HAVE I SCENE THIS BEFORE? SPATIALLY GROUNDED CONVERSATIONAL MEMORY FOR COMPLEX QUERIES IN EGOCENTRIC ASSISTANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personal assistants are evolving from search and text interfaces toward interactive multimodal systems that retain spatial and conversational history to support ongoing tasks. Effective assistants must therefore reason jointly over observed scenes and past dialogue to address complex requests. We formalize this challenge as *Spatially grounded Conversational Reasoning* (SPACR), covering cross-scene, recall-oriented, and counterfactual spatial queries that require accurate geometric reasoning and the integration of dialogue with spatial evidence. Existing approaches are poorly suited to this setting. Direct VLMs incur high inference costs and context limits as multimodal histories grow. Keyframe selection and retrieval can omit evidence needed for complete recall, while memory systems must preserve both conversational facts and object geometry to support spatial reasoning. To bridge this gap, we propose Spatially grounded Conversational Memory (SPAC-MEM), an object-centric working memory that applies 3D reconstruction and segmentation to egocentric observations, thereby spatially grounding conversational context directly to physical objects. It compresses long multimodal histories while preserving spatial and conversational evidence and allowing object-specific information to be updated through later dialogue. We further introduce EGO-SPACR, an egocentric SPACR dataset comprising 620 video sessions from ScanNet, augmented with 95 conversations averaging 230 dialogue turns each, and 3,100 evaluation queries. Experiments show that SPAC-MEM achieves the highest overall answer accuracy among the evaluated methods and improves object recall, while requiring substantially fewer reference input tokens than native-video baselines. Analysis shows that removing 3D spatial information substantially degrades performance, highlighting the importance of preserving both spatial and conversational evidence for an egocentric assistant.

1 INTRODUCTION

As personal assistants move beyond search and text interfaces toward egocentric multimodal systems, they must remain coherent with both what the user sees and what they say over time (Wang et al., 2023). Users naturally refer to objects in view while expressing facts such as ownership and preferences that cannot be inferred visually. As interactions span rooms, sessions, and days, tasks requiring object comparisons, alternative arrangements, or a complete inventory naturally arise, even when the relevant evidence is no longer visible or readily recoverable. Thus, an effective assistant should preserve both historical spatial and conversational context to support ongoing tasks, without requiring the user to remember or revisit earlier interactions. Figure 1 demonstrates how a user might identify friends' belongings while moving through an apartment, then later ask a spatially grounded question dependent on those same references, such as *"How much of my friend's stuff on the tables earlier can I fit in this box?"* Answering such complex cross-scene spatial queries requires more than retrieving a relevant frame: the assistant must recover conversational facts tied to persistent objects, combine them with prior geometric, qualitative relational, or counterfactual spatial evidence, and account for all relevant objects across the history. We refer to this as *Spatially grounded Conversational Reasoning (*SPACR*)*.

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.

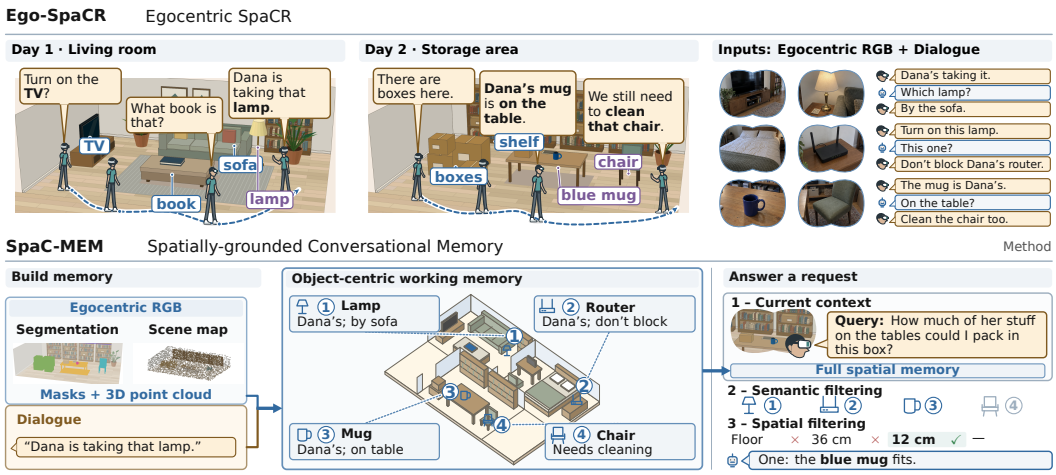

Figure 1: SPACR involves complex spatial queries that require connecting user-stated facts to spatial evidence about the same physical objects across the interaction history. Here, identifying which of Dana's belongings could fit in a box requires recalling ownership established through conversation and checking previously observed locations and object dimensions. SPAC-MEM links this conversational and spatial evidence through persistent objects in a unified working memory.

Where Existing Approaches Fall Short. Direct vision-language models (VLMs) repeatedly process growing multimodal histories, exhausting context windows and incurring high inference costs (Tang et al., 2025; Zhu et al., 2026). Keyframe selection and multimodal Retrieval-Augmented Generation (RAG) reduce context but may miss objects or evidence required for questions requiring complete recall. Spatial scene memories primarily retain visual observations without evolving conversational context (Gu et al., 2024; Yang et al., 2025b). Multimodal conversational memories such as M2A retain raw interactions alongside semantic records (Feng et al., 2026), but do not maintain the explicit object geometry needed for complex spatial reasoning. Existing datasets likewise lack key requirements of SPACR (Table 3). VideoQA approaches (Yang et al., 2025a; 2026; Gu et al., 2026) do not evaluate spatial reasoning over conversationally established facts, while situated dialogue datasets (Wu et al., 2023) focus on single-scene interactions with limited spatial reasoning.

A Spatially Grounded Working Memory. To address this gap, we propose Spatially grounded Conversational Memory (SPAC-MEM), a unified object-centric *working memory* that grounds user-specific conversational information to physical objects in historical egocentric interactions while preserving their spatial evidence (e.g., object size) via 3D geometric reconstruction with segmentation. SPAC-MEM compresses long frame-based histories into a compact representation, while allowing conversational information to be accumulated and updated over time (e.g., changing a chair from discarded to kept). We further introduce EGO-SPACR, a benchmark for SPACR over egocentric video, augmented with task-oriented dialogue. Our experiments show that SPAC-MEM achieves the highest overall answer accuracy among the evaluated methods and improves object recall while requiring substantially fewer reference input tokens than native-video baselines. Removing explicit 3D spatial information substantially degrades performance, supporting the use of spatial evidence alongside conversation in memory. Our contributions are summarized as follows:

- We introduce Spatially grounded Conversational Memory (SPAC-MEM), an object-centric working memory that connects detailed spatial and conversational information via persistent physical objects across historical egocentric user interactions.
- We introduce EGO-SPACR, a benchmark for SPACR over egocentric video and task-oriented dialogue, with 95 task-driven conversations in 620 real-world video sessions and 3,100 evaluation questions across varying classes of spatial complexity and recall breadth.
- We conduct a controlled analysis across EGO-SPACR, revealing where existing methods fail as queries require increasingly complex spatial reasoning and recall over object sets.
- We show that SPAC-MEM consistently outperforms existing methods in accuracy and object recall, while substantially reducing input token cost compared to in-context methods.

2 Related Work

2.1 Egocentric Video and Dialogue Benchmarks

Egocentric video captures the environment from the wearer's viewpoint, recording what they see and interact with (Damen et al., 2018; Grauman et al., 2022). EPIC-KITCHENS (Damen et al., 2018) and Ego4D (Grauman et al., 2022) capture longer-term first-person experiences across different environments. ScanNet further provides RGB-D observations of real scenes together with geometry and object annotations (Dai et al., 2017) as the foundation of spatial understanding evaluation.

Video question answering (VideoQA). Egocentric VideoQA benchmarks, such as EgoVQA and EgoSchema, cover first-person activities over longer video contexts (Fan, 2019; Mangalam et al., 2023). OpenEQA and SQA3D evaluate questions in situated 3D environments (Majumdar et al., 2024; Ma et al., 2023); VSI-Bench, VSI-SUPER, and VSI-Super-Wild focus on spatial reasoning, long-term spatial recall, and counting (Yang et al., 2025a; 2026; Gu et al., 2026). These VideoQA benchmarks evaluate questions about recorded observations without an evolving history of user-stated facts tied to persistent physical objects. In SpaCR, language is integral to the evolving history, supplying user-specific facts by referring to an object in view, such as ownership or preferences. The assistant must connect those facts to the same object's spatial evidence over time.

Long-term Dialogue and Situated Conversation. Long-term dialogue benchmarks such as LongMemEval (Wu et al., 2025) evaluate assistants' ability to use evolving conversation but do not ground it in spatial environments. Situated dialogue (SIMMC-VR (Wu et al., 2023) and SIMMC 2.0 (Kottur et al., 2021)) connect conversation with spatial properties in simulated environments, but focus on short, single-scene interactions with limited spatial complexity. They do not require reasoning across scene spaces to answer the complex counterfactual spatial queries that users may request.

2.2 Spatial Understanding with Video Input

In-Context Spatial Understanding. Recent multimodal models directly process video and language; Gemini supports native video understanding, while Qwen3-VL extends long-context vision-language modeling to long videos and temporal grounding (Google DeepMind, 2026; Bai et al., 2025). These models keep visual and conversational history in context instead of organizing it around persistent physical objects. As a result, recovering every object linked to conversational facts remains difficult. Query-aware frame selection methods such as AKS and FOCUS select informative frames before VLM inference (Tang et al., 2025; Zhu et al., 2026). But selection is conditioned primarily on queries and visuals, without considering the evolving conversation.

Agentic Spatial Understanding with Object Memory. Object memory methods reconstruct persistent scenes from visual observations. ConceptGraphs builds open-vocabulary 3D scene graphs from RGB-D input (Gu et al., 2024). However, they primarily target visual understanding. Recent agentic systems (e.g., VideoAgent) combine temporal captions, object tracking, re-identification, and tool-based memory in hybrid agentic retrieval (Fan et al., 2024); Gemini's agentic video understanding searches and inspects relevant segments. While these systems combine language-guided reasoning with video memory, they still keep conversation and visual memory explicitly parallel.

Multimodal RAG and Dialogue Memory. Retrieval-based methods select stored evidence for answering: RAG retrieves text (Lewis et al., 2020), VideoRAG retrieves video evidence (Jeong et al., 2025). Our *Multimodal RAG* baseline retrieves dialogue and frames using embeddings, while *Summary RAG* retrieves textual summaries of both. M2A retains raw interactions alongside semantic memory and uses agentic retrieval to manage long-term context (Feng et al., 2026). Retrieval can miss evidence needed for complete recall, while summaries can omit geometry required for SpaCR.

3 SpaCR: Spatially Grounded Conversational Reasoning

We distinguish SpaCR from existing settings through four required properties: (i) *bimodal histories*, since users naturally introduce non-visual information while referring to objects they see; (ii) *spatial grounding*, so that this information remains attached to the same physical objects across views, sessions, and revisits; (iii) *complex spatial reasoning*, because later requests may depend on geometric

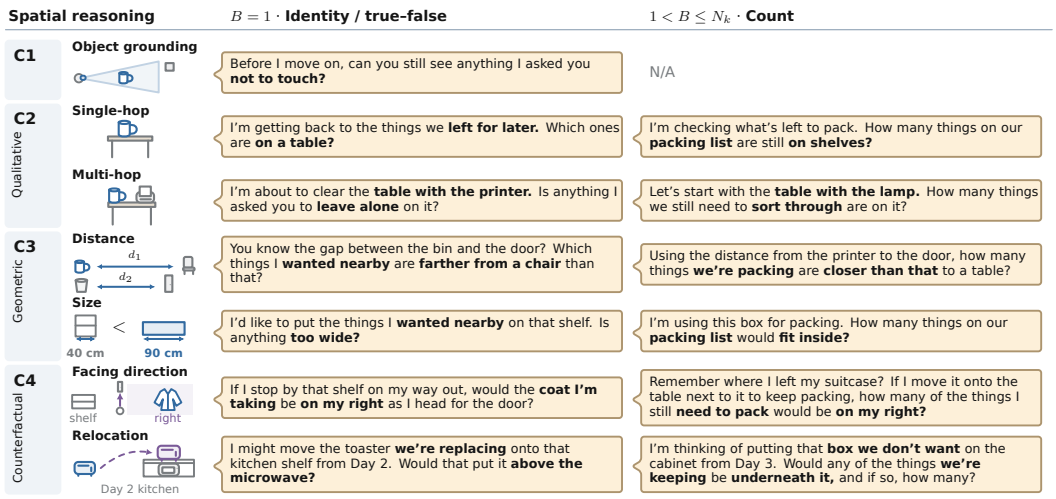

Figure 2: Example SPACR queries by recall type (columns) and reasoning level (rows). Unlike VideoQA, each query requires user-stated facts from conversation alongside spatial evidence.

relations or hypothetical arrangements that are not directly visible; and (iv) *recall breadth*, as tasks may require accounting for the complete relevant object set rather than retrieving isolated evidence.

(i) Bimodal histories. An egocentric assistant receives two histories: a sequence of egocentric RGB observations $\mathcal{V} = (v_1, \ldots, v_T)$ and a concurrent dialogue history $\mathcal{C} = (c_1, \ldots, c_K)$ spanning multiple scenes. Each frame v_t contains a set of visible physical objects $\mathcal{O}_t$, and each turn c_k occurs at time t_k with v_{t_k}. The objects the assistant has observed by turn k are therefore $\mathcal{O}_{\leq t_k} = \bigcup_{t \leq t_k} \mathcal{O}_t$. Object identities persist within a scene; objects from different scenes remain distinct. We represent each user-stated semantic fact as $\phi = (o, p, x)$, where object o has property p with value x, such as $(o, \text{owner}, \text{Dana})$, capturing information unavailable from $\mathcal{V}$.

As users change decisions over time, an existing (o, p) may change in two ways: a *correction* retracts the old value and applies a new one retroactively, while a *state change* preserves it until the update. If a lamp changes from *keep* to *donate*, a later query about its initial status (e.g., *"What did I first decide about the lamp?"*) resolves to *donate* under a correction but *keep* under a state change.

(ii) Spatially grounded queries. Users may ask for help at any point during an ongoing task, based on what they previously said and encountered. An egocentric assistant must therefore answer from the available interaction history. We define a user request at turn k as a query q_k that can use only $\mathcal{V}_{\leq t_k}$ and $\mathcal{C}_{<k}$. Let $\Phi_{<k}$ contain the facts established before q_k, with corrections and state changes resolved for the time the query refers to. Each query defines a conversational condition $h_k(o)$ and a spatial condition $r_k(o)$, giving the matching set $\mathcal{S}_k = \{o \in \mathcal{O}_{\leq t_k} : h_k(o) \wedge r_k(o)\}$. Depending on the user's need, the assistant may identify a specific object from $\mathcal{S}_k$ or account for the full set through $|\mathcal{S}_k|$, as in tasks such as packing or organizing. This yields a two-dimensional space defined by the *spatial complexity* required by $r_k(o)$ and the *recall breadth* needed to recover $\mathcal{S}_k$ (Fig. 2).

(iii) Spatial complexity. Users may ask whether a prior object is currently visible, whether objects fit geometrically, are closer to one another, or satisfy hypothetical spatial arrangements. Supporting such requests requires increasingly complex spatial evidence. We define four levels through $r_k(o)$:

(C1) *Object grounding* tests whether the assistant can connect a conversational condition $h_k(o)$ established earlier to the same physical object in the current view, with $r_k(o) = \mathbf{1}[o \in \mathcal{O}_{t_k}]$. Users may refer back to previously discussed objects as they reappear during an ongoing task.

(C2) *Qualitative relations* tests whether the assistant can evaluate single- or multi-hop spatial relations R, such as *on* or *near*, over conversationally grounded objects. For objects on the same table as a conversationally qualified lamp, $r_k(o) = \exists u, w \in \mathcal{O}_{\leq t_k} : (u : \text{table}) \wedge (w : \text{lamp}) \wedge g_k(w) \wedge R_{\text{on}}(o, u) \wedge R_{\text{on}}(w, u)$. Here g_k qualifies the anchor independently of the target condition h_k: for *"a lamp that my roommate owns,"* $g_k(w) = \mathbf{1}[(w, \text{owner}, \text{Roommate}) \in \Phi_{<k}]$; without an anchor qualification, $g_k(w) = 1$.

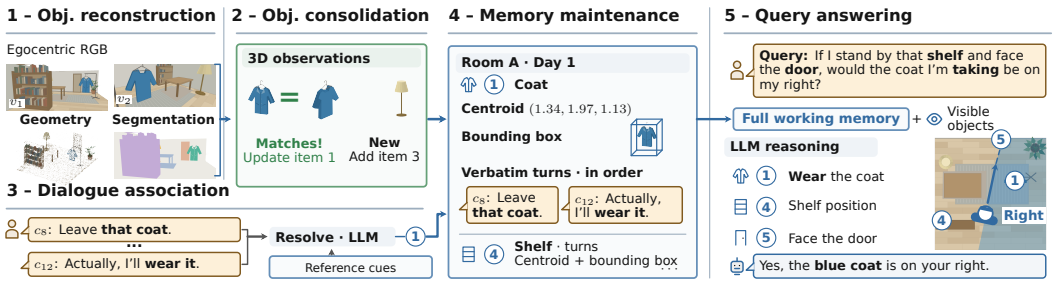

Figure 3: SPAC-MEM consolidates egocentric observations into persistent 3D object records, associates dialogue turns, and supplies the resulting working memory with visible objects to an LLM. The example combines updated intent and geometry for hypothetical viewpoint reasoning (C4).

(C3) *Geometric properties* tests whether the assistant can accurately recover the geometric properties of objects and relations, such as size or distance. Users may ask whether objects fit or compare their distances for tasks like packing, where qualitative relations alone are insufficient.

(C4) *Counterfactual spatial reasoning* tests whether the assistant can evaluate spatial relations under a hypothetical viewpoint or placement. Answering may require combining geometric evidence from earlier observations to reason about an arrangement that was not directly observed.

(iv) Recall breadth. Let $\mathcal{R}_k$ denote the target objects supporting the answer, excluding anchors used only to evaluate spatial conditions. For a count query, $\mathcal{R}_k = \mathcal{S}_k$; an identification or positive existence query needs a single target or witness. We define *recall breadth* as $B_k = |\mathcal{R}_k|$. Counting requires recovering the complete matching set. For zero-count and negative existence answers, $\mathcal{R}_k = \varnothing$, but all candidates must be ruled out; recall breadth therefore measures target-set size.

4 OBJECT-CENTRIC SPATIALLY GROUNDED CONVERSATIONAL MEMORY

As interaction histories grow, assistants that keep the full bimodal history in context face increasing difficulty and cost in SPACR. We therefore propose Spatially grounded Conversational Memory (SPAC-MEM), an object-centric conversational *working memory* that spatially grounds past dialogue in persistent physical objects. Recording each object once keeps the memory far smaller than the raw history while preserving the spatial and conversational detail SPACR needs.

SPAC-MEM consists of five stages, illustrated in Fig. 3: (1) *object reconstruction*, where incoming frames $\mathcal{V}$ are converted into frame-level 3D object observations; (2) *object consolidation*, where repeated observations update existing objects in a persistent pool and newly observed objects are added; (3) *dialogue association*, where incoming turns from $\mathcal{C}$ are grounded to the corresponding objects and attached to their records; (4) *memory maintenance*, which preserves the latest conversational state while retaining the relevant interaction history for each object; and (5) *query answering*, where a downstream LLM reasons directly over SPAC-MEM to answer SPACR queries. For full implementation details, see Appendix E.

(1) Object reconstruction. SPAC-MEM first extracts frame-level object observations $\hat{o}_{t,i}$ and their geometry from the RGB frames available at time t, $\mathcal{V}_{\leq t}$, without requiring depth measurements or supplied camera poses. A *feed-forward geometry reconstruction model*, Depth Anything 3 (Lin et al., 2025), serves as f_{geo} and estimates the depth map $\hat{D}_t$, the point map giving each pixel's 3D position $\hat{P}_t$, and the camera pose in the environment. SPAC-MEM transforms these components into a shared coordinate system, with axes aligned to dominant scene planes.

A segmentation model f_{seg} takes the current frame v_t and a predefined object vocabulary $\mathcal{L}$, and returns n_t object instances $\{(M_{t,i}, \ell_{t,i})\}_{i=1}^{n_t}$, where $M_{t,i}$ is the mask of instance i and $\ell_{t,i} \in \mathcal{L}$ is its class label. SPAC-MEM lifts each mask into 3D through the reconstructed point map, $\hat{P}_{t,i} = \{\hat{P}_t(u) : u \in \tilde{M}_{t,i}\}$, where $\tilde{M}_{t,i} \subseteq M_{t,i}$ retains only pixels whose depth is consistent with the object, reducing background contamination near mask boundaries. Each nonempty $\hat{P}_{t,i}$ defines an object observation $\hat{o}_{t,i}$ with spatial record $s_t(\hat{o}_{t,i}) = (\ell_{t,i}, \boldsymbol{\mu}_{t,i}, \mathbf{B}_{t,i})$, containing its class label, 3D

centroid $\boldsymbol{\mu}_{t,i}$, and an outlier-robust axis-aligned bounding box $\mathbf{B}_{t,i}$. A key design choice is whether SPAC-MEM should preserve this detailed spatial evidence or retain only object class and persistent identity. We evaluate this choice in RQ3.

(2) Object consolidation. SPAC-MEM consolidates the frame-level observations into a persistent object pool $\hat{\mathcal{O}}_{\leq t}$ for each scene. For each observation $\hat{o}_{t,i}$, we consider previously reconstructed objects in $\hat{\mathcal{O}}_{\leq t-1}$ with the same class label $\ell_{t,i}$ and select $a_t(\hat{o}_{t,i}) = \arg\max_{\hat{o}} A(\hat{o}_{t,i}, \hat{o})$, where A measures overlap between their reconstructed 3D regions. The association is accepted only when the best match is unambiguous, reducing erroneous merges between nearby objects. If $\hat{o}_{t,i}$ matches no existing object, it initializes a new persistent object. For a matched object, its accumulated spatial record $s_{\leq t}(\hat{o})$ is updated with the new observation. The pool is periodically refined by merging overlapping fragments that never co-occur in the same frame and are supported by appearance or tracking evidence, while rarely observed objects are removed. A key design choice is *object-centric* vs. *per-frame* structure that receives the same reconstructed geometry before cross-frame association. Object-centric structures prevent the downstream context from growing with every observed frame while preserving detailed spatial information that exceeds the context window.

(3) Dialogue association. After object consolidation establishes persistent object identities, this step grounds each dialogue turn to the physical objects it refers to, allowing conversational information to accumulate on the same object records over time. For a turn c_k at visual time t_k, an association model maps $\hat{\mathcal{A}}_k = f_{\text{assoc}}(c_k, \hat{\mathcal{O}}_{t_k}, \hat{\mathcal{O}}_{\leq t_k}, \mathcal{C}_{<k})$, where $\hat{\mathcal{O}}_{t_k}$ is the subset of persistent objects observed at frame t_k, and $\hat{\mathcal{A}}_k \subseteq \hat{\mathcal{O}}_{\leq t_k}$ contains the objects about which the turn makes a statement. The model resolves references using prior dialogue, object classes, spatial descriptions, pronouns, and previously established names.

Each $\hat{o} \in \hat{\mathcal{A}}_k$ is then associated with the original turn c_k; we retain the original dialogue turn rather than replace it with a generated summary (Hu et al., 2026), preserving the user's wording while allowing later turns to revise information attached to the same persistent object. In RQ3, we evaluate the reliability of LLM-predicted object associations with human-annotated references.

(4) Memory maintenance. Together, object reconstruction, consolidation, and dialogue association produce a persistent working memory that connects each physical object with its accumulated spatial and conversational information. Before query q_k, SPAC-MEM represents this memory as one record per object $\hat{o} \in \hat{\mathcal{O}}_{\leq t_k}$: $\mathcal{W}_k(\hat{o}) = \left(s_{\leq t_k}(\hat{o}), [c_j]_{j<k:\, \hat{o} \in \hat{\mathcal{A}}_j} \right)$, where $s_{\leq t_k}(\hat{o})$ is the consolidated spatial record and the list contains all preceding dialogue turns associated with that object. Records are organized by scene and session while preserving conversational order. As new observations and turns arrive, the corresponding record is updated rather than adding separate frame-level entries.

(5) Query answering. For a query q_k, the answering agent π_{LLM} receives the current memory, the objects visible in the frame $\hat{\mathcal{O}}_{t_k}$, and the query q_k, producing $\hat{a}_k = \pi_{\text{LLM}}(q_k, \hat{\mathcal{O}}_{t_k}, \mathcal{W}_k)$. The complete working memory is provided without retrieval, allowing the answerer to examine all stored objects when a query requires a complete set rather than a relevant subset.

5 EXPERIMENTS

To understand the performance and limitations of existing methods in SPACR, we study three RQs:

- **RQ1** Does SPAC-MEM improve downstream answering model performance on SPACR compared with existing approaches?
- **RQ2** Across object identification, geometric estimation, and spatial reconstruction, which abilities limit current methods on SPACR?
- **RQ3** How do SPAC-MEM's design choices, including object-centric structure, dialogue association, and geometric reconstruction, contribute to overall performance?

We compare against three baseline families spanning direct in-context reasoning, agentic video memory, and retrieval-based conversational memory: (i) *In-context spatial understanding.* Native-video baselines use GEMINI 3.7 FLASH (Google, 2026) and QWEN3.8-27B-FP8 (Qwen Team, 2026), each given the video prefix and timestamped conversation before the query at 1 fps. Frame-based baselines use uniform sampling, AKS, and FOCUS (Tang et al., 2025; Zhu et al., 2026), with selected frames capped by a fixed visual-token budget and provided to GPT-5.6 TERRA (OpenAI,

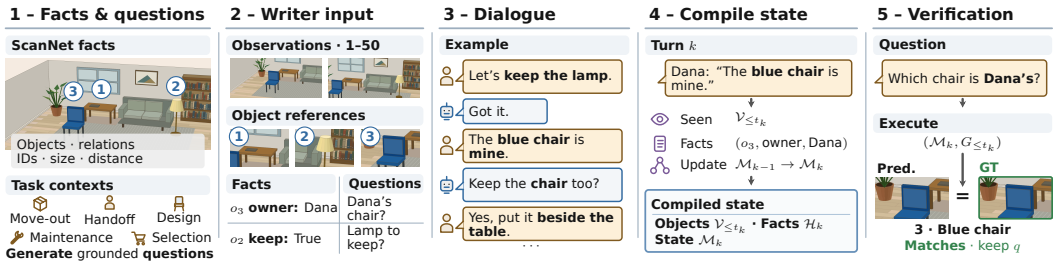

Figure 4: EGO-SPACR's pipeline for natural, verifiable dialogue over egocentric histories.

Table 1: Summary of 95 conversations spanning 620 video sessions and 3,100 evaluation questions in real ScanNet rooms (mean ± population s.d.; brackets give [min, max]).

Statistic	Value	Statistic	Value
Task types	5	Grounded objects per conversation	49.8 ± 24.3
Multi-session conversations	95	Object classes per conversation	27.4 ± 9.8 (60 total)
Video sessions (clips)	620 (6.4 h)	Evaluation questions	3,100
Semantic facts per conversation	201.2 ± 118.8	Questions per conversation	32.6 ± 14.9
Sessions per conversation	6.5 ± 2.7 [2, 14]	Complexity mix C1 / C2 / C3 / C4 (%)	6.0 / 24.3 / 45.9 / 23.8
Video per conversation (min)	4.1 ± 1.9	Answer type count / object / boolean (%)	85.4 / 13.4 / 1.2
Dialogue turns per conversation	230.1 ± 96.3 [64, 505]	Question length (words)	34.0 ± 13.8
Dialogue words per conversation	$5{,}315 \pm 2{,}534$	Dialogue turns / words before a question	158.8 / 3,797

2026) and GLM-5.3 FLASH (Z.ai, 2026). (ii) *Agentic spatial understanding with object memory.* We evaluate *VideoAgent* (Fan et al., 2024) and the agentic video mode of GEMINI 3.7 FLASH. (iii) *Multimodal RAG and conversational memory.* We evaluate *Multimodal RAG*, *Summary RAG*, and *M2A* (Feng et al., 2026), using GPT-5.6 TERRA and GLM-5.3 FLASH. We also include Random and Conversation Only controls (Appendix C).

5.1 DATASETS AND METRICS

EGO-SPACR. Existing benchmarks do not comprehensively evaluate SPACR (cf. Sec. 2). We therefore construct EGO-SPACR, an egocentric benchmark built from ScanNet (Dai et al., 2017) that covers the four properties of SPACR defined in Sec. 3. ScanNet provides real egocentric observations with annotated object identities and geometry, enabling verifiable spatial evidence and analysis of reconstruction error. Construction follows five stages (Fig. 4).

(1) *Facts & questions.* Five everyday tasks and compatible personas define plausible object-specific semantic facts. Clips from selected ScanNet scenes are composed into multiple visits, forming a shared fact pool Φ. We then construct queries by combining conversational conditions with spatial operations, each represented by a deterministic program G_k that produces a ground-truth answer from corresponding facts and ScanNet geometry. For count queries near geometric boundaries, we allow a bounded answer interval to account for uncertainty in the annotated mesh (cf. App. B).

(2) *Writer input.* For each scene-level batch, the Writer receives the relevant observations, grounded object references, facts to establish, selected queries, and preceding dialogue context.

(3) *Dialogue.* An LLM *Writer* generates a natural continuation that establishes the required facts before each query while maintaining coherence across visits.

(4) *Compile state.* An independent LLM *Reader*, without access to the Writer's annotations, processes the conversation turn by turn to reconstruct grounded facts and updates available.

(5) *Verification.* We retain a query only if all required facts precede q_k, all references resolve, and executing G_k over the Reader-recovered state reproduces the target answer, followed by human verification. We evaluate all methods using the following two metrics.

Answer accuracy. Our primary metric is final answer accuracy. For object-identification questions, the model returns a natural-language description of the selected object. We use a VLM only to map this description to a ground-truth physical object, using annotated reference images, bounding

Table 2: Answer Accuracy across selected baselines and SpaC-MEM (%); bold and underlined entries are the best and second-best values per column.

Method	Model	C1	C2	C3	C4	All
Random	–	4.3	12.0	21.9	7.9	15.1
Conversation Only	GLM-5.3 FLASH	41.7	39.0	22.0	23.2	27.6
In-Context Spatial Understanding						
Native Video	GEMINI 3.7 FLASH	74.3	55.9	31.5	37.3	41.4
	QWEN3.8 27B	64.2	46.3	11.3	8.5	22.3
Uniform	GLM-5.3 FLASH	38.3	43.3	20.5	16.8	26.2
	GPT-5.6 TERRA	63.6	54.8	29.3	33.1	38.5
AKS	GLM-5.3 FLASH	41.5	44.9	21.0	18.2	27.3
	GPT-5.6 TERRA	64.5	55.5	28.2	34.4	38.5
FOCUS	GLM-5.3 FLASH	38.8	43.0	20.0	17.3	26.1
	GPT-5.6 TERRA	59.0	55.9	28.9	34.8	38.7
Agentic Spatial Understanding with Object Memory						
VideoAgent	GLM-5.3 FLASH	36.2	38.7	16.3	13.4	21.0
Gemini Agentic	GEMINI 3.7 FLASH	72.2	55.9	35.4	37.3	43.1
Multimodal RAG and Conversational Memory						
Multimodal RAG	GLM-5.3 FLASH	41.7	23.0	4.1	2.4	10.6
	GPT-5.6 TERRA	35.3	26.0	5.3	3.9	11.8
Summary RAG	GLM-5.3 FLASH	26.2	32.0	8.6	14.5	16.8
	GPT-5.6 TERRA	31.6	31.3	8.7	13.0	16.6
M2A	GLM-5.3 FLASH	55.6	39.6	10.8	13.1	21.0
	GPT-5.6 TERRA	46.0	44.9	14.8	21.0	25.5
Ours						
SpaC-MEM	GLM-5.3 FLASH	85.0	**59.6**	38.2	60.3	51.5
	GPT-5.6 LUNA	**87.2**	59.0	**39.9**	**61.4**	**52.5**

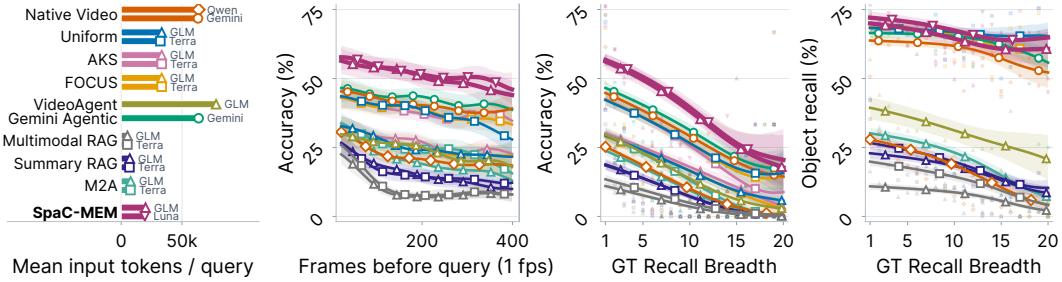

Figure 5: Left: reference input tokens across methods and accuracy as a function of input frames before queries. SPAC-MEM achieves the highest accuracy while using substantially fewer reference input tokens. Right: object recall and answer accuracy across recall breadth. Colors and markers are consistent across panels; SPAC-MEM maintains the high recall and accuracy as breadth increases.

boxes, and spatial information. Accuracy is then determined by whether the mapped object matches the ground-truth object (see Appendix F for details).

Object recall. To verify that a method recovers the objects needed rather than guessing the count, we ask each answer model to also return the object descriptions supporting its count and map them to physical objects using the same VLM. Let $\mathcal{R}_k$ be the ground-truth objects required for count query q_k, and $\hat{\mathcal{R}}_k$ the objects returned by the model. For queries with $|\mathcal{R}_k| > 0$, object recall is $\frac{|\hat{\mathcal{R}}_k \cap \mathcal{R}_k|}{|\mathcal{R}_k|}$ the fraction of required objects recovered, averaged over nonempty count queries. We compare the context supplied to the final answer model using reference input tokens under a common scheme across providers (Appendix D). For SPAC-MEM, we use DEPTH ANYTHING 3 (Lin et al., 2025) for reconstruction and SEGVGGT (Qu et al., 2026) for segmentation.

5.2 RESULTS

RQ1. We first compare SPAC-MEM with existing methods. As shown in Tab. 2, SPAC-MEM achieves the highest overall accuracy on EGO-SPACR (52.5%), outperforming the strongest baseline (43.1%) using GPT-5.6 Luna as the answer model. The advantage becomes larger on more demanding spatial queries: SPAC-MEM reaches 61.4% on C4, compared with 37.3% for the strongest baselines, while remaining competitive on the rest. SPAC-MEM also maintains stronger accuracy as recall breadth and the number of prior frames before query increase. Importantly, it achieves this performance using only about one quarter of the reference input tokens required by native-video methods, showing that an explicit object-centric memory can preserve the information needed for SPACR without repeatedly processing the full multimodal history.

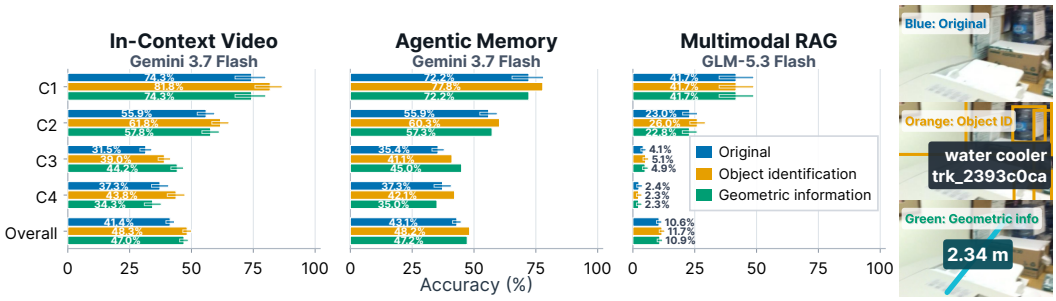

Figure 6: Oracle ablations between the original frame, oracle object box, and oracle geometric measurement (illustrated on the right panel). Ground-truth boxes yield only marginal gains, while oracle geometric measurements mainly improve C3 but not C4.

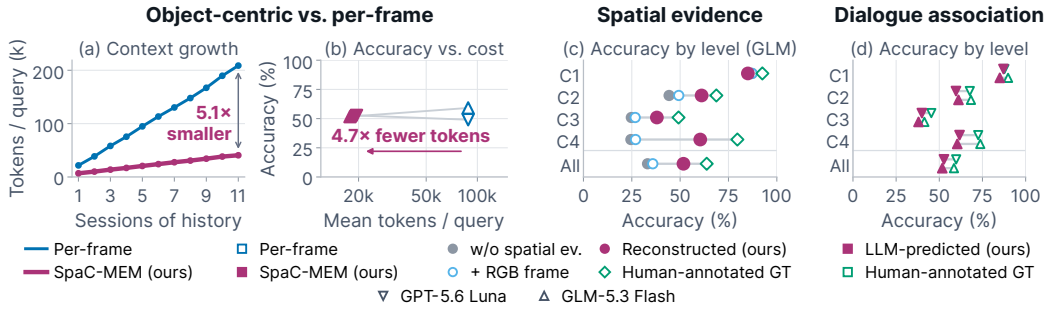

Figure 7: Ablations of SPAC-MEM: object-centric memory needs 4.7× fewer tokens than per-frame memory (a, b), removing spatial evidence lowers accuracy (c), and LLM-predicted dialogue association stays close to human-annotated GT (d).

RQ2. To separate object-grounding errors from spatial-reasoning errors, we perform controlled oracle ablations for representative baselines in each method group. (1) *Oracle object bounding boxes* provide ground-truth boxes to remove object localization and separation errors. As shown in Fig. 6, this yields only a marginal improvement, suggesting that object identification is not the primary bottleneck. (2) *Oracle geometric measurements* additionally provide the spatial quantities required by each query. This improves in-context and agentic baselines on C3, which directly tests geometric reasoning, but produces little improvement on C4 counterfactual queries. These results suggest that the main limitation of existing baselines relative to SPAC-MEM is not object detection or geometric measurement alone, but the lack of a consistent spatial representation that can preserve and combine spatial evidence across observations for SPACR.

RQ3. Fig. 7 evaluates design choices in SPAC-MEM. (1) *Object-centric vs. per-frame memory:* SPAC-MEM reduces context by 5.1× at 11 sessions and uses 4.7× fewer input tokens, with comparable accuracy. (2) *Geometric vs. non-geometric memory:* removing spatial detail reduces accuracy from 51.5% to 34.9%, and C4 from 60.3% to 24.5%; adding each object's RGB view recovers little of this loss. The oracle reconstruction provides a further reference for the quality of the object records. (3) *LLM vs. human dialogue association:* the method retains 89% of the accuracy obtained with human-annotated associations, showing that most conversational grounding is preserved.

6 CONCLUSION

We introduced EGO-SPACR to evaluate spatially grounded conversational reasoning across egocentric observations and evolving dialogue, and SPAC-MEM to connect both histories through persistent object records. SPAC-MEM achieves the highest overall answer accuracy among the evaluated methods while requiring fewer reference input tokens than native-video baselines. Ablations show that retaining object geometry supports both answer accuracy and object recall.

AI USE STATEMENT

Language models are used in benchmark construction to generate dialogue and independently extract its semantics, and vision-language models support selected reference and answer judgments. Their outputs are subject to the execution and evidence checks described in Section 5.1. AI assistance was also used to support research planning and literature discovery, draft and edit prose, and prepare LaTeX. Model identifiers for construction and evaluation are given in Appendix B and Section 5.1.

ETHICS STATEMENT

The proposed use case involves personal spaces and information about users and bystanders. The benchmark composes existing research scans with synthetic conversations; synthetic ownership, preferences, and roles must not be interpreted as facts about the people who originally occupied those spaces. Any release of source imagery must follow ScanNet's terms and applicable privacy requirements. The work does not establish safe deployment or justify continuous recording.

REPRODUCIBILITY STATEMENT

Sections 5.1 and 4 describe the data construction and memory interfaces. Appendices B and E give further construction and method details. The intended artifacts include sampled clip manifests, task and persona settings, prompts, executable query definitions, public inputs, hidden evaluation records, and exact model configurations. Raw ScanNet data remain subject to their original access conditions.

REFERENCES

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

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop VideoQA in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025. doi: 10.1609/aaai.v39i2.32214. URL https://ojs.aaai.org/index.php/AAAI/article/view/32214.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1702.04405.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision*, pp. 720–736, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Dima_Damen_Scaling_Egocentric_Vision_ECCV_2018_paper.html.

Chenyou Fan. EgoVQA: An egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. URL https://openaccess.thecvf.com/content_ICCVW_2019/html/EPIC/Fan_EgoVQA_-_An_Egocentric_Video_Question_Answering_Benchmark_Dataset_ICCVW_2019_paper.html.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, and Wentao Zhang. M2A: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://video-mme.github.io/.

Google. Gemini 3.7 flash. Google AI for Developers, 2026. URL https://ai.google.dev/gemini-api/docs/models.

Google DeepMind. Gemini 3.7 flash: Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/. Published August 13, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

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. In *IEEE International Conference on Robotics and Automation*, pp. 5021–5028, 2024. doi: 10.1109/ICRA57147.2024.10610243. URL https://arxiv.org/abs/2309.16650.

Tianjun Gu, Tianyu Xin, Kuan Zhang, Bowen Yang, Kok-Chung Chua, Peize Li, Xinran Zhang, Yupeng Chen, Qiyue Zhao, Qinlei Xie, Jianhang Liu, Yucheng Lu, Yinan Han, Marco Pavone, and Yiming Li. Towards spatial supersensing in the wild. In *The Nineteenth European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.13681.

Sen Hu, Yuxiang Wei, Jiaxin Ran, Zhiyuan Yao, Xueran Han, Huacan Wang, Ronghao Chen, and Lei Zou. Does Memory Need Graphs? A Unified Framework and Empirical Analysis for Long-Term Dialog Memory, February 2026. URL http://arxiv.org/abs/2601.01280. arXiv:2601.01280 [cs.CL].

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1096.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4903–4912,

2021. doi: 10.18653/v1/2021.emnlp-main.401. URL https://aclanthology.org/2021.emnlp-main.401/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.11401.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

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*, pp. 16488–16498, 2024. URL https://open-eqa.github.io/.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

OpenAI. Gpt-5.6 terra. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra.

Jinyuan Qu, Hongyang Li, and Lei Zhang. SegVGGT: Joint 3d reconstruction and instance segmentation from multi-view images. *arXiv preprint arXiv:2603.19926*, 2026.

Qwen Team. Qwen3.8. Alibaba Group, 2026. URL https://github.com/QwenLM/Qwen3.8.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025. URL https://arxiv.org/abs/2502.21271.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20270–20281, October 2023.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2407.15754.

Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6273–6291, 2023. doi: 10.18653/v1/2023.acl-long.345. URL https://aclanthology.org/2023.acl-long.345/.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. 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. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rBFDvZu6pb.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3D-Mem: 3D scene memory for embodied exploration and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17294–17303, June 2025b.

Z.ai. GLM-5.3-Flash/FlashX: Model overview. Z.ai Developer Documentation, 2026. URL https://docs.z.ai/guides/vlm/glm-5.3-flash.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a010577f93a8301dada61f0b97cda923-Abstract-Conference.html.

A Related Work Comparison

Table 3: Benchmark coverage of visual input, conversation, object identity, and evaluation. ✓: explicit support; ◐: related or limited support; ×: not specified.

	Visual input		Conversation		Objects	Evaluation			
						Spatial reasoning			
Benchmark	Egocentric video	Real scenes	Dialogue history	Object-specific user semantics	Cross-view object continuity	View grounding	Spatial relations	Geometric properties	Recall breadth
First-person perception									
EPIC-KITCHENS (Damen et al., 2018)	✓	✓	×	×	×	◐	×	×	×
Ego4D (Grauman et al., 2022)	✓	✓	×	×	✓	◐	◐	◐	✓
Situated dialogue and spatial understanding									
SIMMC-VR (Wu et al., 2023)	✓	×	✓	✓	✓	✓	✓	×	×
OpenEQA (Majumdar et al., 2024)	✓	✓	×	×	×	◐	✓	◐	✓
SQA3D (Ma et al., 2023)	✓[†]	✓	×	×	×[‡]	◐	✓	×	✓
VSI-Bench (Yang et al., 2025a)	✓	✓	×	×	×[‡]	◐	✓	✓	✓
VSI-SUPER (Yang et al., 2026)	✓	✓	×	×	✓	◐	×	×	✓
VSI-Super-Wild (Gu et al., 2026)	✓	✓	×	×	✓	◐	◐	×	✓
Video question answering									
EgoVQA (Fan, 2019)	✓	✓	×	×	×	◐	◐	×	✓
EgoSchema (Mangalam et al., 2023)	✓	✓	×	×	×	×	×	×	×
MULTIHOP-EGOQA (Chen et al., 2025)	✓	✓	×	×	✓	◐	◐	×	✓
LongVideoBench (Wu et al., 2024)	×	✓	×	×	✓	◐	×	×	×
Video-MME (Fu et al., 2025)	×	✓	×	×	×	◐	◐	×	✓
Conversational memory									
LoCoMo (Maharana et al., 2024)	×	×	✓	×	×	×	×	×	✓
LongMemEval (Wu et al., 2025)	×	×	✓	×	×	×	×	×	✓
EGO-SPACR (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

[†] SQA3D includes an evaluated egocentric-video input option. [‡] Static scene annotations alone are not treated as explicit cross-view object continuity.

B Additional Benchmark Details

B.1 ScanNet observations and episode composition

We build the visual history from ScanNet RGB-D recordings (Dai et al., 2017). RGB frames form the observations shown to the assistant. Camera poses, frame-level instance annotations, and the annotated 3D mesh give us persistent object identities, visibility, and spatial measurements for constructing and checking questions. The instance labels link the egocentric views to the reference geometry, letting us ask about objects seen in video and check their spatial properties in a consistent physical scene.

Each conversation combines clips from selected scenes into several visits. The visits share a task and a pool of object-specific semantic facts Φ, while recording gaps allow a conversation to continue in another room or return to an earlier one. We preserve object identity when a scene is revisited and keep objects from different scenes distinct. The rooms and their geometry remain those of the source scans; task roles, semantic facts, dialogue, and the order of visits are composed for the benchmark. For dataset splits, all clips from a physical scene stay in the same split. We use shards 01–19 of the v1.0 release: 95 conversations with 620 video sessions. Of 3,218 released questions, 3,100 enter the official evaluation after excluding 118 questions that do not require spatial evidence. Conversation statistics use the full histories; query statistics use this evaluation subset.

Table 4: Five everyday tasks used to guide the conversational facts in each episode. Each episode keeps one task and a compatible persona across all selected clips; Appendix B.2 describes their role in conversation generation.

Task	Purpose	Example properties p	Example persona
Inventory walkthrough	Record what is in the space for later use.	Owner, provenance, history, and alias.	Roommate; new tenant.
Guided room tour	Explain the space so another person can use it later.	Function, history, owner, and alias.	Welcoming host; prospective tenant.
Lab or office setup	Prepare a shared work area without disturbing fixed items.	Constraint, responsibility, maintenance, and priority.	Lab member; organizer.
Workspace setup	Decide how to arrange a usable workspace.	Preference, intention, constraint, and function.	Remote worker; interior designer.
Cleaning and organizing	Decide what should be cleaned, kept, moved, or discarded.	Maintenance, status, decision, and planning.	Organizer; new tenant.

B.2 Tasks, personas, and conversational state

The five task settings give the conversation a reason to retain information about particular objects. An inventory can concern ownership and provenance; a room tour can establish functions and local names. Lab or office setup introduces responsibilities and access; workspace setup introduces intended uses and preferences; and cleaning and organizing distinguishes condition from a decision about what to do. We pair each task with a compatible persona and one of its three authored narrative variants. The role and way of speaking remain stable across visits. A task or persona frames the dialogue; an object acquires a semantic property only when the conversation establishes it.

We write an object-specific fact as $\phi = (o, p, x)$, with physical object o, property p, and value x, and retain when it was stated. The fact pool may include later updates. A *correction* replaces an earlier value, including for retrospective questions; a *state change* starts a new period while preserving the earlier value for questions about the past. The query program uses the value applicable to its temporal scope.

References are grounded in the observation and dialogue history available at the turn where they occur. A class name can identify an object when the context makes it unique. Otherwise, an appearance, spatial relation, established name, or clarification distinguishes it from similar objects. Descriptions tied to a particular view are used in the window where they were checked; an established name can carry the reference into later visits. This gives the Reader a recoverable link between conversational facts and ScanNet objects.

B.3 Conversation generation

Model versions. We use GPT-5.6 Sol as the Writer and GPT-5.6 Terra as the independent Reader, with low reasoning effort for the Reader. GPT-5.6 Luna resolves free-form object descriptions during evaluation.

We select queries from Φ whose semantic prerequisites and spatial evidence can be available before the question. The selected queries cover the complexity classes and recall breadths defined in the main text. Each has a deterministic program G_k, a temporal scope, and a placement window in which its answer is stable. Required facts are scheduled before that window. Other compatible facts can enter the dialogue as the task develops.

An LLM *Writer* generates the conversation in scene-level batches. It receives the task and persona, the preceding dialogue and summary, relevant RGB observation windows, grounded reference options, and the facts to establish. The query specification tells it what condition to ask about and when; the answer and qualifying object set are withheld. Carrying the prior dialogue across batches lets a later visit continue the same task and refer back to established facts. Tab. 5 summarizes the principal inputs.

Table 5: Selected inputs to the Writer, with illustrative field values from one organizing task. Spatial context supports object references without disclosing the spatial outcomes being tested. Visual descriptions are checked independently before inclusion; query answers are never supplied.

Task and conversation context
Task and persona: The continuing goal and the speaker's role and manner.
Example: Cleaning and organizing; a considerate organizer; continue in the study.
Prior context: Earlier dialogue, established object facts, and unfinished discussion.
Example: The living room lamp was marked for cleaning; the walkthrough continues.
Observation windows
Time and view: A video interval and its representative frame.
Example: Study window w_1, frames 120–180; representative frame 145.
Visibility changes: Objects in view, newly encountered, revisited, or leaving view.
Example: Two lamps and a cabinet appear; a previously seen sofa remains visible.
Spatial context: Relations available in the window that can support object descriptions.
Example: Cabinet near sofa, available in w_1 after both objects have been observed.
Object references
Object identity: The object's class, earlier sightings, and similar objects to distinguish.
Example: Two study lamps are distinct from each other and from the living room lamp.
Permitted descriptions: Natural references and the observations needed to use them.
Example: "The cabinet near the sofa" becomes usable after both objects are observed.
Visual references for ambiguous objects
Checked description: A VLM describes a marked object; an independent VLM, without the target cue, must identify only that object from the phrase.
Example: "The lamp with the red shade" uniquely resolves to the intended study lamp.
Valid frame: The exact view where the phrase was verified; later references require an established name or another valid description.
Example: Use the phrase at frame 145; name it "reading lamp" in that turn for later references.
Fact pool
Required facts: Object properties that must be expressed before their dependent questions.
Example: Reading lamp: maintenance = needs cleaning.
Optional facts: Additional task information the Writer may include naturally.
Example: Cabinet: comment = cluttered.
Timing and dependencies: When a fact can be introduced and which statements must precede it.
Example: Establish the lamp's cleaning fact in w_1 before the question in w_2.
Question intents
Meaning and answer type: The conversational condition, spatial test, and requested result.
Example: Marked for cleaning; on a table; return the number of matching objects.
Scope and placement: Which observations the question covers and when it may be asked.
Example: All rooms visited so far; ask in w_2 after all required facts are established.

An independent *Reader* then processes the realized conversation turn by turn. Using the visual and dialogue history available at each turn, it resolves mentions and reconstructs the stated facts and updates. We use this recovered state for answer execution, so the benchmark answer follows the conversation that was written.

B.4 Query construction and recall

The four spatial complexity classes follow Sec. 3. Qualitative and geometric predicates operate on objects and measurements from the same source scene, even when a question aggregates results across scenes. Counterfactual predicates use objects already observed and evaluate a hypothetical viewpoint or placement against the available scene geometry. Recall-oriented questions require the complete set of objects satisfying both the conversational and spatial conditions. We retain that set alongside the final answer so a correct count can be checked against the objects that support it.

B.5 Ground truth and answer evaluation

Let $\Phi_{<k}$ be the Reader-recovered facts established before q_k, resolved for its temporal scope, and $\mathcal{G}_{\leq t_k}$ the ScanNet spatial evidence available up to t_k. We compute the canonical answer by executing the question program over that prefix:

$$a_k = \mathrm{Exec}(G_k; \Phi_{<k}, \mathcal{G}_{\leq t_k}). \tag{1}$$

Execution also records the qualifying objects and the spatial measurements that support the answer. This matters for a count: two methods can return the same integer while identifying different objects.

For count questions involving measured geometric boundaries, we retain the literal comparison in G_k and record how far each candidate lies from the boundary. Candidates clearly inside contribute to the lower count L_k; those within the measurement tolerance also contribute to the upper count U_k. The accepted answer is any integer in $[L_k, U_k]$. We use a 0.04 m tolerance for distance and extent comparisons, reflecting the resolution of the annotated ScanNet mesh. During query selection, we bound the interval width to two counts. When no candidate lies near the boundary, $L_k = U_k = a_k$. This treatment preserves the stated spatial condition while allowing for small measurement differences at its edge.

Models still return a single count. We score it by membership in the accepted interval when one applies; other count answers and Boolean answers use exact matching. For object identification, a VLM maps the model's natural-language description to a physical object using the observations available before the question. The mapped object is compared with the target fixed by G_k. Ambiguous descriptions remain unresolved. Models see neither the hidden object identifiers nor a candidate answer list.

B.6 Verification and quality review

We check each realized question against the Reader-recovered prefix. Its required facts must have been stated before the question, its references must resolve at the turns where they are used, and its objects and spatial evidence must be available by the query time. We execute the program again at the realized placement and retain the question when the result agrees with the planned answer. This also checks that conversational updates and optional facts have preserved the intended temporal and semantic conditions. Visual descriptions used for object answers must provide enough evidence for reliable grounding.

Human review examines the resulting video and dialogue for semantic grounding, reference resolution, temporal scope, conversational plausibility, and answer correctness. The automated checks and this review address different failure modes: execution checks the formal conditions, while the review checks that the question and conversation express them clearly.

C Control Baselines

Random. For count and Boolean questions, this control returns the mean canonical answer over the evaluated set, rounded to the nearest valid integer or Boolean value. For object identification, it samples uniformly from the distinct objects in the available memory before the query. Sampling uses seed 20260925 with a separate deterministic draw for each query. Count and Boolean predictions follow the standard answer rules; object predictions are compared directly with the target identity.

Conversation Only. We give GLM-5.3 Flash the timestamped public dialogue strictly before the query, followed by the question and its answer format. The system prompt and output format match the frame-based baselines, with all images and the frame-index block removed. We use low reasoning effort and a maximum output length of 32,768 tokens.

D Reference Input Cost

Provider-reported input-token counts are not directly comparable across our baselines because different models use different text tokenizers and visual tokenization schemes. We therefore report *reference input tokens*, a model-independent measure of the context delivered to the final answer

model. The goal is not to reproduce provider billing, but to compare how much textual and visual context each method requires at query time.

For method m answering query q, we define

$$T_{\mathrm{ref}}(m,q) = |\mathrm{tok}_{\texttt{o200k}}(s_m \| u_{m,q})| + c_{\mathrm{frame}} F_{m,q},$$

where s_m and $u_{m,q}$ are the system and user prompts actually provided to the answer model, and $F_{m,q}$ is the number of RGB frames provided with that request. We tokenize all text using the same `o200k_base` tokenizer. All visual inputs are evaluated at 640×480 resolution and assigned a common reference cost of one token per 32×32 patch, giving $c_{\mathrm{frame}} = \lceil 640/32 \rceil \lceil 480/32 \rceil = 300$ reference tokens per frame. For image-based methods, $F_{m,q}$ is the number of attached images; for native-video methods, it is the number of frames sampled into the causal video input at 1 fps.

This measure is computed offline from saved answer requests and is applied identically across methods. It measures the context exposed to the final answer model rather than total pipeline compute: retrieval embeddings, summary generation, memory updates, agent tool calls, output tokens, and evaluation calls are excluded. Internal frame re-inspection by agentic methods is likewise not charged again because it does not change the context of the final answer request. For text-only runs whose complete prompts were not retained, we use the recorded input-token count only when the serving tokenizer matches `o200k_base`. We report the mean T_{ref} over all scored answer requests for each method.

E Additional Method Details

E.1 Object reconstruction and consolidation

Geometry. We use the DA3-Nested-Giant-Large checkpoint of Depth Anything 3 (Lin et al., 2025) as f_{geo}, yielding approximately metric geometry. SpaC-MEM places all frames of a scene in one shared coordinate system, with camera intrinsics, poses and depth all estimated from the RGB frames alone. Only pixels whose depth in $\hat{D}_t$ is estimated with high confidence are lifted into 3D. The scene's axes are aligned with its dominant planes: the vertical axis with the floor plane or, when too little floor is visible, with walls facing different directions or the largest horizontal surface, and the horizontal axes with the dominant wall direction, so that axis-aligned boxes follow the room's layout. The floor is placed at height zero.

Segmentation and lifting. Wall and floor are excluded from $\mathcal{L}$, since they form the scene's structure rather than objects. Instances are kept when f_{seg} is sufficiently confident, which keeps spurious detections out of the memory. Where the model also ranks alternative classes, the next most likely ones are kept as runner-up classes. To decide which mask pixels are consistent with the object, SpaC-MEM uses the mask's interior, away from its boundary, as a reference: $\tilde{M}_{t,i}$ keeps the pixels of $M_{t,i}$ whose depth lies within the interior's central depth range, widened by a small tolerance. Edge pixels whose depth matches the object are therefore kept, including thin parts such as chair legs. The centroid $\boldsymbol{\mu}_{t,i}$ is the mean of $\hat{P}_{t,i}$, and the box $\mathbf{B}_{t,i}$ spans $\hat{P}_{t,i}$ along each axis of the scene's coordinate system, excluding the most extreme few points, which makes it robust to stray points.

Observation matching. Each observation $\hat{o}_{t,i}$ is compared with the objects in the scene's pool that carry its class label. Their overlap $A(\hat{o}_{t,i}, \hat{o})$ is the share of the points of $\hat{o}_{t,i}$ that lie within 10 cm of the accumulated points of $\hat{o}$. The match $a_t(\hat{o}_{t,i})$ is accepted when this share is large enough and clearly exceeds that of every other candidate. All observations of a frame are matched jointly, and each object takes at most one of them, since an object appears only once in a view. This keeps neighbouring objects of the same class apart even when they are seen together. A matched observation adds its points to $\hat{o}$, updating $s_{\leq t}(\hat{o})$. The class label of $\hat{o}$ is the one best supported across its observations, and other well-supported classes are kept as runner-up classes.

Pool refinement. Two objects in the pool are merged when their 3D points overlap substantially, they never co-occur in the same frame, and they either share a track from f_{seg} or have similar appearance. Overlapping pieces that never co-occur are likely views of one object, and the shared track or appearance keeps distinct objects that happen to overlap apart. Each object is kept once it

is confirmed by several observations. Isolated points are removed before the final box is computed, so that each box follows the object's main body.

Class labels and runner-up classes. Each persistent object carries one class label together with up to two runner-up classes, retained as secondary label evidence. Both the association model and the answering agent are told that these describe a single persistent object rather than additional objects, and that an object carrying a queried class label outright is preferred over one listing it as a runner-up. An object with no credible runner-up carries none.

Coordinates. Each scene retains its own levelled coordinate frame with approximately metric scale. Spatial predicates are evaluated within a scene, and cross-scene queries aggregate their results, so a shared origin and orientation across scenes are unnecessary.

E.2 Dialogue association and working memory

Association model. For each query, f_{assoc} receives the preceding dialogue and objects observed by the query time, grouped by scene. Each object has an identifier, class labels, centroid and bounding box. Turns are supplied in order with objects visible within two seconds, restricted to the available query prefix; the query itself is withheld. The model resolves class descriptions, pronouns and established names using this context. It returns a turn, object, referring phrase and role for each association, with identifiers constrained to the supplied inputs. Only *subject* roles enter $\hat{\mathcal{A}}_k$: in *"the monitor beside the laptop is Sam's,"* the monitor is the subject and the laptop is a *locator*. A unique class mention can help resolve a reference but does not establish a subject role by itself. References without an identifiable subject remain unassigned.

Contents of the working memory. $\mathcal{W}_k$ holds one record per persistent object observed up to t_k, grouped by scene and session and ordered by class label within a scene. Each record carries the object's class label and runner-up classes, its centroid and axis-aligned bounding box, the frame at which it first came into view, and the camera position at that frame, which the egocentric queries use as the position the user was standing at. Beneath the record sit the dialogue turns associated with the object, verbatim and in conversational order, each annotated so that the phrase referring to this object is marked. A turn about several objects therefore appears under each of them with a different phrase marked, and the answering agent is instructed to read the clause carrying the mark as the statement about that object. A table mapping days to scenes precedes the records, and $\hat{\mathcal{O}}_{t_k}$ follows them.

F VLM Judge for Object Identification

Object-identification queries require a method to name a single physical object in free-form text, for example, "Which single thing I said was an early priority was near a copier?" Correctness must therefore be determined by resolving the generated description to the ground-truth object. A target-only judge that sees only evidence for the correct object is too permissive in this setting: a generic answer such as "the table" may be accepted whenever the target is the only table in the displayed frame, even when several tables satisfy the query's conversational condition. Such a judge can therefore reward class-level guessing without establishing that the method resolved the spatial condition.

Candidate-resolution procedure. For each identification query, we first construct the candidate set from objects satisfying its conversation-derived condition. This set is fixed by the benchmark program before method evaluation and always contains the ground-truth target. Every candidate is then presented uniformly: one RGB frame is selected where its ground-truth instance mask is largest, only that object is boxed and assigned a letter, and candidates are shuffled with a fixed per-query seed. The judge is not told which candidate is correct.

Each candidate is accompanied by its object class, conversation facts, first-seen day, and a compact scene graph computed from ground-truth geometry. The graph contains objects within 3 m of the centroid, ordered nearest first, together with *on* and *supports* relations; the benchmark definitions

use *near* $\leq$ 1.5 m and *far* $\geq$ 3 m. This context allows descriptions such as "the pillow by the whiteboard" to be checked even when the referenced neighbour is outside the selected frame.

Given a model answer, the judge returns every candidate consistent with the answer's own description using object class, appearance, dialogue names and facts, and location claims. To avoid solving the query on behalf of the evaluated method, the judge does not apply the query's selection rule itself. Thus, a class-only answer or a restatement of the conversational condition remains compatible with every otherwise matching candidate. The judge also ignores location claims that merely repeat the query's own reference object; only additional location evidence supplied by the answer may eliminate a candidate, and only when that claim is clearly inconsistent with the candidate's scene graph.

An answer is correct if and only if the resolved candidate set contains exactly the ground-truth target. We use three independent judge calls at temperature 0 and take the majority verdict. Count and Boolean queries remain scored deterministically; the candidate-resolution judge is used only for object-identification queries.

To validate the judge, 500 identification answers sampled across 12 baselines were independently verified by a human annotator using the same candidate presentation.

Table 6: Answer Accuracy across selected baselines and SpaC-MEM (%) with Wilson 95% confidence interval half-widths in gray; bold and underlined entries are the best and second-best values per column.

Method	Model	C1	C2	C3	C4	All
Random	–	4.3 ±3.0	12.0 ±2.3	21.9 ±2.1	7.9 ±1.9	15.1 ±1.3
Conversation Only	GLM-5.3 FLASH	41.7 ±7.0	39.0 ±3.5	22.0 ±2.2	23.2 ±3.0	27.6 ±1.6
In-Context Spatial Understanding						
Native Video	GEMINI 3.7 FLASH	74.3 ±6.2	55.9 ±3.5	31.5 ±2.4	37.3 ±3.5	41.4 ±1.7
	QWEN3.8 27B	64.2 ±6.8	46.3 ±3.6	11.3 ±1.6	8.5 ±2.0	22.3 ±1.5
Uniform	GLM-5.3 FLASH	38.3 ±7.0	43.3 ±3.6	20.5 ±2.1	16.8 ±2.7	26.2 ±1.6
	GPT-5.6 TERRA	63.6 ±6.8	54.8 ±3.5	29.3 ±2.4	33.1 ±3.4	38.5 ±1.7
AKS	GLM-5.3 FLASH	41.5 ±7.1	44.9 ±3.6	21.0 ±2.1	18.2 ±2.8	27.3 ±1.6
	GPT-5.6 TERRA	64.5 ±6.9	55.5 ±3.6	28.2 ±2.4	34.4 ±3.4	38.5 ±1.7
FOCUS	GLM-5.3 FLASH	38.8 ±7.0	43.0 ±3.6	20.0 ±2.1	17.3 ±2.7	26.1 ±1.6
	GPT-5.6 TERRA	59.0 ±7.1	55.9 ±3.6	28.9 ±2.4	34.8 ±3.4	38.7 ±1.7
Agentic Spatial Understanding with Object Memory						
VideoAgent	GLM-5.3 FLASH	36.2 ±7.1	38.7 ±3.6	16.3 ±2.0	13.4 ±1.8	21.0 ±1.8
Gemini Agentic	GEMINI 3.7 FLASH	72.2 ±6.4	55.9 ±3.5	35.4 ±2.5	37.3 ±3.5	43.1 ±1.7
Multimodal RAG and Conversational Memory						
Multimodal RAG	GLM-5.3 FLASH	41.7 ±7.0	23.0 ±3.0	4.1 ±1.0	2.4 ±1.1	10.6 ±1.1
	GPT-5.6 TERRA	35.3 ±6.8	26.0 ±3.1	5.3 ±1.2	3.9 ±1.4	11.8 ±1.1
Summary RAG	GLM-5.3 FLASH	26.2 ±6.3	32.0 ±3.3	8.6 ±1.5	14.5 ±2.5	16.8 ±1.3
	GPT-5.6 TERRA	31.6 ±6.6	31.3 ±3.3	8.7 ±1.5	13.0 ±2.4	16.6 ±1.3
M2A	GLM-5.3 FLASH	55.6 ±7.0	39.6 ±3.5	10.8 ±1.6	13.1 ±2.4	21.0 ±1.4
	GPT-5.6 TERRA	46.0 ±7.1	44.9 ±3.5	14.8 ±1.8	21.0 ±2.9	25.5 ±1.5
Ours						
SpaC-MEM	GLM-5.3 FLASH	<u>85.0</u> ±5.1	**59.6** ±3.5	<u>38.2</u> ±2.5	<u>60.3</u> ±3.5	<u>51.5</u> ±1.8
	GPT-5.6 LUNA	**87.2** ±4.8	<u>59.0</u> ±3.5	**39.9** ±2.5	**61.4** ±3.5	**52.5** ±1.8

G ADDITIONAL EXPERIMENTAL RESULTS

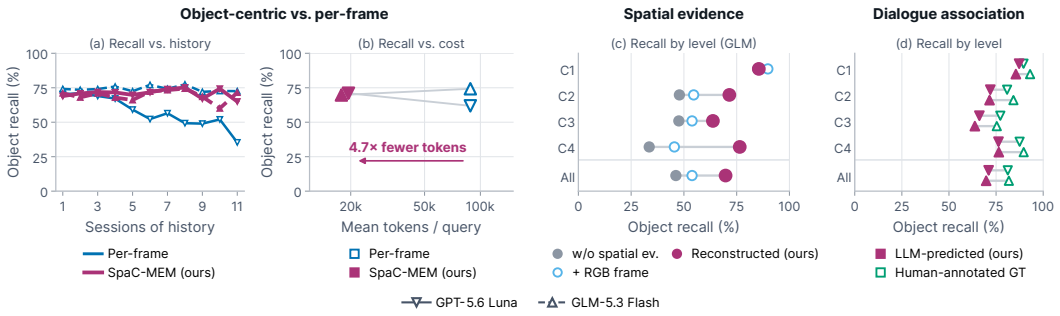

Figure 8: Object recall for the ablations in Fig. 7: SPAC-MEM keeps about 70% recall as history grows while per-frame memory degrades (a) at $4.7\times$ the token cost (b), spatial evidence drives recall on spatial queries (c), and LLM-predicted dialogue association retains most of the recall of human-annotated GT association (d).

H ADDITIONAL ANALYSIS OF OBJECT-CENTRIC MEMORY

H.1 OBJECT RECALL IN THE ABLATIONS

Fig. 8 reports the fraction of ground-truth objects recovered by the answering model. *Object-centric vs. per-frame memory:* SPAC-MEM maintains roughly 70% recall as interaction history grows, while using $4.7\times$ fewer input tokens; with GPT-5.6 Luna, per-frame recall drops from 73% in the first session to 35% after 11 sessions. *Geometric vs. non-geometric memory:* preserving 3D spatial records raises recall from 46.2% to 69.7% overall and from 33.6% to 76.5% on C4, showing that spatial evidence is critical for recovering the correct object set. *LLM vs. human dialogue association:* automatic associations retain 87% and 85% of the recall achieved with human annotations for Luna and GLM-5.3 Flash, respectively.

H.2 SCALING WITH MEMORY SIZE

Fig. 9 relates answer accuracy and object recall to the number of persistent objects in the working memory $\mathcal{W}_k$ at query time. Across the range covering 95% of queries, object recall remains near 70% for both answering models, while accuracy changes gradually with memory size. In the object-centric design of Sec. 4, repeated observations update the same object record, and its associated dialogue remains available to the answerer.



Figure 9: Answer accuracy and object recall as the number of objects in $\mathcal{W}_k$ grows. Solid and dashed lines use GPT-5.6 Luna and GLM-5.3 Flash, respectively. Curves are smoothed trends with 95% episode-bootstrap intervals; markers show means within 25-object bins. The horizontal axis covers 95% of queries.

	0.25 m	0.5 m	1.0 m
Instance recall (%)	70.7	**91.1**	97.3
Semantic recall (%)	65.6	**82.2**	85.4
Instance precision (%)	52.0	67.0	71.6
Median centroid error of matched objects: 16.2 cm			
Recall at 3D box IoU 0.25: 51.7%			

Table 7: Reconstructed object records compared with ScanNet annotations across 39 scenes and 864 annotated objects. Instance matching uses centroid distance; semantic recall also requires agreement on the ScanNet200 class.

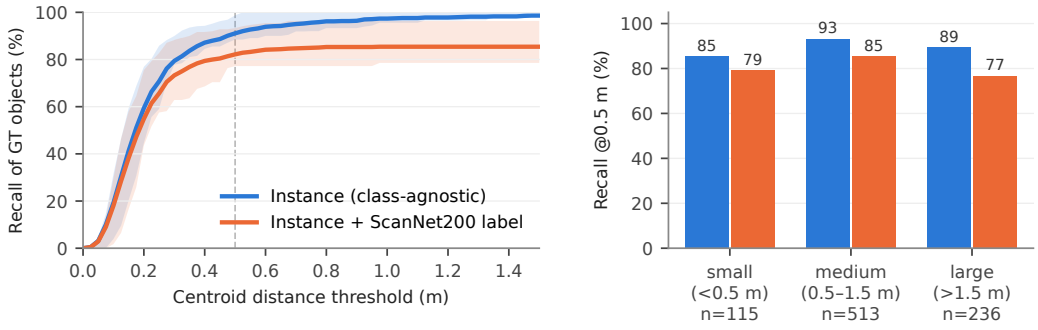

Figure 10: Reconstruction recall by centroid-distance threshold (left) and annotated object size at 0.5 m (right). The bands show the per-scene interquartile range. Instance recall is class-agnostic; semantic recall also requires the correct ScanNet200 label.

H.3 Fidelity of reconstructed objects

The object reconstruction and consolidation stages build the persistent pool $\hat{\mathcal{O}}_{\leq t}$ from RGB observations. We compare its spatial records $s_{\leq t}(\hat{o})$ with annotated ScanNet objects to assess how much object evidence reaches the working memory. Each of the 39 benchmark scenes is evaluated once, giving 864 annotated objects in the ScanNet200 vocabulary. We align the reconstructed and ScanNet coordinate systems using corresponding camera positions, then match objects one-to-one by centroid distance. Matching is independent of class; semantic recall additionally requires the reconstructed class label to agree with the annotation.

Tab. 7 reports object-level results at three distance thresholds. At 0.5 m, the persistent pool recovers 91.1% of annotated objects, and 82.2% are recovered with the correct class label. The median centroid error among matches is 16.2 cm. Fig. 10 shows how recall varies with matching distance and object size.