VTEM: VIEW-TRANSITION EGOCENTRIC MEMORY FOR LONG-HORIZON EMBODIED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon embodied reasoning is fundamental to the reliable deployment of embodied agents, enabling informed decision-making from observations accumulated during long-term interactions. However, existing memory-based approaches often fail to establish effective spatiotemporal associations across observations, causing redundant cross-view content to accumulate in memory and retrieval to return context-misaligned evidence. To address these limitations, we propose the View-Transition Egocentric Memory framework (VTEM), which models view transitions to support traceable memory construction and reliable evidence retrieval. VTEM comprises two core components: the View-Transition Memory Chain (TMC) and Transition-Guided Hierarchical Retrieval (THR). TMC derives spatiotemporal associations from sampled observations and uses them to assess memory value, retaining observations that contribute novel information in a hierarchical scene-view memory chain, thereby enabling efficient memory construction. THR decomposes user questions into hierarchical retrieval targets and association cues, then combines scene-view matching with transition-guided multi-hop reasoning to retrieve fine-grained, context-aligned evidence. Extensive experiments on OpenEQA, ECBench-MultiScene, and EnvQA demonstrate improvements of 7.89, 3.44, and 4.57 points over the strongest baselines, validating the effectiveness of VTEM for efficient and reliable long-horizon embodied reasoning.

1 INTRODUCTION

Embodied intelligence is central to building agents that can operate autonomously and reliably in the physical world, supporting applications ranging from household service to environmental inspection (Savva et al., 2019; Shridhar et al., 2020; Varghese et al., 2026). Real-world embodied tasks normally require agents to make decisions using observations accumulated during long-term interactions, making long-horizon embodied reasoning essential for practical deployment and a major focus of current research (Kim et al., 2024; Fan et al., 2025; Kim et al., 2026).

Current approaches pursue long-horizon embodied reasoning through either training-based adaptation or memory-based augmentation. Training-based methods adapt multimodal models with long-horizon embodied or spatial reasoning data, but require costly supervision and remain limited in processing long histories (Driess et al., 2023; Chen et al., 2024). Memory-based methods instead construct external memories from accumulated egocentric observations and retrieve question-relevant evidence to support subsequent reasoning (Fan et al., 2025; Ali et al., 2025; Zhang et al., 2026; Sohn et al., 2026), thereby reducing the need to process the full observation history directly.

Nevertheless, as shown in Fig. 1, these memory-based methods fail to efficiently establish spatiotemporal associations across observations, causing redundancy and fragmented memories, thereby degrading the efficiency and reliability of long-horizon reasoning. This limitation manifests in two aspects: **(1) Cross-view memory redundancy.** During memory construction, repeated observations of the same object or region across viewpoints are difficult to match against and merge into existing memory. As a result, duplicate and low-value entries accumulate, increasing memory and retrieval costs while obscuring useful evidence. **(2) Context-misaligned memory retrieval.** During memory retrieval, insufficient associations across observations make it difficult to identify memory fragments that match the intended purpose of the user question and its spatiotemporal context, such

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.

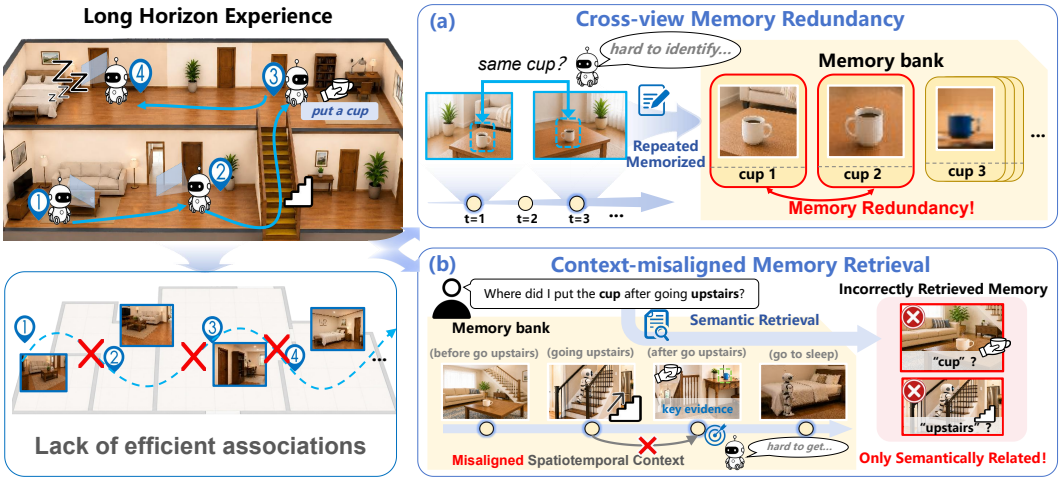


Figure 1: Motivation. Without view-transition associations, memory construction may repeatedly write redundant cross-view observations, while semantic retrieval can select evidence from a similar but incorrect spatial or temporal context.

as the location associated with an object or the time associated with an event. As a result, retrieval may favor semantically similar but context-misaligned evidence.

While some methods seek to model cross-view spatial associations through 3D reconstruction, this process incurs online computation and scales poorly to long egocentric histories (Sohn et al., 2026). In contrast, we observe that view transitions provide a more direct source of spatiotemporal cues: view changes relative to the surrounding scene reveal cross-view spatial relations, enabling efficient spatiotemporal associations among different observations (Jayaraman & Grauman, 2015).

Inspired by this observation, we propose the View-Transition Egocentric Memory framework (VTEM), which models view transitions to support efficient long-horizon memory construction and retrieval. VTEM comprises the View-Transition Memory Chain (TMC) module and the Transition-Guided Hierarchical Retrieval (THR) strategy. During memory construction, TMC derives spatiotemporal associations from view transitions across observations and performs association-guided value assessment to construct a high-value memory chain. Specifically, TMC compares sampled observations to infer changes in the view relative to surrounding scene elements, thereby characterizing cross-view spatial relations. Based on these associations, TMC evaluates each observation's memory value according to whether it contributes new information rather than repeating stored content, retaining high-value observations as scene and view-level memories within the chain for traceable coarse-to-fine retrieval. During memory retrieval, THR decomposes the user question into scene and view-level retrieval targets together with association cues, then hierarchically matches these targets against the scene-view memory to localize evidence. When hierarchical matching is insufficient, THR performs transition-guided multi-hop reasoning from the localized memories along the memory chain to retrieve context-aligned evidence from related observations.

The main contributions are as follows:

- We propose **V**iew-**T**ransition **E**gocentric **M**emory (VTEM), which exploits modeled view transitions for long-horizon memory construction and retrieval.
- We develop the *View-Transition Memory Chain* (TMC), which derives spatiotemporal associations from successive observations and performs association-guided value assessment to construct a traceable hierarchical chain of high-value memories.
- We introduce *Transition-Guided Hierarchical Retrieval* (THR), which decomposes questions for hierarchical memory matching and retrieves context-aligned evidence through transition-guided multi-hop reasoning.

- Experiments on OpenEQA, ECBench-MultiScene, and EnvQA show improvements of 7.89, 3.44, and 4.57 points over the strongest baselines, respectively.

2 Related Work

Our work is most closely related to long-horizon embodied question answering and memory-augmented agents that store and retrieve egocentric experience.

2.1 Long-Horizon Embodied Question Answering

Embodied question answering extends visual question answering (Antol et al., 2015) to interactive environments, requiring agents to observe and reason over embodied experience. Early work formulates EQA as navigation for acquiring visual information (Das et al., 2018). Recent benchmarks and methods further study open-ended, long-horizon, and multi-hop reasoning over embodied or egocentric observations (Majumdar et al., 2024; Chen et al., 2025; Kim et al., 2026), while other approaches improve information acquisition through global-local relevance modeling (Zhang et al., 2026). These studies mainly focus on task formulation, information acquisition, or answer generation. Our work instead investigates how view transitions can support the organization and retrieval of long-horizon egocentric observations.

2.2 Memory-Augmented Embodied Agents

Memory-augmented embodied agents store historical observations to support persistent scene understanding and long-horizon reasoning. Existing approaches construct 3D scene memory, object-centric memory, dynamic scene graphs, or 4D spatiotemporal representations (Yang et al., 2025; Fan et al., 2025; Ali et al., 2025; Sohn et al., 2026). Although these memories improve spatial grounding and dynamic scene modeling, they are mainly organized around scenes, objects, or temporal observations, without explicitly preserving associations induced by successive view transitions. As a result, repeated cross-view observations may accumulate, while similar memories from different contexts may be incorrectly matched. Our TMC uses view-transition associations to guide selective memory construction, while THR exploits hierarchical indices and retained transition relations for context-consistent retrieval.

3 Method

Existing memory-based methods lack explicit spatiotemporal associations across observations, causing cross-view memory redundancy and context-misaligned memory retrieval (Fan et al., 2025; Ali et al., 2025; Zhang et al., 2026; Sohn et al., 2026). To address these limitations, we propose the **V**iew-**T**ransition **E**gocentric **M**emory framework (VTEM), comprising the View-Transition Memory Chain (TMC) and Transition-Guided Hierarchical Retrieval (THR). TMC models the view transition-guided association to provide compact and traceable memory construction, whereas THR supports fine-grained, context-aligned memory retrieval for long-horizon embodied reasoning.

3.1 Preliminary

Following prior works (Das et al., 2018; Majumdar et al., 2024; Fan et al., 2025), we formulate the task as reasoning over long-horizon egocentric experience to answer user questions. During the experience phase, an agent obtains a long observation sequence $\mathbf{O}_{1:T} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, whose content varies across viewpoints and scenes. After the experience phase, the agent receives a user question $\mathbf{q}$ and reasons over the accumulated experience to generate an answer $\hat{\mathbf{y}}$. To reduce the substantial memory and inference costs of reasoning over the full observation sequence, existing memory-based methods construct an external memory $\mathcal{M}_{\text{ext}}$ and retrieve question-relevant evidence for answer generation (Fan et al., 2025; Ali et al., 2025; Zhang et al., 2026; Sohn et al., 2026):

$$\begin{aligned} \mathcal{M}_{\text{ext}} &= \mathcal{G}(\mathbf{O}_{1:T}), \\ \hat{\mathbf{y}} &= \mathcal{F}_{\text{g}}\left(\mathbf{q}, \mathcal{R}_{\text{r}}(\mathbf{q}, \mathcal{M}_{\text{ext}})\right). \end{aligned} \tag{1}$$

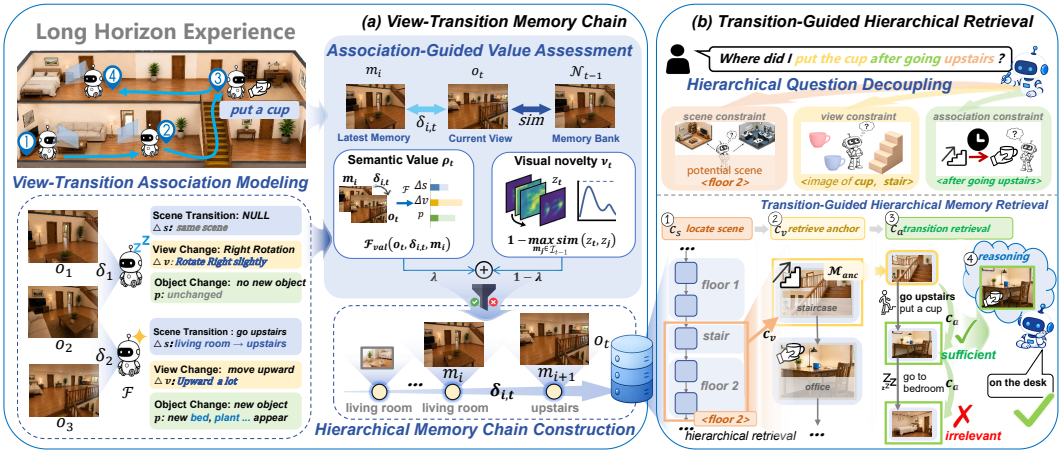


Figure 2: Overview of VTEM. TMC uses view-transition associations to retain high-value scene-view memories, and THR localizes evidence hierarchically before transition-guided expansion.

Here, $\mathcal{G}(\cdot)$, $\mathcal{R}_{\mathrm{r}}(\cdot)$, and $\mathcal{F}_{\mathrm{g}}(\cdot)$ denote memory construction, memory retrieval, and answer generation, respectively. This formulation highlights that answer reliability depends on both what information is retained during construction and which evidence is retrieved for reasoning.

However, existing memory designs overlook view-transition associations (Fan et al., 2025; Ali et al., 2025; Yang et al., 2025; Sohn et al., 2026), limiting value-aware memory construction and context-aligned retrieval; we therefore introduce VTEM to explicitly model these associations.

3.2 View-Transition Memory Chain

Cross-view repetitions often remain unresolved during memory construction, leading to the accumulation of redundant entries (Fan et al., 2025; Ali et al., 2025; Yang et al., 2025; Sohn et al., 2026). To address this issue, we introduce the View-Transition Memory Chain, which models view-transition associations to guide compact, traceable memory construction. TMC models transitions among sampled observations to establish spatiotemporal associations across views. These associations guide observation-value assessment, after which high-value observations are organized into a scene-view hierarchical memory chain for coarse-to-fine retrieval.

View-Transition Association Modeling

To establish spatiotemporal associations, TMC progressively aggregates local view changes between adjacent sampled observations instead of relying only on endpoint comparison.

Let $\mathcal{I}_{t-1}^{\mathrm{ret}}$ be the retained indices before time t, and let $i = \max \mathcal{I}_{t-1}^{\mathrm{ret}}$ denote the latest retained observation. TMC maintains an accumulated view-transition descriptor $\boldsymbol{\delta}_{i,t}$ from $\mathbf{o}_i$ to $\mathbf{o}_t$. This structured semantic descriptor records accumulated scene changes $\Delta s_{i,t}$, viewing-direction changes $\Delta v_{i,t}$, and cross-view object changes $\mathbf{p}_{i,t}$, rather than an explicit geometric state. TMC initializes $\boldsymbol{\delta}_{i,i} = \mathbf{0}$ and recursively updates it as

$$\boldsymbol{\delta}_{i,t} = (\Delta s_{i,t}, \Delta v_{i,t}, \mathbf{p}_{i,t}) = \mathcal{F}_{\mathrm{rel}}\left(\mathbf{o}_{t-1}, \mathbf{o}_t, \boldsymbol{\delta}_{i,t-1}\right). \tag{2}$$

This update provides the relational basis for subsequent association-guided value assessment.

Association-Guided Value Assessment

Reliable memory writing requires measuring the useful new information introduced by a candidate observation. TMC therefore combines association-guided semantic value ρ_t with visual novelty ν_t to assess its memory value from two complementary perspectives: relational-semantic information gain and visual information gain.

Specifically, for association-guided semantic value ρ_t, TMC uses $\mathbf{o}_t$, $\boldsymbol{\delta}_{i,t}$, and the latest retained node $\mathbf{m}_i$ to measure incremental scene, object, or event information relative to the associated memory. For visual novelty ν_t, TMC compares the candidate visual feature $\mathbf{v}_t$ with each retained feature $\mathbf{v}_j$:

$$\begin{aligned} \rho_t &= \mathcal{F}_{\text{val}}\left(\mathbf{o}_t, \boldsymbol{\delta}_{i,t}, \mathbf{m}_i\right), \\ \nu_t &= 1 - \max_{j \in \mathcal{I}_{t-1}^{\text{ret}}} \text{sim}\left(\mathbf{v}_t, \mathbf{v}_j\right). \end{aligned} \tag{3}$$

A lower maximum visual similarity indicates greater visual novelty. Both scores are normalized to $[0, 1]$ for scale alignment before aggregation.

A unified memory-write decision requires balancing these complementary information gains. TMC combines them using a weight $\lambda \in [0, 1]$ to obtain the overall storage value:

$$\alpha_t = \lambda \rho_t + (1 - \lambda)\nu_t. \tag{4}$$

A candidate is selected for memory construction when $\alpha_t \geq \tau$, where τ denotes the memory-write threshold. The first sampled observation directly initializes the memory, and each retained candidate resets the transition anchor as $i \leftarrow t$.

Hierarchical Memory Chain Construction

To preserve coarse-grained scene context together with fine-grained view evidence, each high-value candidate selected by value assessment is instantiated as a scene-view memory node. The node consists of its scene context s_t, the previously constructed multimodal view representation $\mathbf{z}_t$, and the original observation $\mathbf{o}_t$:

$$\mathbf{m}_t = \left(s_t, \mathbf{z}_t, \mathbf{o}_t\right). \tag{5}$$

Here, s_t denotes the scene context associated with $\mathbf{o}_t$ and updated from $\Delta s_{i,t}$; $\mathbf{z}_t$ represents the multimodal view content, and $\mathbf{o}_t$ retains the complete observation.

To organize high-value observations into a traceable chain, TMC connects each newly retained node to the latest retained node through the modeled view-transition association. Let $\mathcal{M}_t = (\mathcal{N}_t, \mathcal{E}_t)$ denote the memory chain, where $\mathcal{N}_t$ contains retained scene-view nodes and $\mathcal{E}_t$ contains transition edges. The edge $\mathbf{e}_{i,t} = (\mathbf{m}_i, \boldsymbol{\delta}_{i,t}, \mathbf{m}_t)$ records the transition from the latest retained node to the current node. The memory chain is updated as

$$\mathcal{M}_t = \begin{cases} \left(\mathcal{N}_{t-1} \cup \{\mathbf{m}_t\}, \mathcal{E}_{t-1} \cup \{\mathbf{e}_{i,t}\}\right), & \alpha_t \geq \tau, \\ \left(\mathcal{N}_{t-1}, \mathcal{E}_{t-1}\right), & \text{otherwise.} \end{cases} \tag{6}$$

When the write condition is satisfied, the current memory node and its transition edge are jointly appended to the memory chain; otherwise, the chain remains unchanged.

To support subsequent coarse-to-fine localization and association-based multi-hop retrieval, the memory chain jointly encodes hierarchical content and view-transition relations. Scene contexts organize retained nodes at the scene level, multimodal view representations support fine-grained localization, and transition edges preserve the spatiotemporal evolution between consecutive memory nodes. The resulting structure forms a compact and traceable scene-view memory chain.

3.3 Transition-Guided Hierarchical Retrieval

During memory retrieval, existing semantic-matching methods often overlook spatiotemporal associations in memory and retrieve evidence that appears semantically relevant but is misaligned with the required spatiotemporal context (Fan et al., 2025; Zhang et al., 2026; Chen et al., 2025; Sohn et al., 2026). Therefore, we introduce Transition-Guided Hierarchical Retrieval (THR), which performs hierarchical question decoupling to extract multi-granularity retrieval targets and association cues, and combines scene-view hierarchical matching with transition-guided reasoning. By separating content localization from relation reasoning, THR precisely localizes fine-grained evidence and recovers related observations that satisfy the required spatiotemporal context.

Hierarchical Question Decoupling

A raw question may contain multi-granularity retrieval constraints involving scenes, objects, events, and their relations. THR maps the user question $\mathbf{q}$ to a structured constraint tuple $\mathbf{c} = (c_s, c_v, c_a)$.

The components c_s, c_v, and c_a denote the scene-level retrieval target, the view-level retrieval target, and the association cue, respectively. The first two specify the relevant scene scope and the required fine-grained content, while c_a captures the cross-view spatiotemporal relation expected between supporting observations, such as the temporal cue "before going upstairs":

$$\mathbf{c} = (c_s, c_v, c_a) = \mathcal{F}_{\mathrm{d}}(\mathbf{q}). \tag{7}$$

The decoupling function $\mathcal{F}_{\mathrm{d}}(\cdot)$ is implemented using an LVLM with a structured prompt. If a retrieval constraint is not expressed in the question, its corresponding component is set to $\emptyset$ and omitted from the associated retrieval operation.

This structured decomposition explicitly separates multi-granularity content constraints from cross-observation relation requirements. The scene- and view-level retrieval targets support scene-to-view hierarchical localization, whereas the association cue provides the relational condition for subsequent transition-guided reasoning.

Transition-Guided Hierarchical Memory Retrieval

Fine-grained evidence is localized within the scene scope specified by the question. Given the retained memory nodes $\mathcal{N}$, THR applies the scene-level and view-level retrieval targets to obtain an anchor memory set $\mathcal{M}_{\mathrm{anc}}$:

$$\mathcal{M}_{\mathrm{anc}} = \mathcal{R}_{\mathrm{hier}}\left(\mathcal{N}, c_s, c_v\right). \tag{8}$$

The hierarchical retrieval operator $\mathcal{R}_{\mathrm{hier}}(\cdot)$ first localizes the relevant scene and then matches the view-level memories within that scope. If a target at either level is unspecified, the corresponding matching stage is omitted. The resulting $\mathcal{M}_{\mathrm{anc}}$ provides anchor evidence with a focused scene scope and fine-grained target relevance, while serving as the starting set for cross-observation reasoning.

Recovering the complete spatiotemporal context associated with the target evidence may require complementary memories connected through view-transition relations. THR therefore evaluates the sufficiency of the anchor set $\mathcal{M}_{\mathrm{anc}}$ and, when additional evidence is needed, expands over the transition-edge set $\mathcal{E}$ under the view-level retrieval target and association cue. The final evidence $\mathcal{M}_q$ is defined as

$$\mathcal{M}_q = \begin{cases} \mathcal{M}_{\mathrm{anc}}, & \mathcal{F}_{\mathrm{suf}}(\mathbf{q}, \mathcal{M}_{\mathrm{anc}}) = 1, \\ \mathcal{R}_{\mathrm{trans}}(\mathcal{M}_{\mathrm{anc}}, \mathcal{E}, c_v, c_a), & \text{otherwise.} \end{cases} \tag{9}$$

The binary function $\mathcal{F}_{\mathrm{suf}}(\cdot)$ determines whether the anchor evidence is sufficient for answer generation. Otherwise, $\mathcal{R}_{\mathrm{trans}}(\cdot)$ performs further multi-hop reasoning over TMC edges, where each hop traverses one transition edge between consecutively retained memory nodes. The association cue c_a further constrains paths compatible with the question's spatiotemporal relation, while the view-level target preserves content relevance. The resulting $\mathcal{M}_q$ aggregates context-aligned evidence across the corresponding related memory nodes.

Overall, by integrating TMC and THR, VTEM addresses two challenges: cross-view memory redundancy and context-misaligned retrieval, yielding compact, traceable memory and fine-grained, context-aligned evidence for reliable long-horizon embodied reasoning.

4 Experiments

In this section, we conduct extensive experiments on three widely-adopted embodied question answering benchmarks to evaluate the effectiveness of VTEM.

4.1 Experimental Settings

Datasets and Metrics

OpenEQA (Majumdar et al., 2024) evaluates open-vocabulary question answering over prior embodied experience. We report LLM-Match scores on the HM3D (Ramakrishnan et al., 2021) and ScanNet (Dai et al., 2017) subsets and their overall average. For ECBench (Dang et al., 2025), we construct ECBench-MultiScene (ECBench-MS) by selecting samples containing at least three

Table 1: Main results on OpenEQA, EnvQA, and ECBench-MultiScene. OpenEQA uses LLM-Match, EnvQA uses role-value IoU-like scores, and ECBench-MS uses ECEval; higher is better for all metrics. Bold and underline denote the best and second-best results.

Method	OpenEQA			EnvQA				ECBench-MS
	HM3D	ScanNet	All	State	Order	Event	All	ECEval
General / Embodied LVLMs								
Qwen3-VL-8B	35.59	58.87	50.94	<u>22.06</u>	44.50	22.26	<u>29.61</u>	42.77
Qwen2.5-VL-7B	41.48	49.77	49.47	19.79	41.50	20.37	27.22	40.48
RoboBrain2.0-7B	47.89	54.84	52.47	19.99	41.00	<u>24.86</u>	28.61	38.78
Video-LLaVA-7B	41.50	34.60	39.20	17.28	38.50	20.88	25.89	36.77
Memory-based Methods								
Embodied VideoAgent	44.88	55.12	51.63	15.36	<u>46.00</u>	20.43	27.26	41.11
CLiViS	47.53	56.33	53.33	21.10	40.50	20.71	27.44	15.76
3D-Mem	47.71	<u>60.40</u>	<u>56.08</u>	20.59	44.50	14.23	26.44	42.72
GeLM	<u>48.20</u>	59.45	55.62	20.47	44.50	16.89	27.29	<u>43.48</u>
VTEM (Ours)	**62.61**	**64.67**	**63.97**	**26.27**	**50.00**	**26.26**	**34.18**	**46.92**

scenes, resulting in 732 samples, and report the official ECEval score. EnvQA (Gao et al., 2021) evaluates object states, temporal orders, and event occurrences; we report State, Order, Event, and overall role-value IoU-like scores.

BASELINES

We compare VTEM with general or embodied LVLMs, including Qwen3-VL-8B, Qwen2.5-VL-7B, RoboBrain2.0-7B (BAAI RoboBrain Team et al., 2025), and Video-LLaVA-7B (Lin et al., 2024), as well as memory-based methods, including Embodied VideoAgent (Fan et al., 2025), CLiViS (Li et al., 2026), 3D-Mem (Yang et al., 2025), and GeLM (Chen et al., 2025).

IMPLEMENTATION DETAILS

VTEM uses TMC to select high-value observations and organize them as a scene-view memory chain, then uses THR for hierarchical localization and transition-guided evidence expansion. All methods follow the same evaluation samples and protocols when applicable for each dataset. In ablations, TMC and THR denote the full View-Transition Memory Chain and Transition-Guided Hierarchical Retrieval modules. All scores are percentages; higher is better.

4.2 MAIN RESULTS

We conduct the main experiments on embodied question-answering benchmarks: OpenEQA, ECBench-MultiScene, and EnvQA. On these benchmarks, VTEM outperforms the strongest baselines by 7.89, 3.44, and 4.57 points. These datasets cover open-vocabulary indoor QA, multi-scene egocentric evidence selection, and structured state/event understanding, so the consistent gains indicate that VTEM is not tied to a single answer format.

OpenEQA. As shown in Table 1, VTEM achieves the best overall OpenEQA score of 63.97, outperforming the strongest baseline, 3D-Mem, by 7.89 points. It also exceeds the strongest competing results on HM3D and ScanNet by 14.41 and 4.27 points, respectively. The consistent improvements across both environments indicate that explicitly preserving view-transition associations helps organize long-horizon observations into more informative evidence, while hierarchical localization and conditional transition-guided expansion improve the retrieval of memories aligned with the question context. The larger margin on HM3D also suggests that transition-aware memory is useful for recovering sparse question-relevant evidence from long egocentric histories. This pattern is important because many OpenEQA queries depend on the recovery of the correct view context, rather than only the most similar object mention.

ECBench-MultiScene. Table 1 shows that VTEM achieves the best ECEval score of 46.92 on ECBench-MultiScene, exceeding GeLM by 3.44 points. ECBench-MultiScene contains observations distributed across multiple scene contexts, where semantic similarity alone can retrieve evidence from an incorrect location or scene. The improvement therefore supports the effectiveness of scene-view hierarchical localization and transition-guided reasoning in reducing interference from semantically related but context-misaligned memories. Because each evaluated sample involves evidence from at least three scenes, the result matches VTEM's overall design of selecting a scene anchor before expanding along related view transitions. Thus, the improvement reflects stronger cross-scene grounding under settings where visually similar memories can otherwise interfere with answer generation in complex long-horizon embodied environments.

EnvQA. As reported in Table 1, VTEM achieves the best overall EnvQA score of 34.18, outperforming Qwen3-VL-8B by 4.57 points. It also ranks first on State, Order, and Event questions, with scores of 26.27, 50.00, and 26.26, respectively. These question types require the model to distinguish observations acquired at different times and viewpoints and to connect complementary evidence across the interaction history. The results indicate that the traceable transition chain preserves relevant state and event changes, while transition-guided retrieval supports reasoning over their temporal and contextual relations. This structured metric further shows that the same transition relations help beyond free-form LLM-Match evaluation. The gains on State and Event questions also indicate that retained transitions preserve changing evidence, not only static object cues.

4.3 Ablation Study

Table 2: HM3D ablation (LLM-Match %).

TMC	THR	LLM-Match (%)
–	–	48.52
√	–	50.81
–	√	54.40
√	√	**62.61**

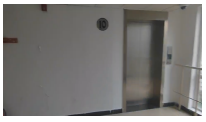
Table 3: Real-video QA (LLM-Match %).

Method	LLM-Match (%)
RoboBrain2.0-7B	<u>70.00</u>
Embodied VideoAgent	61.25
3D-Mem	67.50
GeLM	67.50
VTEM (Ours)	**81.25**

We conduct an ablation study on the widely-used dataset HM3D of OpenEQA to evaluate the roles of TMC and THR. The baseline retains sampled observations and performs flat global retrieval. As shown in Table 2, adding either module consistently improves performance: TMC yields a 2.29-point gain by selectively constructing a view-transition memory chain that removes redundant observations while preserving memories with incremental scene, object, or event information; THR provides a larger 5.88-point improvement by first localizing relevant scene-view anchors and then expanding along transition relations only when additional context is needed. Combining TMC and THR achieves the best result, improving the baseline by 14.09 points. These results demonstrate that the two modules are complementary: TMC constructs hierarchical indices and transition relations, while THR explicitly exploits this structure to retrieve context-consistent and complementary memories, together enabling more accurate and efficient retrieval.

4.4 Results on Real-World Videos

To examine transfer beyond benchmarks, we evaluate VTEM on 20 manually paired QA samples from real robot trajectory videos. As shown in Table 3, VTEM reaches 81.25 LLM-Match, outperforming the strongest baseline, RoboBrain2.0-7B, by 11.25 points. The recordings contain viewpoint changes, visually similar corridors, spatial relations around signs, and fine-grained object states, where question-relevant evidence is sparse and distributed. These results indicate that TMC and THR remain effective on real embodied observations; Fig. 3 shows representative Video 2 cases.



Q: What is the wall number beside the elevator?
GT / VTEM: 10 / 10.



Q: Where are the shower toiletries stored?
GT / VTEM: black wall shelf / black shelf.

Figure 3: Real-world Video 2 examples.

Table 4: TMC object coverage and recurrence.

Method	Cov. ↑	AvgRec ↓	MaxRec ↓
Dense (all 120)	79.34%	29.32	70.30
Random (1/4)	72.62%	7.94	18.02
Uniform (1/4)	74.44%	7.77	17.50
Uniform (1/10)	70.93%	3.37	6.80
TMC (Ours)	78.33%	7.02	16.30

4.5 Memory Redundancy Analysis

We assess TMC on 10 representative ScanNet videos by uniformly sampling 120 frames per video, manually annotating salient visible objects, and using a fixed VLM to count their occurrences in retained frames. Coverage is the fraction of annotated objects observed, while AvgRec and MaxRec are the mean and maximum retained-frame occurrences per annotated object. As shown in Table 4, TMC retains 78.33% coverage, only 1.01 points below Dense Sampling, while reducing AvgRec and MaxRec by 76.1% and 76.8%. At the same 1/4 retention ratio, TMC improves coverage and recurrence over random and uniform sampling; Uniform Sampling at 1/10 has lower recurrence but loses substantial object information.

4.6 Hyperparameter Analysis

We analyze the balancing coefficient λ and memory-write threshold τ in $\alpha_t = \lambda\rho_t + (1-\lambda)\nu_t$, where ρ_t is association-aware semantic value and ν_t is CLIP-based (Radford et al., 2021) visual novelty. On a stratified 50-sample OpenEQA ScanNet subset covering all seven question categories and 50 trajectories, we vary λ with $\tau = 0.45$ and vary τ with $\lambda = 0.50$. Figure 4(a) shows that visual novelty alone is insufficient: $\lambda = 0.00$ yields 40.00, while combining semantic value and visual novelty reaches 50.00 for $\lambda \in [0.50, 0.90]$. This supports the descriptor's role in separating meaningful changes from viewpoint-induced appearance variation. Figure 4(b) shows that $\tau = 0.45$ and 0.55 perform best; we use $\lambda = 0.50$ and $\tau = 0.45$ by default.

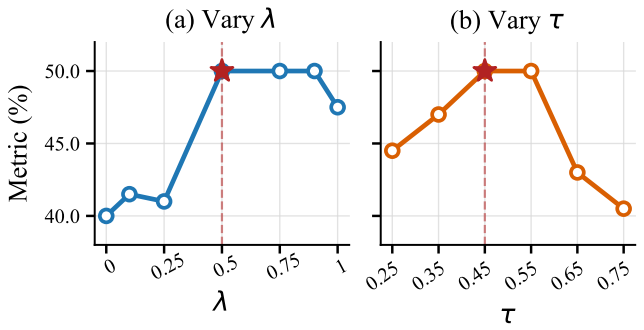


Figure 4: Hyperparameter analysis.

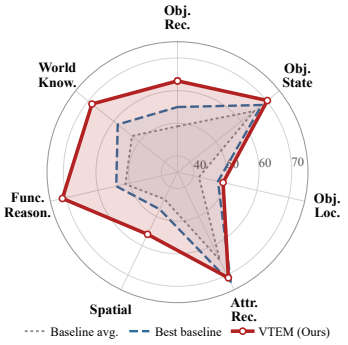


Figure 5: Question-type analysis.

4.7 Question-Type Analysis on OpenEQA

We evaluate VTEM across seven OpenEQA question categories. As shown in Figure 5, VTEM outperforms the strongest baseline in six categories and exceeds the average of all non-VTEM baselines in every category. Gains are largest for functional reasoning and spatial understanding, with improvements of 16.94 and 8.53 points, where THR integrates evidence across scenes and views. VTEM also improves object localization and object-state recognition through traceable scene-view chains, while attribute recognition remains limited by local visual discrimination.

5 Conclusion

This work presents VTEM, a unified memory framework for reliable long-horizon embodied reasoning over egocentric experience. It combines TMC, which constructs compact and traceable memories through view-transition modeling and value assessment, with THR, which performs hierarchical localization and conditional multi-hop retrieval over transition relations. By jointly optimizing memory construction and retrieval, VTEM reduces cross-view redundancy and context-misaligned retrieval. Experiments on OpenEQA, ECBench and EnvQA show consistent improvements over memory-based methods, validating view transitions as an effective signal for building compact and traceable memory systems for embodied agents.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, LaTeX formatting, and internal review of presentation and clarity. The authors reviewed all AI-assisted outputs and take responsibility for the final text, claims, experiments, and artifacts.

ETHICS STATEMENT

The real-world robot-video QA samples are used only for evaluating embodied question answering. The evaluation does not require reasoning about personal identifying attributes, and any release of videos or annotations should be anonymized and checked for privacy-sensitive content.

REPRODUCIBILITY STATEMENT

We report the task formulation, datasets, metrics, baselines, implementation details, and ablation settings in the main paper. Additional implementation configurations, structured prompts, memory-construction and retrieval pseudocode, data preprocessing details, evaluation prompts, and baseline reproduction settings will be provided in the appendix or supplementary material.

REFERENCES

Muhammad Qasim Ali, Saeejith Nair, Alexander Wong, Yuchen Cui, and Yuhao Chen. GraphPad: Inference-time 3d scene graph updates for embodied question answering. *arXiv preprint arXiv:2506.01174*, 2025.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

BAAI RoboBrain Team, Wenhai Wang, Yongqiang Zhang, Tianming Xie, Quanxin Deng, Ruiyuan Guo, Peng Cai, Lingzhi Zhang, Jiajun Tan, Tao Lin, Qingyun Li, Da Lu, Peihao Zhou, Zhipeng Luo, Tianyu Huang, Zhongyue Xu, Yuhui Lin, Qi Dai, Hanwei Chen, Jiazhao Wang, Changan Cai, Ji Shen, Ruowei Gao, Ya Wang, Hao Su, Jie Tang, Song-Chun Zhu, Xizhou Zhu, Jiaqi Guo, Qi Tian, and Liqiang Zhang. RoboBrain 2.0 technical report, 2025. URL https://arxiv.org/abs/2511.08295.

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

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.

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 (CVPR)*, pp. 5828–5839, 2017.

Thao Minh Dang, Thang Viet Nguyen, Duy Phung, Richard Baraniuk, and Toan Nguyen. ECBench: Can multi-modal foundation models understand the egocentric world?, 2025. URL https://arxiv.org/abs/2505.14890.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, June 2018.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-e: An embodied

multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied VideoAgent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6342–6352, October 2025.

Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-QA: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1675–1685, 2021.

Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1413–1421, December 2015.

Kangsan Kim, Yanlai Yang, Suji Kim, Woongyeong Yeo, Youngwan Lee, Mengye Ren, and Sung Ju Hwang. Ma-egoqa: Question answering over egocentric videos from multiple embodied agents. *arXiv preprint arXiv:2603.09827*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Wenhui Li, Shiyu Xiong, Yanwei Wang, Yue Yang, Jian Wang, Zongqing Li, and Yizhou Zheng. CLiViS: Unleashing cognitive map for lifelong vision-and-language navigation, 2026. URL `https://arxiv.org/abs/2605.18409`.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2024.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

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

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Tin Stribor Sohn, Maximilian Dillitzer, Jason J. Corso, and Eric Sax. R4: Retrieval-augmented reasoning for vision-language models in 4d spatio-temporal space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38711–38721, June 2026.

Subin Varghese, Joshua Gao, Asad Ur Rahman, and Vedhus Hoskere. Bridgeeqa: Virtual embodied agents for real bridge inspections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8163–8173, June 2026.

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*, pp. 17294–17303, 2025.

Haochen Zhang, Nirav Savaliya, Faizan Siddiqui, and Enna Sachdeva. FAST-EQA: Efficient embodied question answering with global and local region relevancy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1664–1673, March 2026.

A Supplementary Material

This appendix provides the reproducibility details and controlled evidence for View-Transition Egocentric Memory (VTEM). The goal of every appendix experiment is to test a specific claim about the method: whether transition-aware memory construction stores more useful evidence, whether transition-guided retrieval uses edge semantics rather than temporal adjacency alone, and whether the final answer gain remains under fixed answer backbones and evidence budgets. Unless otherwise stated, all OpenEQA appendix experiments use the same Qwen2.5-VL-7B-Instruct answer generator, the same CLIP ViT-L/14-336px feature backbone, the same GPT-4-1106-preview automatic evaluator, and the same 50-question subset selected from four OpenEQA trajectories. The organization below follows three reviewer-facing questions: whether view-transition relations are necessary, whether the memory is genuinely more compact, and whether the gains remain under controlled answer and evidence budgets.

A.1 Implementation and Reproducibility Details

VTEM separates preprocessing from the two method modules. The CLIP gate first filters incoming frames. The first VTEM module, Transition-aware Memory Construction (TMC), creates a temporary association for each accepted candidate view, uses that association to adjust the view value, and writes the view into memory only when it is high value. The second module, Transition-guided Hierarchical Retrieval (THR), retrieves anchor memories and then expands along retained transition edges only when the anchor evidence is insufficient. Low-level sampling and gate constants are fixed across all relevant variants and are reported only for reproducibility; the validation experiments below change transition structure, memory construction, or retrieval logic while holding these constants fixed.

Transition descriptor schema. For a candidate view $\mathbf{o}_t$ and the latest retained memory $\mathbf{m}_i$, TMC builds a temporary descriptor $\delta_{i,t}$. The descriptor has the following persisted fields only if $\mathbf{o}_t$ is written:


```
{
  "source_node_id": "m_000001 or null",
  "source_floor_id": "floor_1 or null",
  "source_scene_id": "living_room or null",
  "source_frame_id": "5 or null",
  "target_floor_id": "floor_1",
  "target_scene_id": "kitchen",
  "description": "entering kitchen from living_room on floor_1",
  "scene_changed": true,
  "floor_changed": false,
  "viewpoint_shift": true,
  "new_detail": true,
  "repeated_view": false,
  "image_similarity": 0.83,
```

Table 5: Implementation-level definition of VTEM operators.

Operator	Input	Implementation	Output
CLIP gate	Sequential RGB frames	Fixed CLIP ViT-L/14-336px prefilter with a periodic guard. This is a preprocessing step, is shared by all VTEM variants, and is not counted as a VTEM module.	Candidate views passed to memory construction.
F_{scene}	Candidate view, scene label set, compact trajectory history	Qwen2.5-VL-7B-Instruct scene summarizer with JSON-only output, temperature 0.1, and 256 max new tokens.	Scene label, confidence, and one compact view description.
F_{rel}	Latest retained memory and current candidate view	Temporary transition association. It compares floor, scene, CLIP image similarity, CLIP text similarity, content-word overlap, relation words, and optional movement hints. The association is cleared if the candidate is not written.	$\delta_{i,t}$ with source/target ids, scene/floor flags, relation text, novelty flags, similarity values, and a bounded semantic adjustment.
F_{val}	Candidate view summary, visual novelty, and $\delta_{i,t}$	Association-aware value assessment. The base semantic value is calibrated from LVLM confidence and description density, then shifted by the transition adjustment and clipped to $[0, 1]$.	Storage score $\alpha_t = \lambda\rho_t + (1-\lambda)\nu_t$ and a write/drop decision.
Visual novelty	Current CLIP image/text features and retained memory features	Cosine-distance novelty against recent retained nodes, using normalized CLIP features.	$\nu_t \in [0, 1]$.
F_d	User question	Qwen2.5-VL text decomposition. It extracts visual target text and a transition or association constraint; invalid JSON falls back to the raw question and an empty association constraint.	Scene, visual, and association constraints (c_s, c_v, c_a).
R_{hier}	Memory nodes and (c_s, c_v)	Two-stage anchor localization: scene filtering followed by CLIP text/image/hybrid matching.	Anchor memory set $\mathcal{M}_{anc}$.
F_{suf}	Question and anchor evidence	Qwen2.5-VL sufficiency judge over the top retrieved images. Answers such as unknown or cannot determine are treated as insufficient.	Binary sufficiency decision.
R_{trans}	Anchors, retained edges, c_a, and hop limit H	Bidirectional graph expansion over retained transition edges. Neighbor scoring combines edge text similarity and a selector prompt. The search stops when evidence is sufficient, no neighbor is selected, the boundary is reached, or H is reached.	Expanded evidence set $\mathcal{M}_q$.
F_g	Question, selected memory images, descriptions, and movement context	Qwen2.5-VL final answer generation with XML answer output and 512 max new tokens.	Final answer string.

```
  "text_similarity": 0.43,
  "semantic_adjustment": 0.18
}
```

The descriptor is temporary during value assessment. If the view is discarded, the descriptor is not stored. If the view is retained, the descriptor is stored only as metadata on the edge from $\mathbf{m}_i$ to $\mathbf{m}_t$.

Value normalization. The semantic value is normalized to the same range as CLIP novelty before storage. Let $\rho_t^{base} \in [0,1]$ be the calibrated value derived from LVLM confidence and description density, and let a_t be the bounded association adjustment. TMC computes

$$\rho_t = \mathrm{clip}_{[0,1]}\left(\rho_t^{base} + w_{\mathrm{assoc}} a_t\right), \qquad \alpha_t = \lambda \rho_t + (1-\lambda)\nu_t.$$

Scene or floor changes, changed egocentric viewpoints, and new visible details increase a_t; repeated views decrease it. Thus a same-looking object observed after a meaningful viewpoint or scene transition can still be retained, while a near-duplicate view can be suppressed.

Algorithm 1 Transition-aware Memory Construction (TMC)

Require: observation sequence $\mathbf{O}_{1:T}$, CLIP gate G, threshold τ, balance λ
Initialize memory nodes $\mathcal{N} \leftarrow \emptyset$ and transition edges $\mathcal{E} \leftarrow \emptyset$
for each observation $\mathbf{o}_t$ **do**
  if $G(\mathbf{o}_t)$ rejects the frame **then**
    continue
  end if
  Summarize the view as scene s_t, confidence, and text $\mathbf{r}_t$
  Build candidate node $\mathbf{m}_t = (s_t, \boldsymbol{\xi}_t, \mathbf{r}_t, \mathbf{o}_t)$
  Build temporary association $\delta_{i,t}$ from the latest retained node $\mathbf{m}_i$ to $\mathbf{m}_t$
  Compute ρ_t from the view value and $\delta_{i,t}$
  Compute CLIP novelty $\nu_t = 1 - \max_{\mathbf{m}_j \in \mathcal{N}} \mathrm{sim}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_j)$
  Compute $\alpha_t = \lambda \rho_t + (1-\lambda)\nu_t$
  if $\mathcal{N}$ is empty, or the scene/floor changes, or $\alpha_t \geq \tau$ **then**
    Add $\mathbf{m}_t$ to $\mathcal{N}$
    Add edge $(\mathbf{m}_i, \delta_{i,t}, \mathbf{m}_t)$ to $\mathcal{E}$ if $\mathbf{m}_i$ exists
  else
    Discard $\mathbf{m}_t$ and clear $\delta_{i,t}$
  end if
end for
return $\mathcal{M} = (\mathcal{N}, \mathcal{E})$

A.2 Controlled and Budget-Matched Comparison Protocol

The appendix subset contains 50 question-answer pairs from four OpenEQA episodes. It was selected to reduce repeated memory construction while retaining both HM3D and ScanNet coverage:

- `hm3d-v0/014-hm3d-rsggHU7g7dh`: 14 questions.
- `hm3d-v0/019-hm3d-y9hTuugGdiq`: 14 questions.
- `scannet-v0/048-scannet-scene0745_00`: 14 questions.
- `scannet-v0/100-scannet-scene0598_00`: 8 questions.

The category distribution is object recognition (8), attribute recognition (8), spatial understanding (8), object state recognition (8), functional reasoning (6), world knowledge (6), and object localization (6). The split metadata is released with the supplementary materials so that all ablations use the same question and episode IDs.

All variants use the same answer model, evaluator, image inputs, CLIP feature backbone, decoding settings, and OpenEQA LLM-match scoring protocol. No variant uses training, depth, pose, or sensor metadata. For retrieval ablations, the constructed memory is reused whenever memory construction is not the ablated target. For construction ablations, THR retrieval settings are held fixed.

Algorithm 2 Transition-guided Hierarchical Retrieval (THR)

Require: question $\mathbf{q}$, memory $\mathcal{M} = (\mathcal{N}, \mathcal{E})$, hop limit H

Decompose $\mathbf{q}$ into scene constraint c_s, visual constraint c_v, and association constraint c_a
Retrieve anchors $\mathcal{M}_{anc}$ with scene filtering followed by view-level text/image matching
if $F_{\text{suf}}(\mathbf{q}, \mathcal{M}_{anc})$ is sufficient **then**
    return reranked $\mathcal{M}_{anc}$
end if
Initialize frontier with anchor ids and mark them visited
for $h = 1$ to H **do**
    Collect unvisited neighbors connected to the frontier by retained transition edges
    Score neighbors using c_a, edge descriptors, and view descriptions
    Select up to K complementary neighbors with the selector LVLM
    Merge selected neighbors into the evidence set and deduplicate by node id
    if no neighbor is selected, boundary is reached, or the sufficiency judge accepts the evidence **then**
        break
    end if
end for
return reranked evidence truncated to the answer budget

Table 6: Controlled comparison protocol for the 50-question OpenEQA appendix subset.

Comparison row	Fixed resources	Changed component	Evidence budget
Full VTEM	Qwen2.5-VL-7B-Instruct generator, CLIP ViT-L/14-336px retrieval features, GPT-4-1106-preview evaluator, same subset and images.	None. TMC and THR are both enabled.	At most 12 final memory candidates per question.
Uniform sampling + flat retrieval	Same answer generator, evaluator, CLIP model, frame stride, and final evidence budget.	Disables CLIP gate, TMC value storage, hierarchical retrieval, and transition expansion.	12 candidates per question.
Flat retrieval	Same constructed VTEM memory and same answer generator.	Replaces THR with global flat retrieval.	12 candidates per question.
No descriptor / temporal / shuffled	Same retained nodes and answer budget whenever construction is not the ablated target.	Removes, degrades, or randomizes transition semantics while preserving comparable retrieval budget.	8.46 to 10.38 retrieved candidates per question, all below the 12-candidate cap.
Submodule ablations	Same dataset, evaluator, and decoding settings.	Enables only one named submodule within TMC or THR while keeping the rest of the pipeline fixed.	Same answer cap and same GPT evaluation.

Table 7: Appendix evidence chain: each experiment is tied to a specific claim.

Evidence block	Intended claim	Main observation
Outcome quality and win/loss	VTEM improves answer quality without sending more final evidence to the generator.	Full VTEM reaches 0.570, improves over flat retrieval by 0.060 and uniform+flat by 0.075, and retrieves 8.38 memories/question instead of 12.00.
Storage compression and retrieval cost	The memory gain is not only an accuracy gain; VTEM stores and retrieves a smaller evidence set than dense or flat controls.	VTEM stores 86 memories from 120 sampled candidate views, a 1.40× reduction, while retrieving 8.38 memories/question.
Transition causal controls	The gain comes from transition semantics rather than ordinary temporal adjacency or graph connectivity.	Shuffling descriptors drops to 0.480 and temporal-only expansion drops to 0.525, both below Full VTEM.
TMC ablation	Temporary association and value assessment are complementary in memory construction.	Removing either branch reduces the metric to 0.530–0.535.
TMC qualitative cases	Value assessment is not a pure visual novelty detector.	Repeated view changes can be discarded, while small but semantic state changes and scene transitions are retained.
THR ablation	Hierarchical anchoring and association-cued expansion are both needed.	Anchor-only improves over flat retrieval, while removing association cues drops to 0.510 despite retrieving 10.38 memories/question.
Mechanism-targeted split	The advantage is concentrated on questions where transition and cross-view evidence should matter.	On spatial/localization questions, Full VTEM reaches 0.554 versus 0.357 for flat retrieval and 0.304 for shuffled transition.
Descriptor audit and corruption	Retained edges are structurally well formed and their semantics affect downstream answers.	All 82 retained edge descriptors pass automatic consistency checks; corrupting descriptor text reduces QA performance.

A.3 Outcome Quality, Memory Efficiency, and Inference-Cost Analysis

Table 8: Outcome quality beyond mean score on the final 50-question runs.

Variant	Metric	Δ	≥4	=5	=1	Ret./Q
Full VTEM	0.570	0.000	27/50	21/50	15/50	8.38
Uniform + flat	0.495	-0.075	23/50	18/50	19/50	12.00
Flat retrieval	0.510	-0.060	23/50	18/50	18/50	12.00
Temporal chain	0.525	-0.045	24/50	20/50	17/50	10.38
Shuffled transition	0.480	-0.090	23/50	18/50	21/50	8.46

Table 8 reports the same final scores in a form that exposes the direction of the effect. Full VTEM has the highest normalized score and the highest number of high-quality answers, while retrieving fewer memories per question than both flat baselines. The strongest transition corruption control is shuffled transition: it keeps the same retained nodes and a nearly identical retrieval size, but it increases score-1 failures from 15 to 21 and reduces the metric from 0.570 to 0.480. This is the clearest evidence that the edge text is not just decorative context.

Table 9: Memory compactness and retrieval cost on the 50-sample subset.

Method	Mean	Metric	Ret./Ep	Edge/Ep	Ret./Q
Uniform + flat	2.980	0.495	24.000	23.000	12.000
Flat retrieval	3.040	0.510	21.500	20.500	12.000
Full VTEM	3.280	0.570	21.500	20.500	8.380

Full VTEM retains 21.5 memory nodes per episode on average, fewer than the uniform storage baseline's 24.0 nodes, while improving the normalized score from 0.495 to 0.570. It also retrieves fewer final evidence memories per question than flat retrieval, 8.38 versus 12.00, so the final answer generator receives a smaller evidence set.

Table 10: Storage compression under the controlled 50-question protocol. Counts are aggregated across the four selected OpenEQA trajectories.

Method	Sampled views	Processed views	Stored memories	Compression
Dense sampled memory	120	120	120	1.00×
Uniform + flat	120	120	96	1.25×
Full VTEM	120	96	86	1.40×

Table 10 reports the actual memory footprint in the controlled protocol. The dense row stores every sampled candidate view, whereas uniform sampling stores a fixed subset and VTEM first applies the CLIP gate and then keeps only high-value observations. VTEM stores 86 memories from 120 sampled views, giving a 1.40× reduction while preserving the best answer score. These counts should be read as controlled storage costs for the appendix subset; the full raw video frame counts vary by episode and are not used as the unit of comparison.

Table 11: Logged construction and question-time LVLM calls on the final 50-question runs.

Variant	Construct	Query/Q	Ret./Q	Metric
Uniform + flat	120	1.00	12.00	0.495
Flat retrieval	96	1.00	12.00	0.510
Full VTEM	96	4.68	8.38	0.570
Temporal chain	96	5.82	10.38	0.525
Shuffled transition	96	4.82	8.46	0.480

The logged call table separates one-time construction calls from question-time calls. Retrieval ablations intentionally reuse the same constructed memory, so their final logs do not double count the construction pass. The construction logs show 96 scene-summary calls for the four selected episodes

Table 12: Question-time retrieval efficiency on the 50-question subset. Query/Q reports logged LVLM calls per question and serves as a hardware-independent cost proxy.

Variant	Metric	Ret./Q	Hop/Q	Query/Q
Flat retrieval	0.510	12.00	0.00	1.00
Temporal chain	0.525	10.38	1.28	5.82
Shuffled transition	0.480	8.46	0.74	4.82
Full VTEM	0.570	8.38	0.72	4.68

under the CLIP-gated VTEM memory and 120 calls for the uniform storage baseline. Full VTEM uses more question-time reasoning calls than flat retrieval, but it sends a smaller final evidence set and outperforms temporal and shuffled controls that use comparable or larger question-time budgets. Therefore the controlled claim is not that every component is cheaper than flat retrieval; it is that VTEM uses the additional structured retrieval steps to select better evidence under the same final answer generator and evidence cap.

Table 12 further shows that THR does not simply increase evidence size. Full VTEM retrieves fewer memories than flat retrieval and also fewer memories than the temporal-chain control, while achieving the highest score. We report logged model calls rather than wall-clock latency or exact token counts because these quantities depend on deployment hardware, batching, and API routing; the call and evidence counts provide the comparable cost proxy available for every ablation.

A.4 Transition Structure Validation and Component Ablations

Table 13: View-transition causal ablation on the 50-sample OpenEQA subset, including no-descriptor, temporal-chain, and random-relation controls.

Variant	Mean	Metric	Ret./Ep	Ret./Q	Hop
Full VTEM	3.280	0.570	21.500	8.380	0.720
No transition descriptor	3.140	0.535	14.000	7.640	0.640
Temporal chain	3.100	0.525	21.500	10.380	1.280
Shuffled transition	2.920	0.480	21.500	8.460	0.740

Table 14: Submodule ablation on the same 50-question subset.

Variant	Assoc.	Value	Anchor	Hop	Mean	Metric
Full VTEM	yes	yes	yes	yes	3.280	0.570
TMC association only	yes	no	yes	yes	3.120	0.530
TMC value only	no	yes	yes	yes	3.140	0.535
THR anchor only	yes	yes	yes	no	3.200	0.550
THR multi-hop only	yes	yes	no	yes	3.200	0.550

Table 15: Fine-grained THR ablation with the same memory construction.

Variant	Mean	Metric	Ret./Q	Expand	Hop
Flat retrieval	3.040	0.510	12.000	0.000	0.000
Hierarchical anchor only	3.200	0.550	8.120	0.000	0.000
Always expand	3.240	0.560	8.540	0.120	1.120
No association cue	3.040	0.510	10.380	0.640	1.280
Full VTEM	3.280	0.570	8.380	0.080	0.720

Note: the no-association-cue row withholds association and transition wording during neighbor selection and answer generation. It expands more often but drops to 0.510, indicating that uncontrolled graph expansion is not a substitute for association-cued evidence selection.

The transition causal ablation isolates whether view-transition semantics contribute beyond ordinary temporal adjacency. Full VTEM reaches 0.570 normalized score. Removing the transition descrip-

Table 16: Paired Full VTEM versus ablation comparison on the same 50 questions.

Ablation	Δ metric	W/T/L	Big wins	$\geq$4 net
Shuffled transition	+0.090	11/35/4	7	+4
Uniform + flat	+0.075	9/39/2	5	+4
Flat retrieval	+0.060	8/37/5	6	+4
Temporal chain	+0.045	8/37/5	3	+3
TMC assoc. only	+0.040	5/42/3	3	+2
TMC value only	+0.035	7/39/4	4	+1
THR anchor only	+0.020	4/43/3	1	+1
Always expand	+0.010	3/44/3	1	+1
No assoc. cue	+0.060	9/35/6	7	+4

Table 17: Mechanism-targeted subset analysis. Metrics are normalized OpenEQA scores.

Variant	All 50	Transition-sensitive	Spatial/localization	Recognition/attribute
Full VTEM	0.570	0.562	0.554	0.589
Uniform + flat	0.495	0.451	0.375	0.607
Flat retrieval	0.510	0.465	0.357	0.625
Temporal chain	0.525	0.472	0.446	0.661
Shuffled transition	0.480	0.424	0.304	0.625
No association cue	0.510	0.417	0.304	0.750
THR anchor only	0.550	0.528	0.464	0.607

tor gives 0.535, using only temporal-chain edges gives 0.525, and shuffling the edge descriptors gives 0.480. The no-descriptor row is the construction-side control that disables transition descriptors during memory building. The shuffled row is the random-relation control: it keeps nodes and endpoints fixed but corrupts which relation is attached to which edge. The no-association-cue row in Table 15 provides the retrieval-side fixed-memory no-edge control: the same constructed memory is used, but association and transition wording are withheld during neighbor selection and answer generation. Because the shuffled setting keeps the same retained nodes and endpoints but corrupts relation text, the performance drop directly supports the role of transition semantics rather than memory size alone.

The paired comparison in Table 16 makes the effect less dependent on a single average. Full VTEM beats shuffled transition on 11 questions while losing on 4, with 7 large wins of at least two score points. It also has positive win/loss balance against uniform+flat, flat retrieval, and temporal-chain controls. The sign-test values are not used as a strong statistical claim because the subset is small; instead, the paired counts show where the final average comes from and prevent isolated examples from driving the interpretation.

The submodule ablation shows that both modules matter under the same answer pipeline. TMC association only reaches 0.530, TMC value only reaches 0.535, THR anchor only reaches 0.550, and THR multi-hop only reaches 0.550. Full VTEM improves over each of these rows. The TMC rows indicate that association alone can retain too many memories and value scoring without transition association can retain too few; the full construction policy gives the strongest downstream answers.

The THR table shows that hierarchical anchoring is the largest retrieval contributor: anchor-only retrieval improves over flat retrieval from 0.510 to 0.550 while using fewer evidence memories. The stricter no-association-cue control withholds association and transition wording during neighbor selection and answer generation; it drops from 0.570 to 0.510 while retrieving 10.38 memories per question. This shows that simply expanding over more graph neighbors is not enough. The temporal-chain row is important for the same reason: it expands much more often and reaches larger hop depth than Full VTEM, yet scores lower, showing that uncontrolled adjacent expansion is not a substitute for transition semantics.

Table 17 asks whether the gains appear on the kinds of questions where transition and cross-view memory should matter. On the 14 spatial/localization questions, Full VTEM reaches 0.554 compared with 0.357 for flat retrieval and 0.304 for shuffled transition. On the broader transition-sensitive split, Full VTEM reaches 0.562 compared with 0.465 for flat retrieval, 0.424 for shuffled

transition, and 0.417 for the stricter no-association-cue control. The recognition/attribute split does not show the same advantage, which is consistent with the mechanism: VTEM improves memory organization and cross-view evidence selection, but it does not remove all fine-grained visual recognition errors from the answer model.

A.5 TMC Value Assessment Analysis

Table 18: Qualitative TMC value-assessment cases. The examples summarize representative decisions observed in the final memory traces.

Case type	Signal pattern	TMC decision	What it validates
View change with little information gain	The view remains in the same scene and the descriptor marks repeated content; CLIP appearance may change because of rotation, but text and relation novelty are low.	Drop the candidate and clear the temporary association.	TMC does not store every camera motion or visual novelty spike as a useful memory.
Small visual change with important state information	The view is visually close to a retained memory, but the description introduces a state or relation needed for QA, such as an open/closed shelf, box, door, or object placement.	Retain the candidate and store the edge descriptor.	Association-aware semantic value can override weak visual novelty when the new observation changes answer-relevant evidence.
Scene or floor transition	The descriptor marks a scene/floor change or a clear entering/leaving relation between two retained views.	Retain the candidate and connect it to the previous retained node.	Transition edges provide the cross-view relation later used by THR for context-aligned expansion.

The qualitative cases in Table 18 complement the numeric TMC ablation. The association descriptor is built before the write decision and is cleared if the candidate is rejected, so it functions as a temporary cue for value assessment rather than a new memory item. This design helps separate three situations that pure visual novelty can confuse: viewpoint motion without new information, small visual changes with answer-relevant state evidence, and genuine scene or floor transitions. The first case prevents memory growth from camera rotation alone, while the latter two cases explain why TMC can keep useful observations even when visual appearance is similar to a retained view.

A.6 Exact Definitions of Ablation Configurations

The THR-only rows are defined by which retrieval submodule is active while the memory-construction side is kept fixed. "Hierarchical anchor only" keeps the scene-view anchor retrieval but disables the sufficiency-triggered graph expansion. "THR multi-hop only" disables the hierarchical decomposition and starts from flat anchors, then uses retained transition edges for expansion. This makes the two THR submodules separable: one tests anchor localization and the other tests transition-guided evidence completion.

A.7 Transition Descriptor Consistency and Robustness

The final full VTEM memory for the 50-question subset contains 86 retained nodes and 82 retained transition edges across four episodes. The descriptor audit is an automatic consistency check, not a claim of human-level semantic accuracy. It verifies whether each retained edge has a complete schema, agrees with its source and target endpoints, has scene/floor flags consistent with those endpoints, and uses a descriptor type consistent with the edge state. All 82 retained descriptors pass these checks. Among the 41 same-scene edges, all are marked as informative same-scene writes

Table 19: Exact definitions of ablation configurations.

Configuration	Gate	Assoc.	Value	Anchor	Suff.	Hop	Definition
Full VTEM	on	on	on	on	on	on	Full TMC memory construction followed by THR anchor retrieval and bounded transition expansion.
TMC association only	on	on	off	on	on	on	Builds temporary associations and retained edges, but disables association-aware value scoring.
TMC value only	on	off	on	on	on	on	Uses value and novelty storage without temporary transition association or association weight.
THR anchor only	on	on	on	on	off	off	Uses hierarchical scene-view anchor localization and directly answers from anchors.
THR multi-hop only	on	on	on	off	on	on	Skips question decomposition for hierarchical anchoring and expands from flat anchors through transition edges.
Temporal chain	on	on	on	on	on	temporal	Uses temporal adjacency for expansion instead of transition descriptor scoring.
Shuffled transition	on	on	on	on	on	on	Keeps nodes and edge endpoints fixed but randomly permutes edge descriptors.
Flat retrieval	on	on	on	off	off	off	Uses full constructed memory but retrieves globally without THR.
Uniform + flat	off	off	off	off	off	off	Uniform frame storage and global flat retrieval.

Table 20: Direct consistency audit of retained transition descriptors in the final full VTEM memory.

Audit item	Pass	Total	Rate
Schema completeness	82	82	1.000
Endpoint agreement	82	82	1.000
Scene/floor flag agreement	82	82	1.000
Descriptor type agreement	82	82	1.000
Same-scene informative writes	41	41	1.000

through new visible detail or association-guided transition value, which is the intended behavior for views that remain in the same room but reveal complementary context.

Descriptor robustness is evaluated downstream by corrupting the descriptor while keeping the node set fixed. Shuffling descriptors reduces the normalized score from 0.570 to 0.480, and replacing semantic descriptors with temporal adjacency reduces it to 0.525. These two controls show that the stored edge semantics are used by THR and that the benefit is not explained by extra nodes, larger evidence budgets, or simple before/after adjacency.

A.8 Full Prompts, Output Schemas, and Additional Examples

This subsection reports the prompt templates used by the two VTEM modules and the OpenEQA evaluator. Curly-brace fields are filled at runtime, and image inputs are supplied as visual messages immediately before the corresponding text prompt. Minor formatting such as item numbering is implementation-dependent and does not change the instructions or output schema.

Scene update and view summary. For each candidate view that passes the CLIP gate, the scene LVLM receives the current image, optional candidate scene labels, and a compact trajectory history. When a fixed scene vocabulary is available, we use the following template:

```
You are analyzing a video frame captured by an embodied agent navigating
↪ an
indoor environment. Based solely on the visual content of this image,
↪ classify
the scene into ONE of the following categories:
{scene_types}
{trajectory_context}
Rules:
- Pick the single best-matching category from the list above.
- Do NOT invent new categories.
- The description must be useful for future question answering, not just
↪ scene
  naming.
- In the description, mention salient small objects, wall/ceiling
↪ fixtures,
  counter/table items, door states, artwork/TVs, and important spatial
  relations when visible.
- Keep the description grounded in the image. If fine details are
↪ unclear,
  describe the visible region and uncertainty instead of omitting it.
Respond ONLY with valid JSON (no markdown fences, no extra text):
{"scene_type": "<one category from the list>",
 "confidence": "<high|medium|low>",
 "description": "<one compact but detailed sentence describing visible
 ↪ objects,
 states, fixtures, and spatial relations>"}
```

The optional trajectory block is:

```
Trajectory History (the agent's previously visited floors and scenes, for
reference):
{history}
Use this history to better infer continuity and context, but ground your
↪ answer
primarily in the current image. Pay attention to the stability and
↪ continuity of
scene changes.
```

If no fixed vocabulary is provided, the first instruction is replaced by "identify the type of scene or location shown", and the output `scene_type` is a concise snake-case label. Invalid JSON is repaired by extracting the first JSON object; if parsing still fails, the scene is set to `unknown`, confidence to `low`, and the raw visual summary is truncated.

Temporary association and value assessment. TMC does not call an extra LVLM for every association. It builds a temporary association from the latest retained memory to the current view using scene and floor continuity, image similarity, text similarity, content-word novelty, and relation-word novelty. The association is used only for the current value assessment and is kept as a transition edge only if the candidate is written to memory. The descriptor templates are:

```
first retained memory candidate
moving from {source_scene} on {source_floor} to {target_scene} on
↪  {target_floor}
entering {target_scene} from {source_scene} on {floor}
repeated view of {scene} on {floor}
observing {scene} on {floor} from a changed egocentric viewpoint with
  complementary visual context
same {scene} context with additional visible detail
```

The retained edge text appends compact source/target visual summaries:

```
{base_transition}; before view in {source_scene}: {source_summary};
after view in {target_scene}: {target_summary}
```

This keeps transition evidence queryable by object and spatial anchors without turning the temporary association itself into a permanent memory node.

Question decomposition. THR decomposes each question into floor, scene, visual-anchor, and association constraints. The floor prompt is:

```
You will be given:
- A user question
- Candidate floor IDs

Your task:
Extract floor ID based on whether information about the floor appears in
↪  the
question. If there is no extremely accurate floor information, choose all
↪  floor
IDs.

Output schema:
{"target_floors": ["target_floors"], "notes": "short reason"}

Question:
{question}

Candidate floors:
{floors}
```

The scene prompt is:

```
You will be given:
- A user question
- Candidate scenes

Your task:
Select target scenes from the candidate list.

Rules:
- Choose specific scenes only if the question gives clear scene
↪  information.
- If the scene is not explicitly stated, select ALL candidate scenes.
- Only return scenes from the candidate list.
- If you do not detect decisive scene keywords, then return all candidate
↪  scenes.

Output schema:
{"target_scenes": ["scene"], "notes": "short reason"}
```

```
Question:
{question}

Candidate scenes:
{scenes}
```

The visual-anchor prompt converts the question into an image-retrieval phrase:

```
You will be given:
- A user question

Your task:
Convert the question into EXACTLY ONE short visual target constraint for image
retrieval. The target constraint should describe the visual evidence that must
be inspected, not the answer.

Rules:
- Output exactly one declarative sentence.
- Do not output a question.
- Never guess the answer category. If the question asks "what is on the
  ceiling", do not write "chandelier" or "light" unless that word appears in
  the question.
- Preserve visible anchors, spatial relations, and necessary locations from the
  question, such as wall, ceiling, counter, right of, underneath, next to,
  upstairs, bedroom, kitchen, hallway.
- For yes/no or state questions, preserve the visual state predicate:
  open/closed, on/off, set, clean, tucked in, full/empty, fenced, available
  space, enough room, or whether an object can fit.
- For unknown targets, describe the region to inspect, e.g. "the center of the
  ceiling" or "the tray on the kitchen counter".
- The statement MUST start with exactly: "a photo showing"
- Keep it short and concrete.

Output schema:
{"statement": "...", "notes": "Kept only the strongest visible anchor."}

Question:
{question}
```

The association prompt is:

```
You will be given:
- A user question

Your task:
Extract the transition/association constraint needed for transition-guided
memory retrieval. This constraint should describe the cross-view relation that
may require expanding from an anchor memory to neighboring memories. It can
include trajectory clues such as entering, leaving, passing through, going
upstairs/downstairs, or before/after a transition, and it can also include
spatial association clues such as opposite, across from, left/right of, behind,
in front of, between, next to, above, or below an anchor object.
```

```
Rules:
- If the question contains no movement, transition, or spatial
↪  association clue,
  return an empty string.
- Do not answer the question.
- Keep the constraint short and concrete.
- Preserve only transition or association words from the question.

Output schema:
{"movement_constraint": "", "notes": "short reason"}

Question:
{question}
```

The decomposer uses deterministic decoding with `max_new_tokens=128`. If JSON parsing fails, the visual query falls back to the raw question and the association constraint is empty.

Sufficiency assessment. The sufficiency judge receives the question and the top retrieved memory images. The system instruction is:

```
You are an embodied-agent question answering model. You are given memory
↪  images
from the agent's past observations. Your job is to answer the question
↪  using the
visual evidence in these images. Do not be overly conservative: if the
↪  images
contain enough partial but reliable evidence to answer the question,
↪  provide the
best concise answer. Only say the answer is unknown if the relevant
↪  evidence is
truly missing or too unclear. Output only the final answer in the format:
<answer>...</answer>
```

The user prompt is:

```
Answer the question using the provided memory images.

Instructions:
- Use only the visual evidence from the images.
- If the answer is visible or can be inferred with high confidence from
↪  the
  images, answer it directly.
- Do not be overly strict: partial but clear evidence is enough.
- If multiple images are provided, combine evidence across them.
- If the answer truly cannot be determined from the images, output:
  <answer>unknown</answer>
- Keep the answer concise.
- Output only one final answer in XML format.

Question: {question}

Return exactly:
<answer>your answer here</answer>
```

Answers such as `unknown`, `cannot determine`, and `unclear` are parsed as insufficient.

Multi-hop selector. When anchor evidence is insufficient, THR expands over retained transition edges. The selector receives text summaries for the current evidence and the neighbor memories connected at hop `h`:

```
You are guiding transition-guided hierarchical retrieval for embodied QA.

Question:
```

```
{question}

Visual anchor constraint:
{visual_anchor_constraint}

Association constraint:
{association_constraint}

Current retrieved memory:
{current_memory_summary}

Neighbor memories connected by retained transition edges at hop {h}:
{neighbor_memory_summary}

Task:
- Decide whether the current memory already contains enough evidence.
- If not enough, select up to {K} neighbor node_ids that are most likely to add
  complementary evidence through the transition edges.
- Prefer neighbors whose transition relation helps with opposite/across/
  left/right/behind/in-front/between/near/above/below relations or before/after
  movement clues.
- Do not select redundant neighbors that repeat the same view.

Return only JSON:
{"sufficient": false,
 "selected_node_ids": ["m_000001"],
 "notes": "short reason"}
```

Parsing failures default to the top scored neighbors. The selector uses `max_new_tokens=160` and deterministic decoding.

Final answer generation. The final generator receives the selected evidence package. Each item is formatted as:

```
Evidence {i}: floor={floor_id}, scene={scene_id}, frame={frame_id};
transition={transition_text}; description={memory_description};
focused_observation={optional_low_confidence_hint}
Evidence {i} full/crop view:
{image}
```

The final prompt is:

```
Question: {question}
Use the full evidence package: images, crops, descriptions, movement context,
and focused observations. The text evidence may name small objects, states, or
relations that are hard to inspect in low-resolution images. If text and images
are consistent or the image is too small to verify a text detail, use the text
detail rather than saying the object does not exist.

Treat movement context as a navigation and retrieval hint that explains why
views are adjacent or complementary; do not treat transition wording alone as
visual proof of an object, color, count, or spatial relation.

Focused observations are low-confidence helper hints for attribute, spatial,
affordance, and state questions. Never use a focused observation as the only
```

```
support for a color, material, shape, object type, spatial relation,
↪  object
absence, or yes/no state. If a focused observation conflicts with the
↪  original
description, movement context, or visible image, ignore the focused
↪  observation.

Inspect small details carefully. If evidence is partial, give the best
↪  concise
answer supported by the memory instead of defaulting to no or unknown.
↪  For
wall/ceiling fixtures, distinguish detectors, vents, lights, switches,
thermostats, and speakers by shape and placement. For functional
↪  questions,
choose the most relevant visible tool, place, control, appliance, or
↪  surface
from the evidence. For usable-space and open/closed questions, inspect
↪  the
exact requested target and state. For attribute questions, answer the
↪  requested
attribute rather than a nearby material or object.

Output only the final answer in the format:
<answer>...</answer>
```

For the no-association-cue ablation, the `transition` field is withheld and the movement-context rule is replaced by: "Association and movement context are unavailable; answer only from the visible images, crops, and original memory descriptions."

Automatic evaluation. OpenEQA automatic evaluation uses GPT-4-1106-preview with temperature 0. We use the following LLM-match scoring prompt:

```
You are an AI assistant who will help me to evaluate the response given
↪  the
question and the correct answer. To mark a response, you should output a
↪  single
integer between 1 and 5 (including 1, 5). 5 means that the response
↪  perfectly
matches the answer. 1 means that the response is completely different
↪  from the
answer.

Example 1:
Question: Is it overcast?
Answer: no
Response: yes
Your mark: 1

Example 2:
Question: Who is standing at the table?
Answer: woman
Response: Jessica
Your mark: 3

Example 3:
Question: Are there drapes to the right of the bed?
Answer: yes
Response: yes
Your mark: 5

Your Turn:
Question: {question}
Answer: {answer}
Response: {response}
```

When extra acceptable references are provided, the evaluator also receives an `Extra Answers` field and a score of 5 is assigned if the response matches any accepted reference. The reported OpenEQA metric is the mean normalized score $(\text{score} - 1)/4$ over valid examples.

Table 21: Main hyperparameters used in the appendix experiments.

Item	Value
Inference model	Qwen2.5-VL-7B-Instruct
CLIP model	ViT-L/14-336px
Initial frame stride	5
CLIP gate window / stat window	8 / 8
CLIP gate cooldown / quantile	1 / 0.70
TMC λ	0.5
TMC write threshold τ	0.45
TMC novelty window	12 retained nodes
TMC association weight	0.45
Scene retrieval top-k	5 text + 5 visual/hybrid
Answer evidence budget	12 memory candidates
Focused observation top-k	4 candidates
Focused observation views	5 views/candidate
THR max hop	2
THR neighbor candidates per hop	12
THR selected neighbors per hop	5
Sufficiency top-k	4 memory images
Answer max tokens	512
Evaluation model	GPT-4-1106-preview
Evaluation workers	16 CPU threads

Hyperparameters.

A.9 Qualitative Retrieval Traces and Failure Cases

The qualitative traces are chosen from the final scored outputs rather than from an earlier development run. They illustrate both sides of the mechanism. For spatial and state questions, Full VTEM can recover the correct referenced object or state when flat, temporal-only, or shuffled-transition retrieval selects incomplete or mismatched evidence. The failure case shows the remaining boundary: when the answer depends on a small color attribute, better memory organization cannot fully compensate for the visual acuity of the answer model.

Table 23 provides a compact textual view of the THR process. Each example starts from an anchor memory found by hierarchical matching and then uses the association cue to decide which transition-connected memories should be added. The important distinction from nearest-neighbor retrieval is that neighbor selection is conditioned on the relation required by the question, such as shelf order, left-of relations, or the wall next to an elevator.

Table 24 broadens the qualitative evidence beyond one OpenEQA episode. The HM3D and ScanNet rows are taken from the same controlled 50-question ablation runs, while the RealEQA rows are taken from the real-video experiment scored with the same GPT-4-1106-preview protocol. These examples are not meant to replace aggregate metrics; they show the concrete retrieval failures behind the averages. VTEM is most visibly better when the question requires preserving a scene-specific state, a spatial relation, or a real-video anchor that visually similar frame-level baselines can confuse.

Figures 6 and 7 attach the corresponding trajectory evidence to these cases. The OpenEQA figure includes six representative questions, while the real-video figure includes two cases from the collected robot videos. In each row, the ranked views expose why the answer is supported: VTEM keeps the frames where the referenced object, relation, or state is visible instead of relying only on adjacent frames or visually similar but mismatched views.

Table 22: Representative retrieval outcomes from the final scored runs.

Question	Category	Full VTEM outcome	Ablation or failure observation
How many boxes are on the fourth shelf?	Spatial understanding	Correct answer `2`, score 5.	Flat retrieval answers `0` and temporal chain answers `three`; both score 1, showing that extra temporal neighbors do not replace relation-aware evidence selection.
What is to the left of maroon books on the top shelf?	Spatial understanding	Correct answer `blue books`, score 5.	Shuffled transition and no-association-cue variants fail to localize the referenced relation and answer that no maroon books are visible.
Are the boxes open or closed?	Object state recognition	Correct answer `closed`, score 5.	Temporal-only expansion answers `open`, score 1, illustrating an error from following adjacency without useful transition semantics.
What color is the trim around the mirror?	Attribute recognition	Full VTEM answers `white`, score 1, while the ground truth is gold.	This failure is dominated by fine-grained color recognition; transition expansion cannot compensate when the answer generator misreads a small attribute.

Table 23: Textual THR traces for relation-guided expansion.

Question	Anchor memory	Transition-guided expansion	Answer effect
How many boxes are on the fourth shelf?	A shelf or closet anchor retrieved by scene-view matching.	THR follows retained transition edges to neighboring shelf views whose descriptors preserve the same storage context, rather than arbitrary adjacent frames.	Selects the box-visible shelf evidence and answers `2`.
What is to the left of maroon books on the top shelf?	A bookshelf/top-shelf anchor with maroon books.	The association cue keeps left-of/top-shelf relation terms active during neighbor selection, avoiding visually similar table or room views.	Selects the bookshelf relation evidence and answers `blue books`.
What number is displayed on the wall next to the elevator?	A hallway/elevator anchor in the real-video memory.	Expansion preserves the elevator-side wall relation and separates it from a nearby sign observed along the same trajectory.	Selects the correct wall-number evidence and answers `10`.

Table 24: Cross-dataset cases with stronger VTEM answers than comparison methods.

Dataset	Question and ground truth	VTEM answer	Weaker comparison and interpretation
OpenEQA-HM3D	Are the drapes in the bedroom open or closed? GT: open.	`open`, score 5.	Flat retrieval and uniform+flat both answer incorrectly, score 1. The retained memory localizes the bedroom state instead of retrieving visually similar but irrelevant room evidence.
OpenEQA-HM3D	Where is the red and white checkered rug? GT: underneath the wooden dining table.	Near the dining table, score 4.	Flat, shuffled-transition, and uniform+flat score 1. The case illustrates that scene-aware memory selection helps preserve the dining-area relation needed for object localization.
OpenEQA-ScanNet	How many boxes are on the fourth shelf? GT: two.	`2`, score 5.	Flat, temporal-chain, anchor-only, and uniform+flat variants score 1. The failure of temporal-only expansion shows that adjacency alone does not recover the needed shelf-level relation.
OpenEQA-ScanNet	What is to the left of maroon books on the top shelf? GT: blue books.	`blue books`, score 5.	Shuffled transition scores 1 and no-association cue scores 3. This is a case-level example where association-guided edge semantics prevent retrieval from settling on the wrong bookshelf relation.
RealEQA video1	Where are the tables and chairs located? GT: beside the large windows near the end of the hallway.	Beside the large windows, score 5.	GeLM, 3D-Mem, and Embodied VideoAgent answer only "in the room", score 2. The VTEM memory keeps the hallway-to-window context that makes the location specific.
RealEQA video2	What number is displayed on the wall next to the elevator? GT: 10.	`10`, score 5.	RoboBrain, GeLM, and 3D-Mem answer `9`, score 1. The case shows that transition-guided retrieval can preserve the correct real-video anchor frame instead of a nearby but wrong elevator cue.

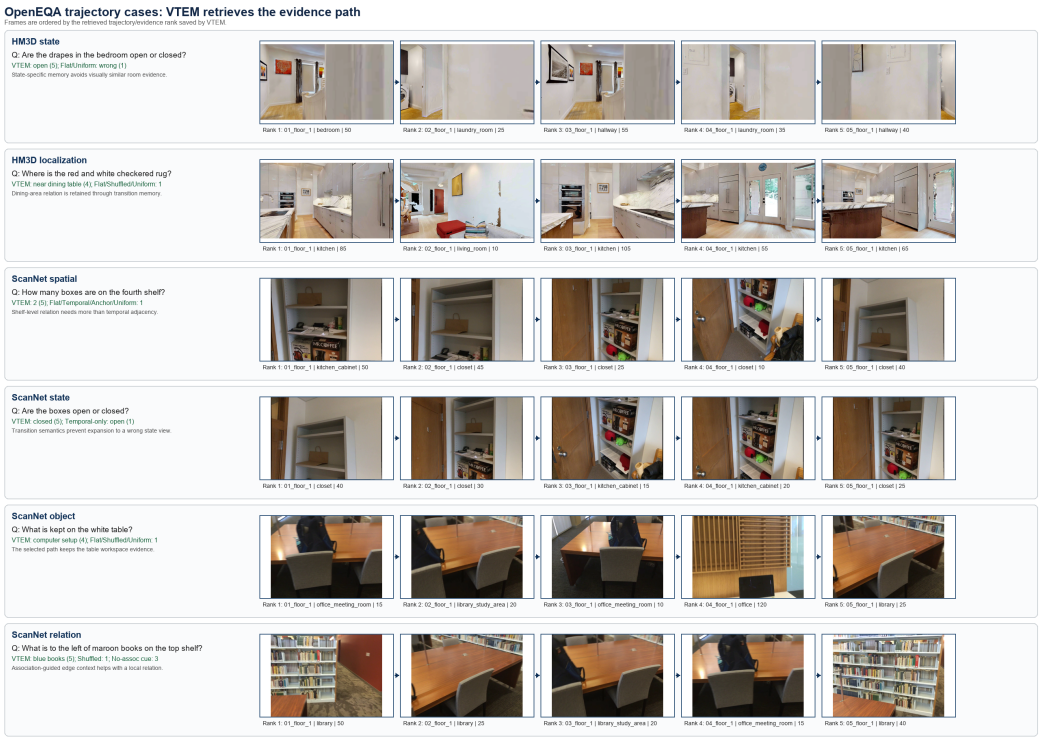


Figure 6: Trajectory evidence for six OpenEQA cases. Each row shows the question, the VTEM answer, and the ranked visual memories returned to the answer model. The examples cover HM3D and ScanNet questions whose answers depend on scene-specific state, object count, or spatial relation evidence.

A.10 Failure Analysis

Transition detection failure. VTEM can fail when the LVLM summary or the temporary descriptor assigns an incorrect scene change, object state, or relation to a transition. In this case, TMC may store a less useful edge, and THR may expand from a correct anchor to a neighbor whose relation text is misleading. The descriptor audit checks schema and endpoint consistency, but it does not guarantee perfect semantic interpretation of every visual transition.

Attribute-heavy questions. The main type analysis shows that attribute recognition benefits less from transition-based retrieval than spatial or functional reasoning. These failures usually occur when the answer depends on a small color, material, trim, or fine object attribute. Better memory organization can recover the right view, but it cannot fully correct an answer model that misreads a tiny local appearance cue.

Ambiguous temporal or repeated-object relations. When visually similar objects appear in several nearby places, the transition chain may contain multiple plausible paths. THR reduces this ambiguity by using scene constraints, view anchors, and association cues, but an underspecified question can still make two paths hard to distinguish. These cases motivate the bounded multi-hop design: expansion is useful when the relation is clear, but it is stopped when additional neighbors become redundant or weakly grounded.

A.11 Limitations

VTEM still depends on LVLM judgments for scene summaries, question decomposition, sufficiency, transition-neighbor selection, focused observations, and answer generation. Memory construction

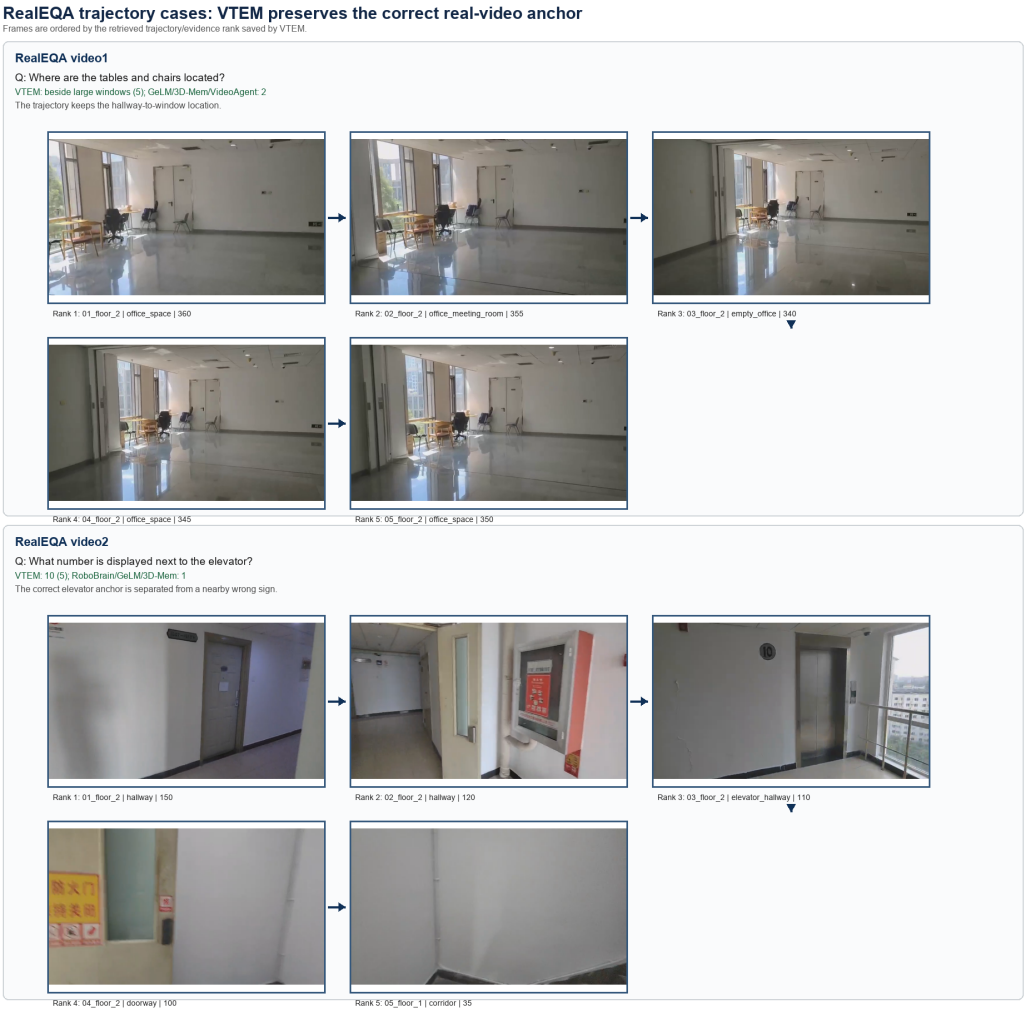


Figure 7: Trajectory evidence for two RealEQA cases. The retrieved memories keep the real-video anchor views needed to distinguish the target location or elevator sign from nearby visually similar frames.

introduces a one-time offline cost compared with uniform frame storage. The transition descriptor is inferred from retained visual summaries and CLIP similarities rather than measured camera motion, so it can be wrong when the scene summary is wrong. The retained chain is linear and therefore only approximates loop closure. Fine-grained colors, small objects, materials, and counting remain limited by image resolution and the visual acuity of the LVLM.

A.12 ECBench-MultiScene Construction

The processed ECBench-MS candidate pool contains 736 multi-scene examples; after the same validity filtering used by the evaluator, 732 scored samples are used in the main experiments. Each item contains `video_name`, `dimensions`, `video_source`, `question_en_v2.2`, `answer_en_v2.2`, `scene_type`, `source_dataset`, and `dataset_pth`. The multi-scene definition follows the `scene_type` list associated with each video: 531 examples contain three scene labels, 197 contain four, and 8 contain five. The sources are HM3D (576), ScanNet (128), hallucination diagnostics (26), and dynamic examples (6). The benchmark covers spatial imagery, distance and azimuth awareness, movement imagery, trajectory review, temporal state reasoning, functional reasoning, object recognition, attribute recognition, and localization.