed category whose initially visible object has its 3D center closest to the table.

A. chair B. art C. storage unit D. barstool

Answer: A - chair

fixed reference relation



12 s | question / current

Q Identify the listed category whose initially visible object is to the left of the tea table, using the image at the time this question was asked as the fixed reference view by selecting an option.

A. art B. tv stand C. armchair D. l-shaped sofa

Answer: D - l-shaped sofa

Figure 10: Current QA sample 6

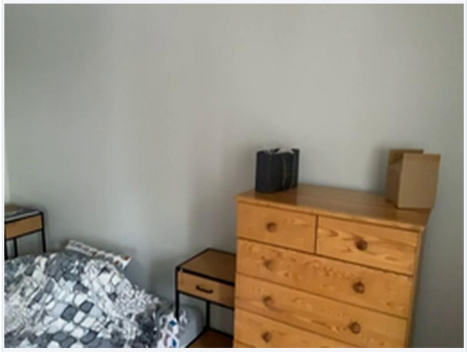
MEMORY

Questions grounded in recorded historical observations

first seen category then timestamp



5 s | historical evidence



52 s | question / current

Q1 Which of the listed object categories was seen first in this episode?

A. armchair B. barstool C. art D. basin E. cabinet

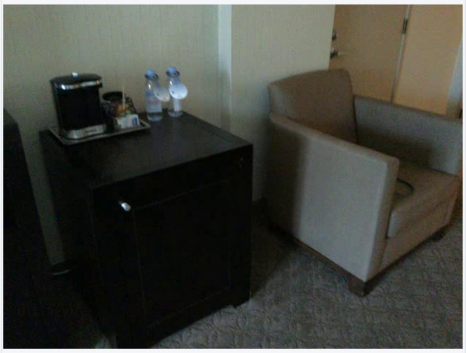
Answer: E - cabinet

Q2 At what timestamp was that category first seen in this episode, in seconds?

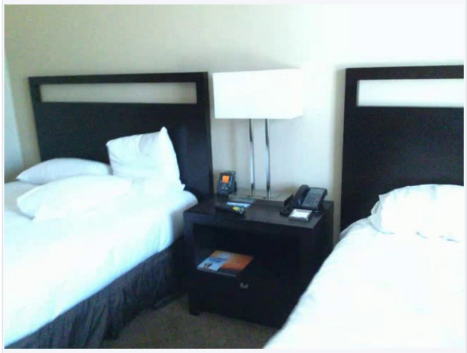
Answer: 5.0 seconds

Memory query: armchair, art, barstool, basin, cabinet. Timestamped memory records support the answer; historical images illustrate the stored evidence.

first seen object category



0 s | historical evidence



47 s | question / current

Q Which of the listed object categories was seen first in this episode?

A. bed B. chair C. curtain D. picture E. monitor

Answer: B - chair

Memory query: bed, chair, curtain, monitor, picture. Timestamped memory records support the answer; historical images illustrate the stored evidence.

Figure 11: Memory QA sample 1

MEMORY

Questions grounded in recorded historical observations

first seen timestamp s



0 s | historical evidence

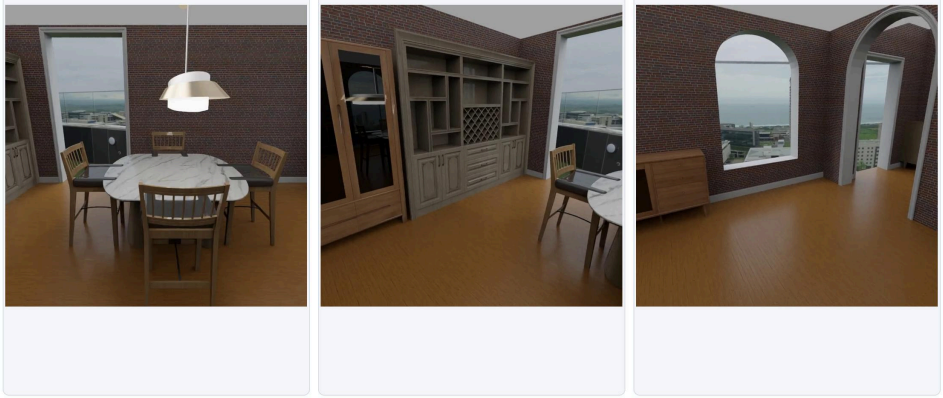
52 s | question / current

Q At what timestamp did the plant category first appear, in seconds?

Answer: 0.0 seconds

Memory query: plant. Timestamped memory records support the answer; historical images illustrate the stored evidence.

seen count by category



0 s | historical evidence

1 s | historical evidence

5 s | question / current

Q How many different dining chair objects have been seen so far?

Answer: 4 count

Memory query: dining chair. Count distinct objects across retrieved history; count repeated sightings of the same object only once.

Figure 12: Memory QA sample 2

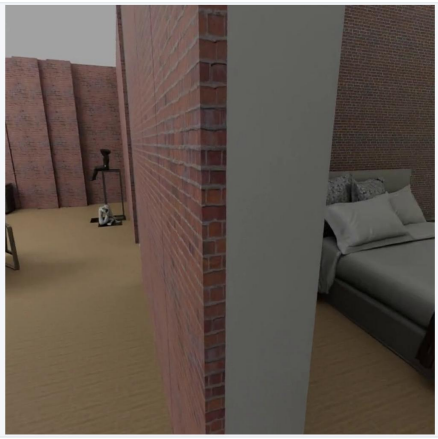
MEMORY

Questions grounded in recorded historical observations

single seen object then closest category



11 s | historical evidence



47 s | question / current

Q Among the other objects visible at 11 seconds with the tea table seen earlier and absent now, which category was nearest to it by 3D center distance?

A. bed B. dining chair C. dining table D. two-seat sofa E. cabinet

Answer: D - two-seat sofa

Memory query: tea table. Memory lookup returns the historical evidence frame; the answer refers to that past observation.

single seen object then closest distance



23 s | historical evidence



56 s | question / current

Q Among the other objects visible at 23 seconds with the monitor seen earlier and absent now, estimate its 3D center distance to the nearest one, in meters.

Answer: 1.179 meters

Memory query: monitor. Memory lookup returns the historical evidence frame; the answer refers to that past observation.

Figure 13: Memory QA sample 3

MEMORY

Questions grounded in recorded historical observations

historical coappearance category



47 s | historical evidence



55 s | question / current

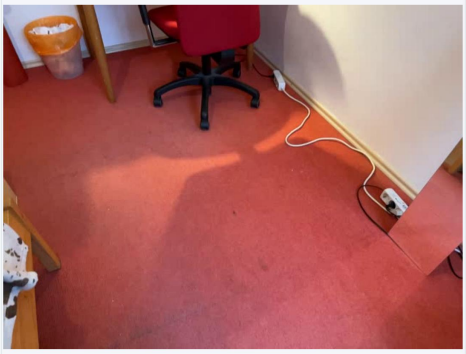
Q Among the choices, which one gives the listed category also visible alongside the chair at timestamp 47 seconds?

A. armchair B. table C. barstool D. art

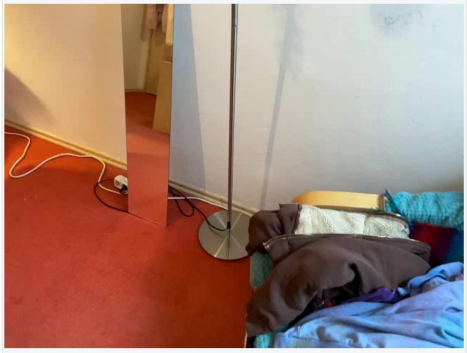
Answer: B - table

Historical records / evidence are retained before the question. No memory write is performed during QA.

historical relative direction



34 s | historical evidence



40 s | question / current

Q What is the image-plane position of the trash can relative to the chair in the observation at timestamp 34 seconds? Choose from the listed options.

A. above B. below C. right D. left

Answer: D - left

Historical records / evidence are retained before the question. No memory write is performed during QA.

Figure 14: Memory QA sample 4

MEMORY

Questions grounded in recorded historical observations

last seen timestamp



39 s | historical evidence



40 s | historical evidence



41 s | historical evidence



45 s | question / current

Q How much is the last timestamp at which the bed category was observed before this question? Answer in seconds.

Answer: 40 seconds

Timestamped memory records support the answer; historical images illustrate the stored evidence.

first appearance order



0 s | first table



3 s | first cabinet



4 s | additional cabinet view



5 s | first sofa

Q Among the choices, which one gives the first-observation order of the categories cabinet, sofa, table since this episode began?

A. table > sofa > cabinet B. table > cabinet > sofa C. cabinet > table > sofa D. cabinet > sofa > table

Answer: B - table > cabinet > sofa

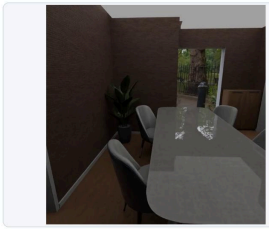
Question at 50 s. First observations: table 0 s, cabinet 3 s, sofa 5 s. The 4 s view further illustrates the cabinet; timestamped memory records establish the order.

Figure 15: Memory QA sample 5

MEMORY

Questions grounded in recorded historical observations

seen count by category



5 s | historical evidence



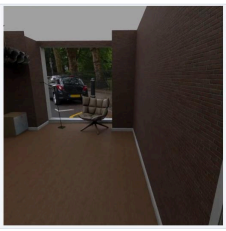
7 s | historical evidence



8 s | historical evidence



9 s | historical evidence



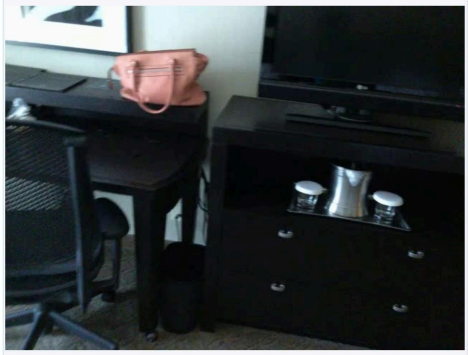
12 s | question / current

Q How many different chair objects have been seen so far?

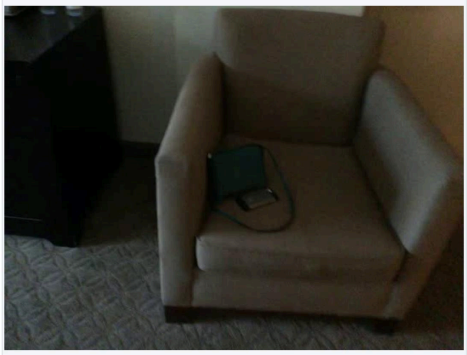
Answer: 5 count

Memory query: chair. Count distinct objects across retrieved history; count repeated sightings of the same object only once.

single seen object then closest category



13 s | historical evidence



67 s | question / current

Q Among the other objects visible at 13 seconds with the monitor seen earlier and absent now, which category was nearest to it by 3D center distance?

A. chair B. curtain C. door D. table E. bed

Answer: D - table

Memory query: monitor. Memory lookup returns the historical evidence frame; the answer refers to that past observation.

Figure 16: Memory QA sample 6

ACTION

Questions answered after acquiring additional visual evidence

object size m



16 s | question / current



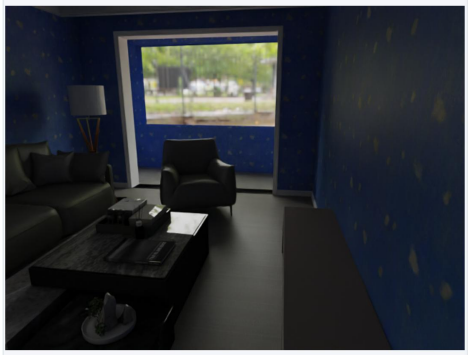
Step 1 | new observation | rotate left 45 deg

Q What is the longest physical side of the 3D bounding box of the bed visible in the initial view, in meters?

Answer: 2.685 meters

Initial question > action > new observation > answer. Metric quantities use scene annotations.

object spatial proximity



24 s | question / current



Step 1 | new observation | rotate left 45 deg

Q Which 3D center-to-center distance range describes the sofa in the middle left of the initial view and the table visible in the initial view?

A. very near (at most 0.75 meters) B. near (1 to 1.75 meters) C. mid-range (2 to 3 meters) D. far (at least 3.5 meters)

Answer: B - near (1 to 1.75 meters)

Initial question > action > new observation > answer. Metric quantities use scene annotations.

Figure 17: Action QA sample 1

ACTION

Questions answered after acquiring additional visual evidence

object to object distance bucket



5 s | question / current



Step 1 | new observation | rotate left 45 deg

Q Which 3D center-to-center distance range describes the l-shaped sofa visible in the initial view and the tea table visible in the initial view?

A. very near (at most 0.75 meters) B. near (1 to 1.75 meters) C. mid-range (2 to 3 meters) D. far (at least 3.5 meters)

Answer: B - near (1 to 1.75 meters)

Initial question > action > new observation > answer. Metric quantities use scene annotations.

object to object distance m



27 s | question / current



Step 1 | new observation | rotate right 45 deg

Q What is the 3D center-to-center distance between the l-shaped sofa visible in the initial view and the tea table visible in the initial view, in meters?

Answer: 1.058 meters

Initial question > action > new observation > answer. Metric quantities use scene annotations.

Figure 18: Action QA sample 2

ACTION

Questions answered after acquiring additional visual evidence

object to object distance m





5 s | question / current

Step 1 | new observation | rotate left 45 deg

Q What is the 3D center-to-center distance between the l-shaped sofa visible in the initial view and the tea table visible in the initial view, in meters?

Answer: 1.037 meters

Initial question > action > new observation > answer. Metric quantities use scene annotations.

object to object distance m





59 s | question / current

Step 1 | new observation | move left 0.5 m

Q What is the 3D center-to-center distance between the three-seat / multi-person sofa visible in the initial view and the single bed visible in the initial view, in meters?

Answer: 2.38 meters

Initial question > action > new observation > answer. Metric quantities use scene annotations.

Figure 19: Action QA sample 3

ACTION

Questions answered after acquiring additional visual evidence

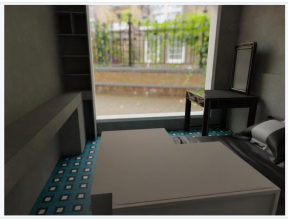
object size m



9s | question / before



Step 1 | forward | 0.5 m



Step 2 | forward | 0.5 m



Step 3 | forward | 0.5 m



Best historical target-pixel observation

Q What is the longest physical side of the 3D bounding box of the dressing table visible in the initial view, in meters?

Answer: 1.672 meters

Target pixel share (audit) vs all prior history: 0.76% > 3.39% > 4.83% > 7.23%. Each retained step improves target pixel evidence over history and prior chosen actions.

object size m



59s | question / before



Step 1 | left | 0.5 m



Step 2 | forward | 0.5 m



Best historical target-pixel observation

Q What is the longest physical side of the 3D bounding box of the three-seat / multi-person sofa visible in the initial view, in meters?

Answer: 1.998 meters

Target pixel share (audit) vs all prior history: 6.66% > 14.81% > 21.09%. Each retained step improves target pixel evidence over history and prior chosen actions.

Figure 20: Action QA sample 4

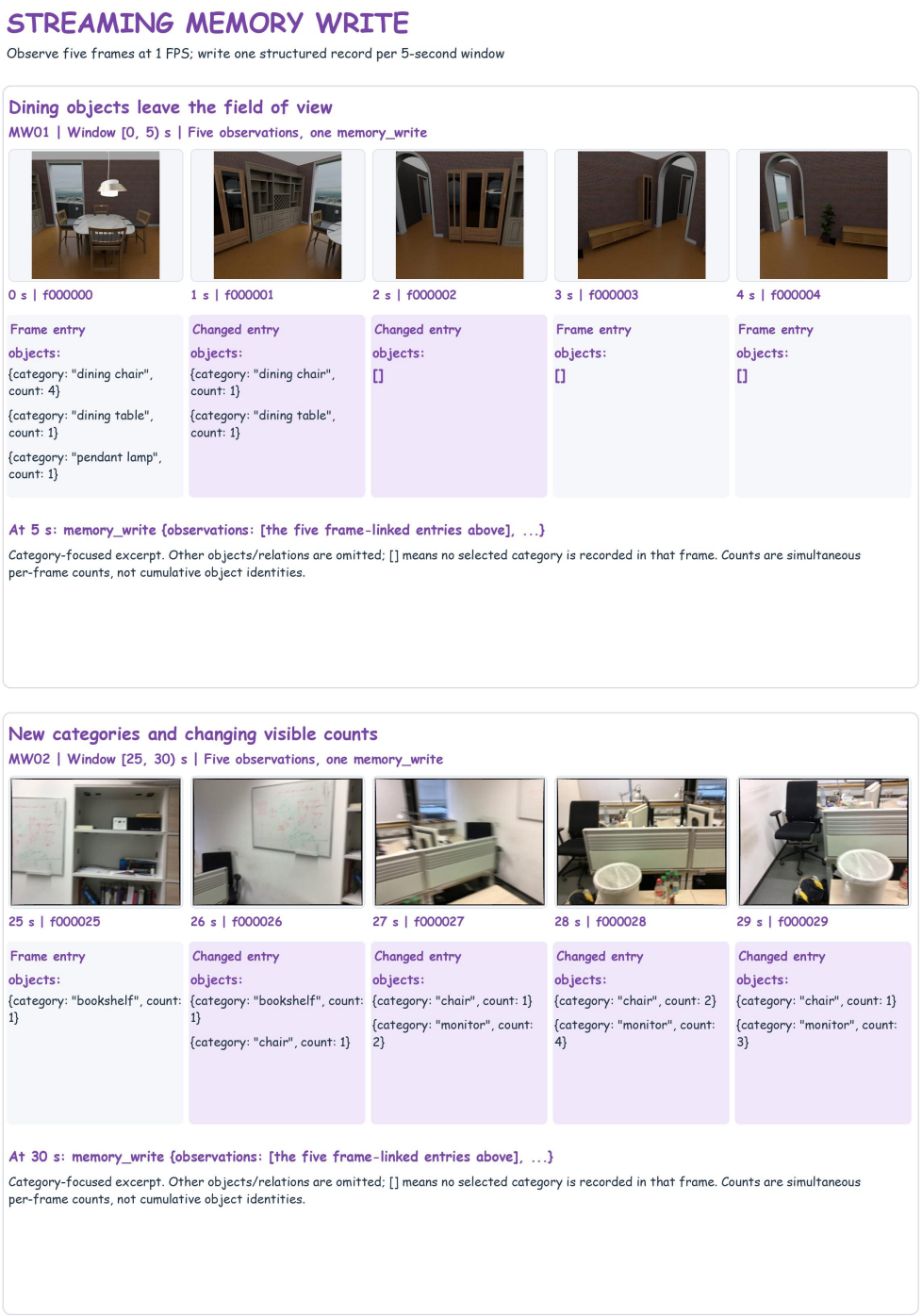


Figure 21: Memory write sample 1

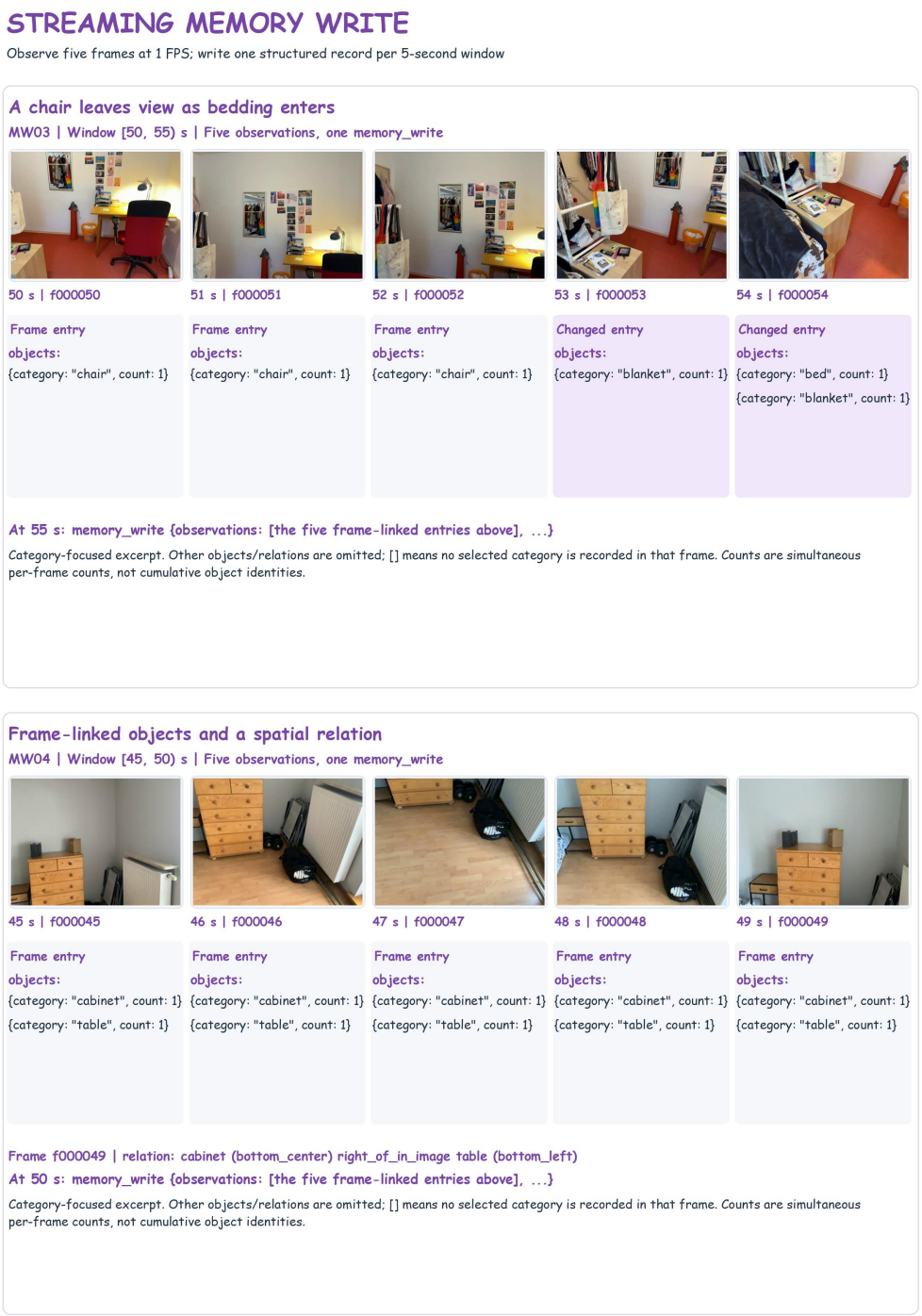


Figure 22: Memory write sample 2