CONNECTING-THE-DOTS: BENCHMARKING REFLECTIVE MEMORY IN LONG-HORIZON DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite substantial progress in long-context modeling, existing benchmarks remain confined to factual memory for explicit recall, failing to measure the reflective memory required to synthesize fragmented, multimodal cues into high-level interpretations. To address this gap, we introduce **RefMem-Bench**, a benchmark for reflective memory in long-horizon dialogue. RefMem-Bench contains 26K annotated QA instances with eight reflective-memory dimensions and three task formats, requiring models to move beyond surface-level retrieval and infer latent meanings from evidence distributed across interaction histories. To enhance reflective memory capability, we propose **RE**flective **M**emory **IND**uction (**REMIND**), a hierarchical framework that treats reflective memory as progressive meaning construction. REMIND couples question-conditioned evidence retrieval, salience-aware grounding, and abstraction-level supervision, and uses Progressive Reflective Alignment to distill high-level reflective reasoning into the factual inference pathway. Experiments show RefMem-Bench poses a substantial challenge to current models, while REMIND consistently improves both answer accuracy and memory recall through progressive evidence perception, grounding, and abstraction.

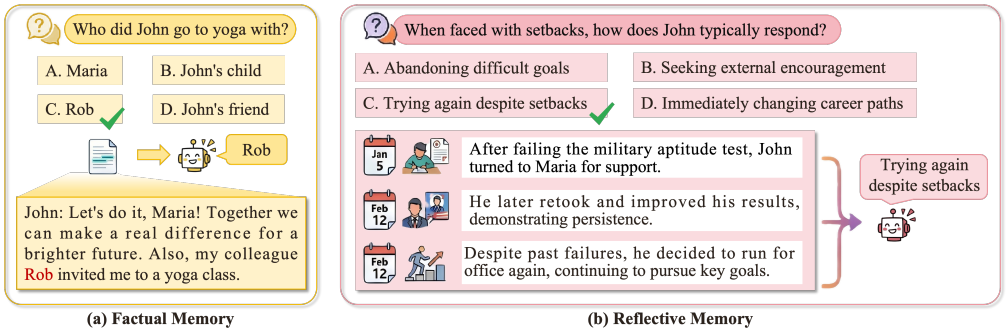
1 INTRODUCTION

As Large Language Models (LLMs) are applied to increasingly long and multimodal contexts, memory has emerged as a key factor in their practical performance (Lewis et al., 2020; Zhong et al., 2024). A strong model must do more than process the immediate input by instead incrementally accumulating user experiences, tracking evolving states, and capturing recurring behavioral patterns across extended turns and sessions (Wang et al., 2023). Recent advances in retrieval-augmented generation, memory-augmented modeling, and long-context training have significantly improved the ability of LLMs to access vast interaction histories (Team et al., 2024; Bulatov et al., 2023). These developments move LLMs closer to becoming persistent companions that can leverage a wealth of prior experiences to produce responses that are both personalized and contextually grounded.

However, existing evaluation paradigms are still largely centered on explicit memory recall, measuring whether a model can retrieve information that has been directly stated in the interaction history. Benchmarks such as LongBench (Bai et al., 2025), LoCoMo (Maharana et al., 2024), and LongMemEval (Wu et al., 2024) are representative examples of this setting. As illustrated in Figure 1 (a), these tasks typically require models to identify surface-level details from prior context, such as a specific name or a visually observable object. While such abilities are important, they capture only a limited notion of memory. From a schema-based view of memory (Alba & Hasher, 1983), effective remembering involves not only retaining details, but also synthesizing distributed signals into structured, higher-level representations. In long-horizon dialogue, this means that *models must go beyond explicit retrieval and integrate scattered historical cues to infer latent states, behavioral patterns, or personal traits*, as illustrated in Figure 1 (b). Yet current benchmarks provide little systematic evaluation of this higher-level synthesis, creating the need for a dedicated benchmark for reflective memory that is essential to building truly intelligent long-term agents.

To bridge this gap, we introduce **RefMem-Bench**, a benchmark for reflective memory in long-horizon dialogue. RefMem-Bench contains 26K annotated question-answer instances covering 1K sessions and 71K dialogue turns. Unlike benchmarks centered on explicit recall, it spans eight reflective-memory dimensions and three task formats, requiring models to integrate distributed evidence across

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.



(a) Factual Memory **(b) Reflective Memory**

Figure 1: Example of factual memory and reflective memory.

extended interaction histories to infer latent user states, behavioral regularities, temporal evolution, and implicit constraints. All instances are evidence-anchored and human-verified to support reliable evaluation of memory reasoning. To improve performance on this capability, we further propose **REMIND** (**RE**flective **M**emory **IND**uction), a hierarchical framework that treats reflective memory as meaning construction rather than evidence retrieval alone. REMIND organizes reasoning through a three-level **Cognitive Pyramid**, consisting of factual, attentional, and reflective states, and is trained with **Progressive Reflective Alignment** to distill high-level reflective reasoning into the factual inference pathway, enabling efficient test-time inference without explicit multi-stage reasoning.

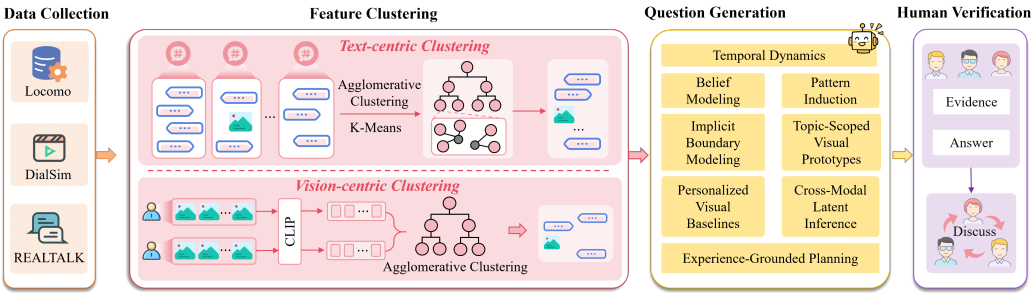
To summarize, our work provides three key contributions: 1) We formalize reflective memory as a distinct target for long-term LLM evaluation, introduce RefMem-Bench [1] as a large-scale multimodal benchmark with evidence-anchored and human-verified annotation; 2) we propose REMIND as a unified framework that integrates factual retrieval, salience grounding, and abstraction induction in a hierarchical training objective; and 3) extensive experiments show that RefMem-Bench is highly challenging for current models, while REMIND consistently improves both answer accuracy and memory recall across task formats and backbone settings.

2 Related Work

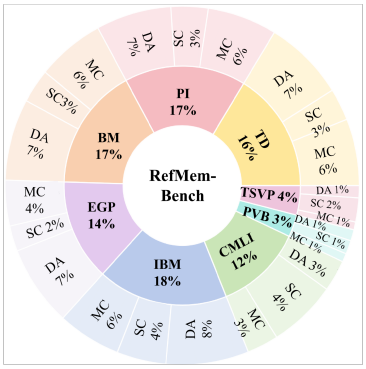
Long-Term Dialogue Benchmarks. Recent long-term dialogue benchmarks have substantially advanced the evaluation of memory in sustained interactions, yet most remain centered on explicit recall. Representative examples include LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2024), which evaluate memory through long-horizon question answering, summarization, temporal reasoning, and knowledge updates. REALTALK (Lee et al., 2025), PersonaMem (Jiang et al., 2025), and LifeBench (Cheng et al., 2026) move toward more realistic long-span interactions and evolving user profiles, while MADial (He et al., 2025), MMRC (Xue et al., 2025), and MemGallery (Bei et al., 2026) extend evaluation to memory-aware generation and multimodal conversation. In contrast, our work focuses on reflective memory: whether models can integrate temporally scattered and cross-modal cues into coherent latent interpretations.

Long-Term Memory Methods. Existing long-term memory methods for LLMs mainly improve interaction over long horizons by introducing external memory structures, latent memory mechanisms, or adaptive retrieval pipelines. Early systems such as MemoryBank (Zhong et al., 2024) and MemGPT (Packer et al., 2023) show that explicit memory storage can help preserve salient user information beyond the immediate context window, while model-level approaches such as LongMem (Wang et al., 2023) and MemoryLLM (Wang et al., 2024) explore cached context and latent memory pools for scalable long-range retention. More recent systems, including Mem0 (Chhikara et al., 2025), MemoryOS (Kang et al., 2025), A-Mem (Xu et al., 2025), and M+ (Wang et al., 2025), further improve scalability, structured storage, retrieval quality, and adaptive memory management. In contrast to the above memory systems, we treat long-term memory as hierarchical meaning construction, emphasizing how models transform retrieved traces into abstraction-level reasoning.

[1] Code and dataset are available at https://anonymous.4open.science/r/RefMem-Bench.



(a) Workflow of dataset construction



Benchmark	Statistics			Benchmark Properties			Reflective Memory Dimensions							
	Tot. Conv.	Avg. Sess.	Avg. Turns	MM	EA	HA	TD	PI	BM	EGP	IBM	CMLI	PVB	TSVP
Dialogue Resources														
DailyDialog	13,118	1.0	7.9*	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
MMDialog	1.08M	1.0	4.6*	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
MSC	1,001	4.0	13.3	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Conversation Chronicles	200K*	5.0	11.7*	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Memory-Oriented Benchmarks														
DialSim	3	1.3K	–	✗	✓	△	✓	△	✗	✗	✗	✗	✗	✗
LoCoMo	10	27.2	21.6	✓	△	✓	✓	△	✗	✗	✗	✗	✗	✗
LongMemEval	50K sess.	–	–	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
REALTALK	10	21.9	40.8	✓	✗	✓	△	✗	✗	✗	✗	△	✗	✗
MMRC	5,120	1.0	15.2	✓	✗	✓	△	✗	✗	✗	✗	△	✗	✗
Mem-Gallery	20*	12.0*	16.51	✓	✓	✓	✓	✗	✗	✗	✗	△	✗	✗
LifeBench	10	361.8	35.2	✗	✓	△	✓	△	✓	✗	✗	✗	✗	✗
RefMem-Bench	**35**	**38.31**	**52.99**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

MM: Multimodal **EA**: Evidence-Anchored **HV**: Human-Verified **TD**: Temporal Dynamics **PI**: Pattern Induction **BM**: Belief Modeling **EGP**: Experience-Grounded Planning **IBM**: Implicit Boundary Modeling **CMLI**: Cross-Modal Latent Inference **PVB**: Personalized Visual Baselines **TSVP**: Topic-Scoped Visual Prototypes ✓: explicit coverage △: partial coverage ✗: the dimension is not a core focus

(b) Data statistic of our proposed RefMem-Bench

Figure 2: Overview of RefMem-Bench. (a) Workflow of dataset construction. (b) Comparison between RefMem-Bench and representative dialogue datasets and memory benchmarks. Statistics marked with * are derived from values reported in the original papers.

3 BENCHMARK

Dataset Curation. RefMem-Bench is built through a semi-automatic pipeline over long-horizon conversations collected from REALTALK (Lee et al., 2025), DialSim (Kim et al., 2024), and LoCoMo (Maharana et al., 2024), with complementary text-centric and image-centric branches (Figure 2). In the text branch, we segment multi-turn dialogues into fine-grained topical units, cluster them with Agglomerative Clustering (Johnson, 1967), and refine large clusters via K-Means (MacQueen, 1967), producing questions for **Temporal Dynamics**, **Pattern Induction**, **Belief Modeling**, **Experience-Grounded Planning**, and **Implicit Boundary Modeling**. In the image branch, we collect speaker-level image histories from REALTALK, encode them with CLIP (Radford et al., 2021), and cluster them to construct questions for **Cross-Modal Latent Inference**, **Personalized Visual Baselines**, and **Topic-Scoped Visual Prototypes**. All generated instances are reviewed by three expert annotators to verify answers and evidence, correct LLM-induced errors, and filter out ambiguous or low-quality questions, with inconsistent cases resolved by an adjudicator following prior annotation protocols (Wu et al., 2024). More details are provided in Appendix D.1.

Task Formulation. We instantiate this formulation with three complementary task formats: **single-choice** (**SC**), where the model selects one answer from four candidates; **multiple-choice** (**MC**), where it identifies all valid options among four candidates; and **direct-answer** (**DA**), where it produces a short free-form response that is matched against a small set of acceptable references.

Data Statistics. RefMem-Bench contains 25,573 annotated QA instances constructed from 35 curated long-horizon conversational contexts. Each instance includes a complete conversational thread and, when available, aligned visual evidence. Figure 2 shows that RefMem-Bench is substantially larger and more comprehensive than prior dialogue datasets and memory benchmarks in reflective-memory coverage. To the best of our knowledge, it is the only benchmark in our comparison that simultaneously supports multimodal, real-world, evidence-anchored, and human-verified evaluation across all eight reflective-memory dimensions. Following recent long-horizon memory-agent bench-

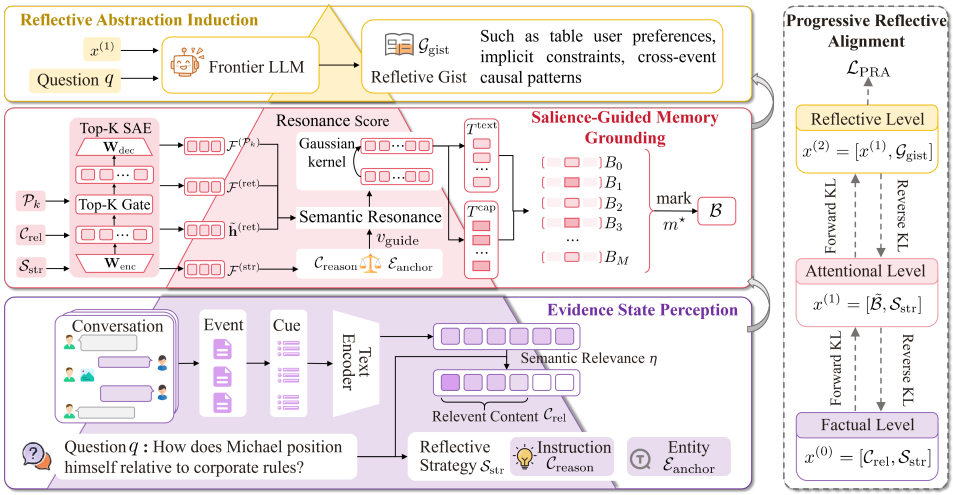


Figure 3: Overview of our REMIND method.

marks, including Mem0 (Chhikara et al., 2025) and Memory-R1 (Yan et al., 2025b), we use a 2:8 train/test split. Detailed statistics by task format and reflective-memory dimension are provided in Appendix D.6.

Evaluation Protocols. Following Maharana et al. (2024) and Wu et al. (2024), we evaluate all task formats using two primary metrics, answer accuracy (**Acc**) and memory recall (**MemR**), and additionally report BLEU-1 (**B1**) and token-level F1 (**F1**) for direct-answer (DA) questions. For SC and MC, Acc is computed by exact match between predicted and gold option sets; for DA, correctness is determined by an LLM-as-a-judge (Yang et al., 2023) using `gpt-4o-2024-08-06`. MemR is unified across the benchmark with modality-specific scoring: semantic evidence matching for text-dominant dimensions and evidence index overlap for visually grounded dimensions, both averaged over valid samples. For DA, B1 and F1 are computed from unigram overlap between predictions and references. Further details are provided in Appendices D.7 and D.8.

4 REMIND METHOD

Given a long-horizon multimodal dialogue history $\mathcal{H}$ and a reflective-memory question q, our goal is to generate an answer y. Unlike standard memory QA, reflective-memory questions require both evidence retrieval and higher-level inference over clues distributed across turns, sessions, and modalities. To address this challenge, we propose **RE**flective **M**emory **IND**uction (**REMIND**), a hierarchical framework with a three-level Cognitive Pyramid, trained via Progressive Reflective Alignment to distill higher-level reasoning into the factual inference pathway used at test time.

4.1 COGNITIVE PYRAMID MODELING

Reflective reasoning in long-horizon dialogue is inherently hierarchical: answering a reflective-memory question requires retrieving relevant evidence, identifying salient clues, and synthesizing them into higher-level interpretations. Consistent with a schema-based view of memory (Alba & Hasher, 1983) and prior work on hierarchical cognitive control (Badre, 2008), we formulate this process as a three-level **Cognitive Pyramid** with factual, attentional, and reflective levels. As shown in Figure 3, these levels correspond to *Evidence State Perception*, *Salience-Guided Memory Grounding*, and *Reflective Abstraction Induction*. Finally, **Progressive Reflective Alignment** distills higher-level reasoning into the factual inference pathway.

Let $x^{(i)}$ denote the state at level i, where $x^{(0)}$, $x^{(1)}$, and $x^{(2)}$ correspond to the factual, attentional, and reflective states, respectively. Given $x^{(i)}$, the model defines the autoregressive next-token distribution $p_i(y_t \mid y_{<t}) \equiv p_\theta(y_t \mid x^{(i)}, y_{<t})$, where $y_{<t} = (y_1, \ldots, y_{t-1})$ and θ denotes the model parameters.

The central idea of REMIND is to model reflective memory as progressive meaning construction rather than evidence retrieval alone. During training, the distribution induced by the factual state is encouraged to approximate those induced by higher-level states, so that attentional grounding and reflective abstraction are distilled into the base inference pathway. This yields efficient test-time reasoning without explicit multi-stage execution.

4.2 Evidence State Perception

We first instantiate the *factual level* by constructing a question-conditioned factual state $x^{(0)}$. This stage couples coarse-grained evidence retrieval with a structured reflective strategy, yielding broad but high-signal memory access for downstream grounding and abstraction. Specifically, we decompose the question into $\mathcal{S}_{\text{str}} = \big(\mathcal{C}_{\text{reason}}, \mathcal{E}_{\text{anchor}}\big)$, where $\mathcal{C}_{\text{reason}}$ is a set of generic reasoning instructions and $\mathcal{E}_{\text{anchor}}$ is a set of explicit answer-critical anchors, such as entities, attributes, or concepts.

Because the full dialogue history $\mathcal{H}$ contains substantial redundancy and multimodal noise, we first partition it into N temporally ordered segments: $\mathcal{U} = \{u_n\}_{n=1}^{N}$. For each segment u_n, we derive a segment-level event cue κ_n tailored to the question type and its cognitive demand from LLM. We then encode both the question and the cue with a text encoder $\phi(\cdot)$ and compute a relevance score η_n that measures the semantic relevance between question q and segment cue κ_n:

$$\eta_n = \text{sim}(q, \kappa_n) = \frac{\phi(q)^\top \phi(\kappa_n)}{\|\phi(q)\|_2 \, \|\phi(\kappa_n)\|_2}. \tag{1}$$

We rank candidate segments by η_n and retain the top-k_1 as $\mathcal{C}_{\text{ret}}$, yielding the factual state $x^{(0)} = [\mathcal{C}_{\text{ret}}, \mathcal{S}_{\text{str}}]$. This question-conditioned retrieval filters irrelevant events while preserving broad evidence coverage for subsequent grounding and abstraction.

4.3 Salience-Guided Memory Grounding

Starting from the coarse factual state $x^{(0)}$, this section highlights strategy-salient evidence and preserves its local interaction structure, yielding the attentional state $x^{(1)}$ for downstream abstraction.

Sparse Encoding. To enable strategy-guided grounding, we map both the reflective strategy and the retrieved context into a shared sparse latent space using a Top-K SAE (Gao et al., 2024). This sparse representation highlights the most salient semantic features of each token, allowing strategy and context tokens to be compared at the feature level rather than only through surface lexical overlap.

Let $\mathbf{h}_t^{(x)} \in \mathbb{R}^{d_{\text{model}}}$ denote the hidden state of token t, where $x \in \{\text{str}, \text{ret}\}$ indicates whether the token belongs to the reflective strategy or the retrieved context. The Top-K SAE maps $\mathbf{h}_t^{(x)}$ to a sparse latent representation $\tilde{\mathbf{h}}_t^{(x)} \in \mathbb{R}^{d_{\text{latent}}}$ and active feature set $\mathcal{F}_t^{(x)}$:

$$\tilde{\mathbf{h}}_t^{(x)} = \text{TopK}\left(\text{ReLU}(\mathbf{W}_{\text{enc}}(\mathbf{h}_t^{(x)} - \mathbf{b}_{\text{dec}}) + \mathbf{b}_{\text{enc}}),\, k\right), \qquad \mathcal{F}_t^{(x)} = \{j \mid [\tilde{\mathbf{h}}_t^{(x)}]_j > 0\}. \tag{2}$$

Here, $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{d_{\text{latent}} \times d_{\text{model}}}$ and $\mathbf{b}_{\text{enc}} \in \mathbb{R}^{d_{\text{latent}}}$ are the encoder parameters, $\mathbf{b}_{\text{dec}} \in \mathbb{R}^{d_{\text{model}}}$ is the decoder bias, $\text{ReLU}(\cdot)$ is the rectified linear unit, and $\text{TopK}(\cdot, k)$ keeps the top-k_1 activated dimensions and zeros out the rest.

Guidance Projection. We next project the reflective strategy $\mathcal{S}_{\text{str}}$ into the same sparse space to obtain a global guidance vector. Specifically, we aggregate feature activations from reasoning spans and anchor spans, while downweighting generic features and emphasizing answer-critical ones:

$$[\mathbf{v}_{\text{guide}}]_j = \max\Big(w_{\text{idf}}(j) \sum_{t \in \mathcal{T}_{\text{reason}}} \mathbb{I}[j \in \mathcal{F}_t^{(\text{str})}],\ \alpha \sum_{t \in \mathcal{T}_{\text{anchor}}} \mathbb{I}[j \in \mathcal{F}_t^{(\text{str})}]\Big). \tag{3}$$

Here, $\mathcal{T}_{\text{reason}}$ and $\mathcal{T}_{\text{anchor}}$ denote reasoning-span and anchor-span token sets from $\mathcal{C}_{\text{reason}}, \mathcal{E}_{\text{anchor}}$, $w_{\text{idf}}(j)$ is the inverse-document-frequency weight of latent feature j, and α controls anchor boosting. The resulting vector $\mathbf{v}_{\text{guide}}$ serves as a compact descriptor of the reflective strategy.

Semantic Resonance. Coarse retrieval provides broad coverage but remains noisy for reflective grounding, since salient clues are often sparse and embedded in local interaction structure. We therefore refine the retrieved dialogue by measuring token-level alignment with the strategy guidance

vector. For each turn $T_k \in \mathcal{C}_{\text{ret}}$, let $\mathcal{P}_k$ and $\mathcal{F}^{(\mathcal{P}_k)}$ denote its speaker and the feature set of its speaker marker, respectively. The token-level resonance score is defined as

$$r_t = \underbrace{\sum_{j \in \mathcal{F}_t^{(\text{ret})}} [\mathbf{v}_{\text{guide}}]_j [\tilde{\mathbf{h}}_t^{(\text{ret})}]_j}_{\text{local alignment}} + \underbrace{\beta \sum_{k=1}^{M} \mathbb{I}[t \in T_k \cup T_{k-1}] \sum_{j \in \mathcal{F}_t^{(\mathcal{P}_k)}} [\mathbf{v}_{\text{guide}}]_j}_{\text{speaker-conditioned broadcast}}, \tag{4}$$

where β controls turn-level broadcasting and $T_0 = \varnothing$. The first term captures direct alignment between a token and the reflective strategy in sparse feature space, while the second propagates speaker-conditioned salience to the current turn and its immediate predecessor, helping recover triggering context for replies whose relevance is implicit.

Turn Selection. To obtain a more stable estimate of local semantic density, we smooth token-level resonance scores with a one-dimensional Gaussian kernel $\mathcal{K}$: $s_t = r_t * \mathcal{K}$. For multimodal dialogue, we partition each turn into a text subset T_k^{text} and a visual-description subset T_k^{cap}, and compute turn salience via weighted fusion:

$$s_{\text{turn}}(T_k) = (1-\lambda)\ \text{Agg}\big(\{s_t \mid t \in T_k^{\text{text}}\}\big) + \lambda\ \text{Agg}\big(\{s_t \mid t \in T_k^{\text{cap}}\}\big), \tag{5}$$

where $\lambda \in [0,1]$ controls modality fusion and $\text{Agg}(\cdot)$ maps token-level salience to subset-level scores.

We retain salient turns $\mathcal{T}_{\text{core}} = \{T_k \in \mathcal{C}_{\text{ret}} \mid s_{\text{turn}}(T_k) > \delta\}$, where δ is a salience threshold, and expand each selected turn with a neighborhood window of size w to obtain $\mathcal{T}_{\text{final}} = \bigcup_{T_k \in \mathcal{T}_{\text{core}}} \{T_{k-w}, \ldots, T_{k+w}\}$. Adjacent turns in $\mathcal{T}_{\text{final}}$ are then merged into contiguous semantic blocks $\mathcal{B} = \{B_i\}_{i=1}^{M}, \quad B_i = [T_{s_i}, \ldots, T_{e_i}]$, where s_i and e_i denote the start and end turn indices of block B_i, respectively. This preserves salient evidence together with the local conversational context required for coherent interpretation.

Rather than discarding non-selected context, we preserve the full sequence and inject soft boundary markers to guide attention. For each block $B_i \in \mathcal{B}$, we insert a special marker $m^\star$ at its boundaries: $\mathcal{C}_{\text{anchor}} = \cdots \oplus m^\star \oplus B_i \oplus m^\star \oplus \ldots$. The final attentional state is then $x^{(1)} = [\mathcal{C}_{\text{anchor}}, \mathcal{S}_{\text{str}}]$. These markers softly bias attention toward strategy-aligned evidence while preserving surrounding context, allowing downstream reasoning to access both salient spans and their triggering structure.

4.4 Reflective Abstraction Induction

While reflective grounding specifies *where* the model should attend, the reflective state must also capture *what* the selected evidence jointly implies. To provide high-level supervision during training, we generate an auxiliary reflective summary using a proprietary frontier LLM: $\mathcal{G}_{\text{ref}} = \text{LLM}(q, x^{(1)})$. Appending this summary to the grounded input yields the third-level state $x^{(2)} = [x^{(1)}, \mathcal{G}_{\text{ref}}]$. $\mathcal{G}_{\text{ref}}$ serves as a compact semantic target that captures the latent structure of the grounded evidence, such as stable user preferences, implicit constraints, or cross-event causal patterns. It is used only during training and is omitted at inference time.

4.5 Progressive Reflective Alignment

Given the factual, attentional, and reflective states $x^{(0)}$, $x^{(1)}$, and $x^{(2)}$, we align adjacent levels of the Cognitive Pyramid via progressive distillation. The goal is to transfer attentional grounding and reflective reasoning from higher-level states into the factual pathway rooted in $x^{(0)}$, so that inference requires only the factual state. Let p_i^ℓ denote the temperature-scaled output distribution at output position ℓ under state $x^{(i)}$, where L_y is the target sequence length. For each adjacent transition $i \to i+1$, the bidirectional distillation loss is defined as

$$\mathcal{L}_{\text{dist}}^{(i \to i+1)} = \frac{1}{L_y} \sum_{\ell=1}^{L_y} \Big[\alpha_i\, \text{KL}(p_{i+1}^\ell \,\|\, p_i^\ell) + (1-\alpha_i)\, \text{KL}(p_i^\ell \,\|\, p_{i+1}^\ell)\Big] \tag{6}$$

where $\alpha_i \in [0,1]$ weights the forward and reverse KL terms. To preserve factual accuracy, we further impose hard supervision on the factual state. The overall PRA objective is

$$\mathcal{L}_{\text{hard}} = -\frac{1}{L_y} \sum_{\ell=1}^{L_y} \log p_0(y_\ell^* \mid x^{(0)}, y_{<\ell}^*), \quad \mathcal{L}_{\text{PRA}} = \lambda_{\text{hard}} \mathcal{L}_{\text{hard}} + (1-\lambda_{\text{hard}}) \tau^2 \sum_{i=0}^{1} \gamma_i \mathcal{L}_{\text{dist}}^{(i \to i+1)} \tag{7}$$

Table 1: Main results on RefMem-Bench. We adopt Qwen3-VL-8B as the base model. Blue / red values denote gains/drops over it. Best and second-best results are highlighted in **bold** and underlined, respectively. "–" indicates that the original paper setting does not support the corresponding metric.

Method \ Task	Multi-Choice		Single-Choice		Direct-Answer			
	Acc	MemR	Acc	MemR	Acc	MemR	B1	F1
Base Model	33.2	45.0	44.9	37.8	20.9	37.9	17.4	20.9
- *w/ NaiveRAG*(2020)	33.6 (+0.4)	45.9 (+0.9)	47.6 (+2.7)	38.6 (+0.8)	19.8 (-1.1)	38.7 (+0.8)	17.0 (-0.4)	19.8 (-1.1)
MemGPT(2023)	38.4 (+5.2)	41.4 (-3.6)	47.6 (+2.7)	39.3 (+1.5)	17.3 (-3.6)	34.3 (-3.6)	16.5 (-0.9)	19.3 (-1.6)
Mem0(2025)	42.5 (+9.3)	–	48.5 (+3.6)	–	21.8 (+0.9)	–	17.5 (+0.1)	19.9 (-1.0)
MemoryOS(2025)	41.2 (+8.0)	48.9 (+3.9)	53.0 (+8.1)	40.9 (+3.1)	18.9 (-2.0)	35.6 (-2.3)	18.6 (+1.2)	21.2 (+0.3)
LangMem(2025)	41.2 (+8.0)	–	57.8(+12.9)	–	21.1 (+0.2)	–	19.9 (+2.5)	22.3 (+1.4)
A-Mem(2025)	46.3(+13.1)	48.5 (+3.5)	57.4(+12.5)	44.2 (+6.4)	22.9 (+2.0)	39.6 (+1.7)	16.4 (-1.0)	18.5 (-2.4)
LightMem(2025)	48.3(+15.1)	48.4 (+3.4)	56.9(+12.0)	44.1 (+6.3)	23.2 (+2.3)	39.9 (+2.0)	20.1 (+2.7)	23.2 (+2.3)
GAM(2025a)	47.0(+13.8)	46.2 (+1.2)	56.4(+11.5)	41.4 (+3.6)	23.2 (+2.3)	36.8 (-1.1)	19.9 (+2.5)	23.0 (+2.1)
MemAgent-7B(2025)	25.0 (-8.2)	40.6 (-4.4)	53.5 (+8.6)	43.9 (+6.1)	12.5 (-8.4)	29.7 (-8.2)	12.1 (-5.3)	14.2 (-6.7)
MEM1-7B(2025)	19.8 (-13.4)	33.9 (-11.1)	49.2 (+4.3)	36.9 (-0.9)	12.6 (-8.3)	24.6 (-13.3)	10.7 (-6.7)	12.9 (-8.0)
Qwen3-VL-30B(2025)	36.3 (+3.1)	46.0 (+1.0)	58.2(+13.3)	40.5 (+2.7)	23.5 (+2.6)	40.4 (+2.5)	19.5 (+2.1)	22.4 (+1.5)
DS-Distill-Qwen-14B(2025)	24.6 (-8.6)	43.8 (-1.2)	54.9(+10.0)	45.7 (+7.9)	19.1 (-1.8)	39.1 (+1.2)	15.4 (-2.0)	18.8 (-2.1)
REMIND(Ours)	**59.3**(+26.1)	**58.1**(+13.1)	**66.2**(+21.3)	**52.3**(+14.5)	**32.9**(+12.0)	**55.3**(+17.4)	**27.6**(+10.2)	**30.4** (+9.5)

where y_ℓ^* denotes the ground-truth token at position ℓ, λ_{hard} balances hard supervision and soft distillation, and γ_i controls the contribution of each pyramid transition. By minimizing $\mathcal{L}_{\text{PRA}}$, REMIND progressively distills attentional grounding from $x^{(1)}$ and reflective abstraction from $x^{(2)}$ into the factual pathway $x^{(0)}$, enabling efficient inference using only $x^{(0)}$ at test time.

5 Experiment

5.1 Experimental Setup

Experimental Details. We instantiate REMIND on Qwen3-VL-8B. GPT-4o-mini is used to derive segment-level event cues and structured strategies, GPT-5-mini generates reflective summaries for training only, and all-MiniLM-L6-v2[2] serves as the text encoder, $\phi(\cdot)$ for relevance scoring. We keep the top 20% ranked segments as $\mathcal{C}_{\text{ret}}$ to form $x^{(0)}$. For salience-guided grounding, we use Top-K SAE, with $\alpha = 6$, $\beta = 4$, and $\lambda = 0.3$. For Progressive Reflective Alignment, we set $\tau = 2.0$, $\alpha_0 = \alpha_1 = 0.5$, $\gamma_0 = \gamma_1 = 3.0$, and $\lambda_{\text{hard}} = 0.2$. More details are shown in Appendix E.2.

Baselines. We compare REMIND on our RefMem-Bench against several representative baselines of conversational memory ability. ① **base settings**: Qwen3-VL-8B (Yang et al., 2025) and NaiveRAG (Lewis et al., 2020). ② **training-free memory methods**: MemGPT (Packer et al., 2023), Mem0 (Chhikara et al., 2025), MemoryOS (Kang et al., 2025), LangMem (LangChain, 2025), A-Mem (Xu et al., 2025), LightMem (Fang et al., 2025), and GAM (Yan et al., 2025a). Following common prior-work settings, these methods use GPT-4o-mini[3] as the backbone model. ③ **trained methods**: MemAgent (Yu et al., 2025), MEM1 (Zhou et al., 2025), Qwen3-VL-30B (Yang et al., 2025), and DS-Distill-Qwen-14B (DeepSeek-AI et al., 2025). More details and open-source implementations are provided in Appendix E.1.

5.2 Main Results

As shown in Table 1, REMIND consistently achieves the best performance across all task formats and backbone settings, with gains that reflect a coherent and systematic improvement pattern rather than isolated metric increases. Notably, answer accuracy and memory recall improve simultaneously across settings, suggesting that the model is better at grounding its predictions in relevant evidence instead of relying on superficial correlations. This advantage becomes most evident in the direct-answer setting,

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

[3]GPT-4o-mini refers to `gpt-4o-mini-2024-07-18` unless otherwise specified.

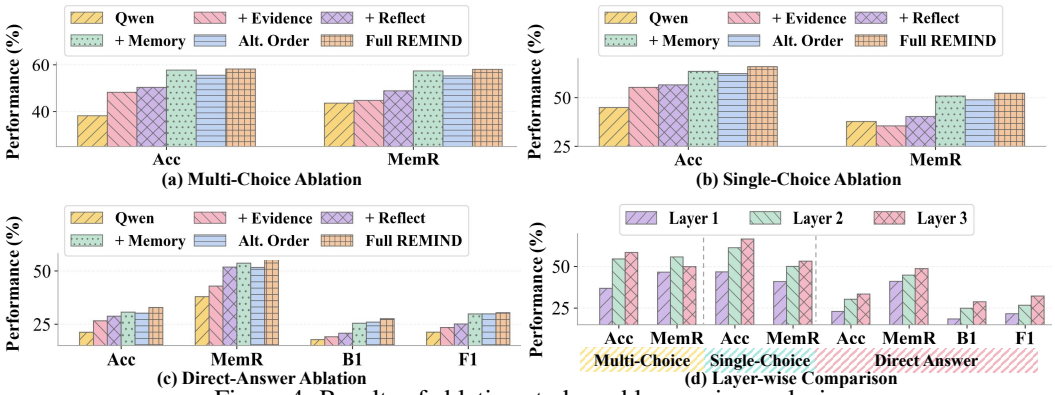


Figure 4: Results of ablation study and layer-wise analysis.

where the model must both retrieve and express reasoning in an open-ended form. In this regime, REMIND shows the largest margins across both retrieval-sensitive and generation-sensitive metrics, with consistent gains jointly observed in MemR, B1, and F1, indicating that improvements in evidence recovery are effectively translated into higher-quality answer generation. Detailed per-dimension results and open-source implementation results are reported in Appendix E.3. These improvements are even more pronounced in the open-source implementation, suggesting that REMIND introduces a general inductive bias that enhances reasoning beyond backbone-specific capabilities. Overall, the results indicate that **REMIND improves not only prediction quality but also the underlying reasoning process by jointly strengthening evidence grounding and abstraction.**

5.3 Ablation Study and Layer Analysis

Ablation Study. To examine the contribution of each stage in REMIND, we conduct ablation experiments by progressively adding modules on top of the bare backbone model (Qwen3-VL-8B), as shown in Figure 4 (a)(b)(c). Adding question-conditioned evidence retrieval (+ *Evidence*) brings substantial gains on both Acc and MemR across all three tasks, showing that question-relevant retrieval provides a strong factual basis for reflective memory reasoning. Further introducing Reflective Strategy (+ *Reflect*) improves accuracy by organizing retrieved evidence in a more question-aware way. Adding Memory Grounding (+ *Memory*) yields more consistent gains, highlighting the role of the attentional level in refining coarse evidence into grounded supporting context. In contrast, *Alt. Order* performs consistently worse, confirming the necessity of the proposed bottom-up hierarchy. Our REMIND achieves the best overall performance by jointly considering these components.

Layer-wise Contribution. To examine whether the three levels of the Cognitive Pyramid provide progressively more useful reflective signals, we perform a layer-wise evaluation on the test set by explicitly constructing the inputs of each layer and feeding them directly into the untuned backbone model without REMIND training. As shown in Figure 4 (d), performance improves steadily from Layer 1 to Layer 3 across all task formats, and the generation quality of direct-answer questions also increases accordingly. This stepwise trend suggests that the hierarchical construction injects increasingly useful reflective information into the backbone, rather than merely reformatting the input. Overall, the results support our core hypothesis that REMIND improves performance by progressively distilling higher-level grounding and abstraction cues into the factual inference pathway.

5.4 Analysis of Complexity and Sensitivity

Complexity Analysis. Existing online memory systems repeatedly invoke the LLM to process the growing dialogue history, update memory, and generate the final response, so their inference cost scales with dialogue length and remains dominated by LLM-based operations. In contrast, REMIND shifts segment-level cue construction offline and reduces test-time inference to lightweight similarity matching over precomputed segment representations followed by a single answer-generation call, $C_{\text{REMIND}} = O(N)C_{\text{sim}} + O(1)C_{\text{LLM}}$. Even with memory grounding, the additional cost is restricted to a small retrieved subset rather than the full dialogue history. This contrast shows that REMIND achieves more favorable long-context efficiency than online memory systems, which is also reflected in Figure 12. Detailed derivations are provided in Appendix E.4.

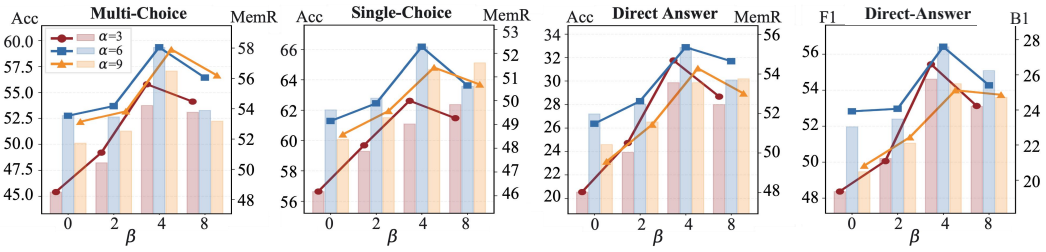


Figure 5: Sensitivity analysis for the hyperparameters α and β. In each subplot, lines correspond to the left-axis metric and bars correspond to the right-axis metric.

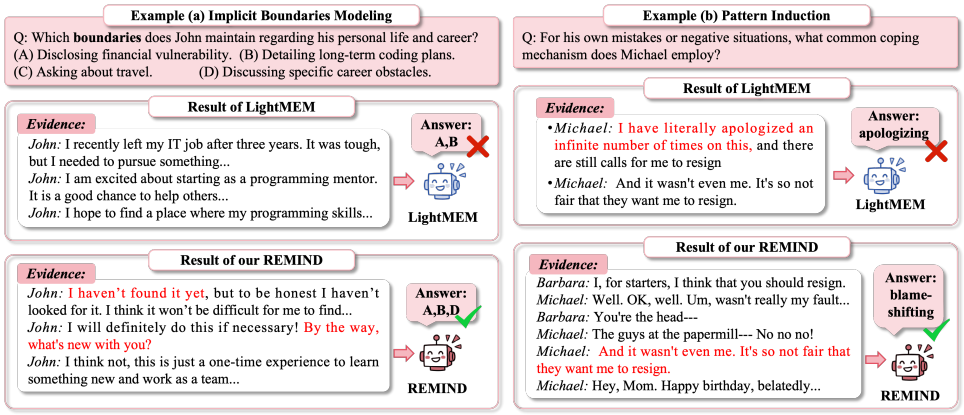


Figure 6: Two examples of the case study.

Sensitivity Analysis. We further examine the robustness of REMIND to the two key hyperparameters α and β in Salience-Guided Memory Grounding. As shown in Figure 5, three tasks exhibit a consistent trend. Performance improves as β increases to a moderate value, but declines when it becomes large, indicating that moderate turn-level broadcasting helps recover implicit triggering context while overly strong broadcasting introduces noise and weakens evidence precision. In contrast, α controls the balance between explicit anchors and broader reasoning cues, with intermediate values yielding the most stable overall behavior across different β settings. The default configuration $(\alpha, \beta) = (6, 4)$ achieves the strongest overall performance, suggesting that reflective memory reasoning benefits from a balanced combination of anchor emphasis and local structural propagation.

5.5 Case Study

To further illustrate the advantage of REMIND over LightMEM, Figure 6 presents two representative reflective memory cases. In *Implicit Boundaries Modeling*, LightMEM retrieves topically relevant but non-diagnostic evidence, whereas REMIND captures more discriminative cues such as avoidance, vulnerability suppression, and topic redirection. In *Pattern Induction*, LightMEM is misled by the surface cue *apologized*, while REMIND identifies the higher-level behavioral pattern of blame-shifting across turns. These examples show that REMIND is more effective at grounding reflective judgments in structurally coherent evidence and abstracting latent patterns from dispersed conversational cues.

6 Conclusion

Existing memory benchmarks mainly focus on explicit recall, yet real-world long-horizon dialogue often requires reflective memory: integrating temporally scattered cues into coherent higher-level interpretations. Therefore, we introduce RefMem-Bench, a benchmark for evaluating reflective memory across diverse reasoning dimensions and task formats, and propose REMIND, a hierarchical framework built on evidence perception, salience-guided grounding, and reflective abstraction. Experiments show that RefMem-Bench is challenging for current models, while REMIND consistently improves both answer accuracy and memory recall, highlighting the promise of hierarchical grounding and abstraction for LLM agents.

AI USE STATEMENT

The core method development in this paper does not involve LLM usage. LLMs were used for question generation during benchmark construction, as described in the paper. We take responsibility for the content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We provide the necessary details in Appendix B.3 and Appendix D.2; all procedures follow the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide the implementation details of both the baselines and our framework in Section 5.1 and Appendix E, including the evaluation setup, training configurations, and LLM prompts. We also provide the source code and LLM prompts in an anonymous repository (see the link in Section 5.1).

REFERENCES

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. Memorybench: A benchmark for memory and continual learning in llm systems. *arXiv preprint arXiv:2510.17281*, 2025.

Joseph W Alba and Lynn Hasher. Is memory schematic? *Psychological Bulletin*, 93(2):203, 1983.

David Badre. Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193–200, 2008. doi: 10.1016/j.tics.2008.02.004.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. *arXiv preprint arXiv:2601.03515*, 2026.

Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, and Mikhail S Burtsev. Scaling transformer to 1m tokens and beyond with rmt. *arXiv preprint arXiv:2304.11062*, 2023.

Zihao Cheng, Weixin Wang, Yu Zhao, Ziyang Ren, Jiaxuan Chen, Ruiyang Xu, Shuai Huang, Yang Chen, Guowei Li, Mengshi Wang, et al. Lifebench: A benchmark for long-horizon multi-source memory. *arXiv preprint arXiv:2603.03781*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948. URL https://arxiv.org/abs/2501.12948.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. Lightmem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*, 2025. doi: 10.48550/arXiv.2510.18866. URL https://arxiv.org/abs/2510.18866.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Junqing He, Liang Zhu, Rui Wang, Xi Wang, Gholamreza Haffari, and Jiaxing Zhang. Madial-bench: Towards real-world evaluation of memory-augmented dialogue generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9902–9921, 2025.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.

Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yeonsu Kwon, Yohan Jo, and Edward Choi. Dialsim: A dialogue simulator for evaluating long-term multi-party dialogue understanding of conversational agents. *arXiv preprint arXiv:2406.13144*, 2024. doi: 10.48550/arXiv.2406.13144. URL https://arxiv.org/abs/2406.13144.

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

LangChain. Langmem SDK for agent long-term memory. https://blog.langchain.com/langmem-sdk-launch/, 2025. Official LangChain blog post.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. Realtalk: A 21-day real-world dataset for long-term conversation. *arXiv preprint arXiv:2502.13270*, 2025.

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. Dialogcc: An automated pipeline for creating high-quality multi-modal dialogue dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1938–1963, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Xinze Li, Ziyue Zhu, Siyuan Liu, Yubo Ma, Yuhang Zang, Yixin Cao, and Aixin Sun. Emembench: Interactive benchmarking of episodic memory for vlm agents. *arXiv preprint arXiv:2601.16690*, 2026.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37: 145–151, 1991. URL https://api.semanticscholar.org/CorpusID:12121632.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

Qingsen Ma, Dianyun Wang, Yaoye Wang, Lechen Ning, Sujie Zhu, Xiaohang Zhang, Jiaming Lyu, Linhao Ren, Zhenbo Xu, and Zhaofeng He. S^3-attention: Attention-aligned endogenous retrieval for memory-bounded long-context inference. *arXiv preprint arXiv:2601.17702*, 2026.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297, 1967.

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

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: towards llms as operating systems. 2023.

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*, pp. 8748–8763, 2021.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024.

Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending memoryllm with scalable long-term memory. *arXiv preprint arXiv:2502.00592*, 2025.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H Chi, et al. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Tingyu Wu, Zhisheng Chen, Ziyan Weng, Shuhe Wang, Chenglong Li, Shuo Zhang, Sen Hu, Silin Wu, Qizhen Lan, Huacan Wang, et al. Knowme-bench: Benchmarking person understanding for lifelong digital companions. *arXiv preprint arXiv:2601.04745*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22477–22503, 2025.

B. Y. Yan, Chaofan Li, Hongjin Qian, Shuqi Lu, and Zheng Liu. General agentic memory via deep research. *arXiv preprint arXiv:2511.18423*, 2025a. doi: 10.48550/arXiv.2511.18423. URL https://arxiv.org/abs/2511.18423.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Liu Yang, Iter Dan, Yichong Xu, Wang Shuohang, Ruochen Xu, and Zhu Chenguang. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv: 2303.16634*, 2023.

Xiaoju Ye. calflops: a flops and params calculate tool for neural networks in pytorch framework. https://github.com/MrYxJ/calculate-flops.pytorch, 2023.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. Memagent: Reshaping long-context LLM with multi-conv RL-based memory agent. *arXiv preprint arXiv:2507.02259*, 2025. doi: 10.48550/arXiv.2507.02259. URL https://arxiv.org/abs/2507.02259.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025.

A Appendix Overview and Organization

This appendix provides supplementary details to support and extend the main paper. The organization of the appendix is as follows:

1. **Limitations, Social Impact, and License & Access (Section B):** This section discusses the limitations, broader impact, ethical considerations, licensing terms, access conditions, and reproducibility commitments of our work.
2. **Additional Related Work (Section C):** This section provides further discussion of prior work on long-term memory benchmarks and long-term memory methods for LLM agents.
3. **Dataset Construction, Annotation, and Evaluation Protocols (Section D):** This section details the construction of RefMem-Bench, including data sources, curation procedures, human annotation, quality control, dataset statistics, and evaluation protocols.
4. **Experimental Setup, Implementation Details, and Additional Analyses (Section E):** This section provides supplementary experimental details, including baselines, implementation settings, detailed results, and additional analysis.

B Limitations, Social Impact, and License & Access

B.1 Limitation

In developing RefMem-Bench, we opted for human annotation of question-answer pairs to ensure higher data quality and richer reflective-memory scenarios. While this approach leads to a more challenging and diverse benchmark, it also introduces certain scalability constraints, as further expanding the dataset requires substantial expert effort. Additionally, during the annotation process, we intentionally avoided rigid templates to encourage greater variety and naturalness in the questions. Although this design makes the benchmark more realistic and expressive, it also makes it more challenging to precisely control the difficulty level across all instances.

For REMIND, we design the Cognitive Pyramid to support goal-directed reasoning tasks that can be naturally organized through evidence perception, salience grounding, and higher-level abstraction. While this structure is well aligned with reflective-memory reasoning and may also benefit broader tasks such as question answering, math, personalized dialogue, adapting it to highly open-ended tasks with less explicit goals may require additional task-specific design.

Despite these inherent trade-offs, we believe that our dataset construction strategy enables RefMem-Bench to better reflect the complexity of long-horizon reflective memory, and that REMIND provides a useful framework for studying how models transform retrieved evidence into abstraction-level reasoning.

B.2 Broader Impact

RefMem-Bench and REMIND may bring both benefits and risks to the research community and broader society. On the positive side, RefMem-Bench provides a challenging testbed for reflective memory in long-horizon multimodal dialogue, and may support progress in personalized assistants, long-term dialogue agents, and other systems that require reliable reasoning over extended interaction histories. REMIND further supports this direction by transforming retrieved evidence into higher-level reflective reasoning.

However, stronger long-term memory capabilities may also raise privacy and fairness concerns, especially when models infer user states, preferences, or behavioral patterns from personal histories. To this end, we encourage researchers and practitioners to use RefMem-Bench and REMIND with appropriate safeguards, ethical data practices, and careful consideration of downstream deployment risks. Meanwhile, we acknowledge the computational cost of evaluating modern LLMs and MLLMs. By releasing our resources publicly, we aim to improve reproducibility and reduce redundant computation. Overall, we welcome constructive feedback and ongoing dialogue to help improve RefMem-Bench and REMIND and mitigate emerging risks.

B.3 Ethics Statement, License and Access

As this research exclusively employs publicly accessible pre-trained models and compensates annotators through biweekly or monthly salaries that cover the working hours dedicated to annotation, it does not present additional ethical concerns. We will release RefMem-Bench under the Creative Commons Attribution 4.0 International License (CC BY 4.0), and require proper disclosure when the dataset is used for model evaluation. This license does not override the original licenses of the source materials, and users must comply with all relevant legal and ethical requirements regarding data subjects. Although we make efforts to ensure the quality and legality of the benchmark, we do not guarantee absolute completeness or correctness, and assume no responsibility for legal or ethical issues arising from misuse, including copyright infringement, privacy violations, or misuse of sensitive information.

B.4 Reproducibility Statement

Our experiments follow baseline configurations from prior work. Implementation details are provided in the Appendix. To support reproducibility, we have anonymously released the code and dataset at this anonymous repository. This work complies with the ICLR Reproducibility Requirements.

C Related Work

Additional Discussion on Long-Term Memory Benchmarks. Despite rapid progress in long-term memory benchmarks for LLM agents, robust evaluation of reflective memory remains limited. Most existing tasks still focus on whether models can recall, update, or apply explicitly stored information (e.g., LoCoMo (Maharana et al., 2024), MADial (He et al., 2025), and LongMemEval (Wu et al., 2024)), rather than whether they can synthesize distributed evidence into higher-level interpretations. Other benchmarks, such as MemBench (Tan et al., 2025), MemoryAgentBench (Hu et al., 2025), MemoryBench (Ai et al., 2025), and Evo-Memory (Wei et al., 2025), evaluate memory from the perspective of operation and adaptation, focusing on storage, updating, continual learning, and test-time memory evolution. Furthermore, recent benchmarks such as REALTALK (Lee et al., 2025), MMRC (Xue et al., 2025), PersonaMem (Jiang et al., 2025), KnowMe-Bench (Wu et al., 2026), EMemBench (Li et al., 2026), Mem-Gallery (Bei et al., 2026), and LifeBench (Cheng et al., 2026) shift evaluation toward realistic users, personalization, and multimodality, but still provide limited support for evaluating higher-level reflective inference over distributed evidence. In contrast, RefMem-Bench is a human-curated benchmark, carefully annotated by three memory-focused experts. It provides comprehensive coverage of eight reflective-memory dimensions and three task formats, requiring models to integrate distributed evidence from extended interaction histories to infer latent user states, behavioral regularities, temporal evolution, and implicit constraints. This benchmark reveals the substantial gap between current LLM agents and human-level reflective memory, and sets a higher standard for evaluating long-horizon memory reasoning in multimodal dialogue scenarios.

Additional Discussion on Long-Term Memory Methods. Existing long-term memory methods for LLM agents mainly improve long-horizon interaction by introducing external memory management, lightweight memory consolidation, or adaptive context construction. Representative systems such as MemGPT (Packer et al., 2023), A-Mem (Xu et al., 2025), Mem0 (Chhikara et al., 2025), and MemoryOS (Kang et al., 2025). Recent works such as LightMem (Fang et al., 2025) and GAM (Yan et al., 2025a) further improve efficiency through lightweight memory consolidation and just-in-time context construction. These methods improve the scalability, organization, and efficiency of long-term memory, but still mainly treat memory as an external resource to be managed. Another line of work, such as MEM1 (Zhou et al., 2025) and MemAgent (Yu et al., 2025), uses reinforcement-learning-based policies to optimize memory updating and long-context reasoning, but remains focused on memory operation efficiency rather than abstraction-level reflective reasoning. In contrast, REMIND organizes reasoning through factual, attentional, and reflective states, and uses Progressive Reflective Alignment to distill high-level reflective reasoning into the factual inference pathway, enabling models to transform retrieved traces into abstraction-level reasoning while maintaining efficiency.

D DATASET

D.1 DETAILS OF THE DATA CURATION

Data Collection. We construct RefMem-Bench from three long-horizon conversational sources: REALTALK (Lee et al., 2025), DialSim (Kim et al., 2024), and LoCoMo (Maharana et al., 2024). These sources are chosen because they provide complementary coverage of long-range interaction phenomena. REALTALK contributes realistic conversational trajectories together with speaker-level visual sharing history, which makes it suitable for multimodal reflective-memory evaluation. DialSim and LoCoMo provide long-range dialogue structure with temporally distributed topics, recurring behavioral signals, and cross-session dependencies, which are useful for text-centric reflective-memory construction. During collection, we retain complete conversational threads rather than isolated local snippets, so that each benchmark instance can be grounded in an extended interaction history and, when available, aligned visual evidence.

Feature Clustering. After collection, we organize the source material through separate text-centric and image-centric clustering pipelines. For the text-centric branch, we first segment multi-turn conversations into fine-grained topical units, so that long conversational histories can be decomposed into semantically coherent local segments. We then group related topics with Agglomerative Clustering (Johnson, 1967; Ward, 1963), which provides a flexible way to aggregate semantically similar segments without requiring a fixed number of groups in advance. Because some clusters remain overly broad after hierarchical grouping, we further refine large clusters with K-Means (MacQueen, 1967; Lloyd, 1982) to improve granularity and reduce topic mixing. For the image-centric branch, we collect speaker-level image histories from REALTALK, encode them with CLIP (Radford et al., 2021), and cluster the resulting image representations with Agglomerative Clustering. This process yields visually coherent image groups that can be aligned with their surrounding conversational context for multimodal question construction.

Question Generation. We generate benchmark items from the clustered dialogue and image groups by focusing on reflective-memory reasoning rather than surface-level fact retrieval. Considering both generation quality and the cost of large-scale data construction, we use Gemini-2.5-Pro for question generation. This model is used only for benchmark construction and belongs to a different model family from that used in our method. In the text-centric branch, clustered dialogue segments are used to construct questions that require reasoning over temporally distributed evidence, recurring behavior, latent user attitudes, experience-conditioned future actions, and hidden interpersonal or personal boundaries. Concretely, this branch is used to generate questions for **Temporal Dynamics**, **Pattern Induction**, **Belief Modeling**, **Experience-Grounded Planning**, and **Implicit Boundary Modeling**. In the image-centric branch, clustered visual histories and their aligned dialogue context are used to generate questions that require models to integrate textual and visual signals, identify user-specific visual regularities, and abstract topic-level visual prototypes. This branch is used to construct questions for **Cross-Modal Latent Inference**, **Personalized Visual Baselines**, and **Topic-Scoped Visual Prototypes**. For direct-answer items, the generation model additionally produces a list of candidate answers to facilitate subsequent human annotation. Across both branches, each generated item is paired with the supporting evidence anchors so that the intended reasoning path remains traceable.

Human Annotation and Verification. To ensure annotation quality, three expert annotators conduct the main annotation process. Given each question and its supporting evidence anchors, the annotators provide answer annotations. For direct-answer items, annotators vote on the generated candidate answers and may additionally provide their own answers when they consider none of the candidates fully correct. They also verify whether each answer is grounded in the underlying dialogue and, when applicable, the aligned visual context. During this process, they correct or flag errors introduced by LLM-based generation. They also filter out low-quality questions and revise cases with ambiguous wording, insufficient specification, or weak grounding in the source material, resulting in the removal of 8.31% of candidate questions. This stage serves three purposes: (1) producing reliable answer annotations; (2) preserving traceability between each question and the evidence required to answer it; and (3) ensuring sufficient reasoning depth, diversity, and grounding. All the annotators are properly briefed with the annotation objective, and a discussion among them is conducted to resolve uncertain

cases. Among items with inconsistent annotations, 5.79% are further discussed and revised through consensus with respect to both their evidence and answers.

D.2 Data Annotation And Quality Control Interfaces

The interfaces for data annotation and quality control are illustrated in Figures 7 and 8. For multi-choice and single-choice items, annotators review the dialogue context and question, select the correct answer option(s), verify or revise the supporting evidence, and assign confidence scores for quality control. For direct-answer items, annotators additionally vote on candidate answers or provide a free-text answer when none of the candidates is fully satisfactory.

D.3 General Guidelines

RefMem-Bench is designed to evaluate reflective-memory reasoning over long-horizon personal dialogues and aligned visual histories. The following guidelines govern the data annotation process:

- **Reflective Reasoning:** All questions focus on reflective-memory reasoning rather than surface-level fact retrieval. Each item requires reasoning over long-range dialogue history, behavioral patterns, user attitudes, or experience-conditioned actions.
- **Dimension Coverage:** For questions in each dimension, annotators assess their quality according to the corresponding dimension definition, ensuring clear distinctions among questions from different reasoning dimensions.
- **Bias Mitigation:** The annotation interface hides the model-generated reference answer from annotators. Annotators make judgments based on the dialogue context, evidence anchors, and their own reasoning.
- **Difficulty, Diversity, and Clarity:** Questions must be clear, answerable by humans, and sufficiently challenging. Annotators are encouraged to remove or revise vague, repetitive, or weakly grounded items.
- **Multimodal Consistency:** For image-centric items, annotators verify that the visual context provides necessary evidence and that the question requires cross-modal reasoning.

D.4 Quality Control and Validation

A rigorous quality control procedure is implemented to ensure the reliability and validity of the benchmark:

- **Independent Annotation:** All generated items are independently annotated by three in-house expert annotators with more than three years of NLP research experience.
- **Evidence Verification:** Annotators verify whether each evidence anchor is grounded in the underlying dialogue and, where applicable, the aligned visual context. Unreliable evidence is flagged and discarded. Missing evidence can also be supplemented to preserve the traceability of the intended reasoning path.
- **Quality Filtering:** Annotators remove or revise questions with ambiguous wording, insufficient specification, weak grounding, or limited reasoning depth. This step ensures that the retained questions are answerable, well grounded, and sufficiently diverse.
- **Inter-Annotator Agreement:** Before discussion and final resolution, we compute Fleiss' κ (Fleiss, 1971) over the three annotators' initial decisions across all question types. The resulting score, $\kappa = 0.849$, indicates strong inter-annotator agreement under the interpretation of Landis & Koch (1977).
- **Disagreement Resolution:** Items with inconsistent annotations undergo structured discussion among all three annotators. Both evidence anchors and answers are jointly revised until consensus is reached.

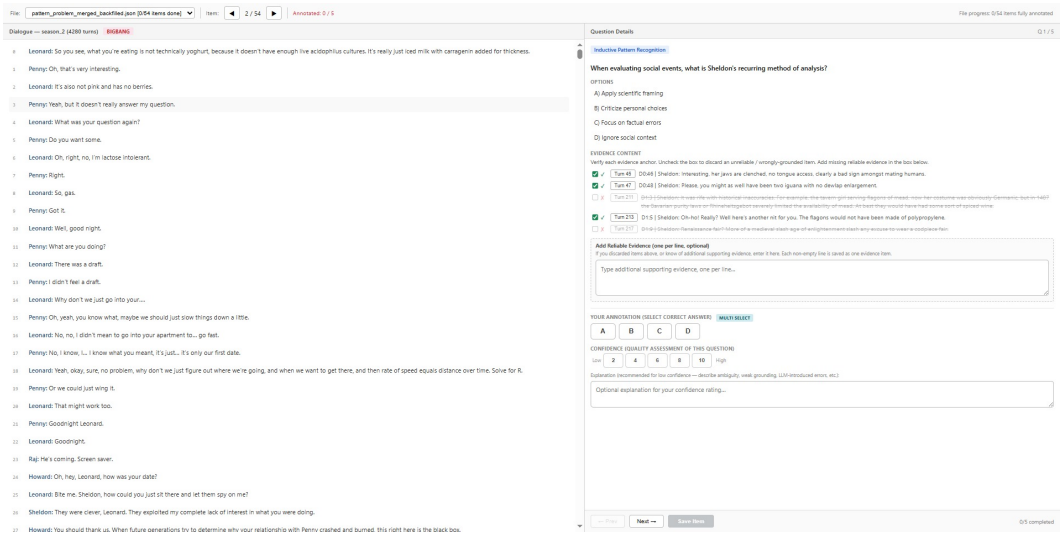


Figure 7: Screenshot of the multi-choice and single-choice annotation interface. Annotators review the dialogue context and question, select the answer option(s), verify or revise supporting evidence, and provide confidence scores for quality control.

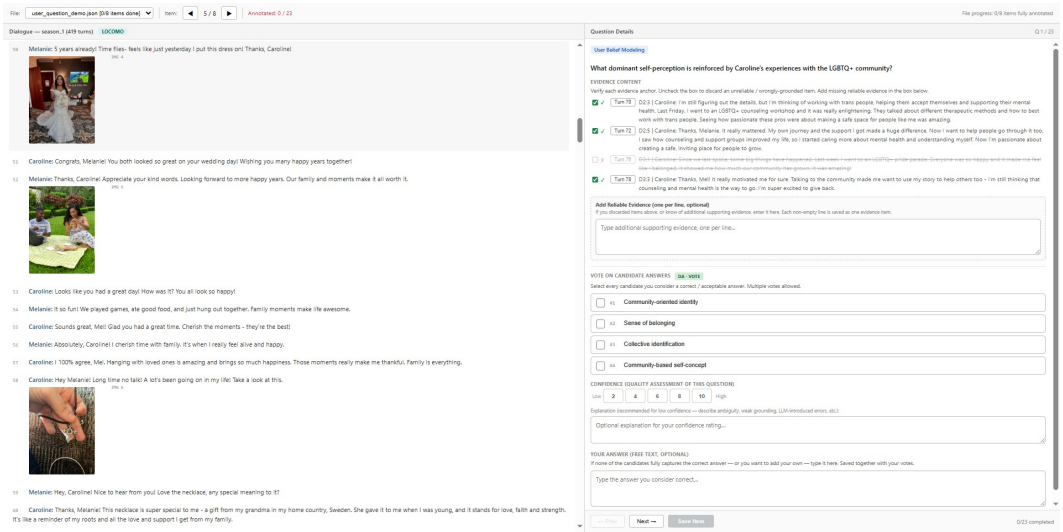


Figure 8: Screenshot of the direct-answer annotation interface. Annotators review the dialogue context and question, verify or revise supporting evidence, vote on candidate answers or provide a free-text answer, and assign confidence scores for quality control.



Figure 9: Visualization of our benchmark.

D.5 Reflective-Memory Dimensions

As shown in Figure 9, RefMem-Bench is organized around eight reflective-memory dimensions. These dimensions are designed to cover complementary forms of abstraction over long-horizon interactions, including temporal state tracking, latent user modeling, behavior-level regularity induction, planning from accumulated experience, and multimodal reflective reasoning.

Temporal Dynamics. This dimension evaluates whether a model can track how relevant states evolve over time. In long-horizon interactions, the correct interpretation of a user, situation, or relationship is often not determined by a single utterance or by the most recent mention alone. Instead, the model must integrate evidence distributed across turns and sessions to recover how attitudes,

constraints, preferences, or circumstances shift over time. Questions in this dimension therefore test whether the model can reason over trajectories of change rather than static snapshots.

Pattern Induction. This dimension tests whether a model can abstract a stable behavioral regularity from multiple, temporally separated observations. The target is not an explicitly stated fact, but a pattern-level summary inferred from repeated instances. In many cases, the relevant evidence is scattered across distant and seemingly unrelated moments, and the answer does not appear verbatim in the source dialogue. Questions in this dimension thus assess whether the model can move from fragmented episodic evidence to a higher-level rule or recurring tendency.

Belief Modeling. This dimension evaluates whether a model can infer a coherent portrait of a user's values, preferences, decision style, or general orientation from repeated interactions. Rather than asking what the user literally said about themselves, these questions ask what kind of person the user appears to be on the basis of their accumulated behavior. The emphasis is on stable latent orientation, such as approval-seeking, risk sensitivity, or preference structure, instead of isolated profile attributes. This makes the dimension a form of user-level reflective abstraction rather than simple persona extraction.

Experience-Grounded Planning. This dimension probes whether a model can use remembered experience to infer plausible future actions, next steps, or recommendations. The answer depends on projecting a previously established behavioral tendency, preference, or practical constraint into a new or prospective situation. Unlike retrieval-style memory questions, these items require the model to convert remembered experience into an action-relevant plan or anticipatory judgment. Success in this dimension therefore depends on whether the model can ground planning in prior interaction history rather than generate generic advice.

Implicit Boundary Modeling. This dimension captures whether a model can infer hidden boundaries or sensitive triggers behind repeated avoidance, deflection, resistance, or conversational shutdown. In many dialogues, important user constraints are not explicitly declared; instead, they must be inferred from recurring negative reactions or consistent evasive behavior. Questions in this dimension therefore test whether the model can identify the latent boundary condition underlying multiple observable responses, rather than merely recalling a single avoided topic or utterance.

Cross-Modal Latent Inference. This dimension evaluates whether a model can infer an unstated capability, state, preference, or background condition by jointly reasoning over dialogue and visual evidence. Neither modality alone is sufficient to determine the answer: the target interpretation emerges only when text and image are considered together. This includes cases where the image clarifies, enriches, or partially contradicts what is said in the dialogue, thereby revealing a deeper implication not explicitly verbalized. The dimension is designed to test cross-modal reflective reasoning rather than simple image grounding.

Personalized Visual Baselines. This dimension requires the model to abstract a speaker-specific visual baseline from repeated image-sharing behavior. Instead of reasoning about a single image in isolation, the model must infer what counts as that user's typical visual style, composition, or habitual way of documenting a recurring kind of experience. The goal is to recover a personalized visual norm from longitudinal evidence. Questions in this dimension therefore test whether the model can learn stable individual-level regularities from a history of shared images.

Topic-Scoped Visual Prototypes. This dimension evaluates whether a model can distill a recurring visual prototype associated with a particular topic. Compared with Personalized Visual Baselines, which are centered on a speaker's general visual tendencies, this dimension is more topic-specific: the model must identify the characteristic visual pattern associated with a given class of content, such as meals, outings, or locations, as it appears across a user's history. The target is thus a topic-conditioned visual abstraction grounded in repeated multimodal evidence.

D.6 Dataset Statistics of our RefMem-Bench

Figure 2 and Table 2 provide a more detailed view of the scale and composition of RefMem-Bench. At the context level, the benchmark contains 35 curated long-horizon conversational contexts, covering 1,341 sessions, 5,604 topic segments, and 71,062 dialogue turns in total. These contexts include 1,124 images and 958,572 words, corresponding to an average of 38.31 sessions, 2,030.34 turns, 27,387.77 words, and 32.11 images per context, with 52.99 turns per session on average. Such statistics indicate that RefMem-Bench is not only large in the number of evaluation instances, but also substantially dense in contextual evidence, making shallow recency-based strategies insufficient.

At the QA level, RefMem-Bench consists of 25,573 question-answer pairs, split into 5,965 training instances and 19,608 test instances. To reduce potential leakage, the split is conducted at the cluster-derived level rather than at the individual QA level. Specifically, all questions generated from the same semantic or visual evidence cluster are assigned entirely to either the training set or the test set. During annotation, we further manually verify the evidence grouping to avoid potential leakage from near-duplicate questions or shared local evidence across splits. After this split, the benchmark includes three task formats: 8,444 multi-choice questions, 6,180 single-choice questions, and 10,949 direct-answer questions. The dimensional distribution is broad rather than concentrated in a single reasoning type: Implicit Boundary Modeling accounts for 4,718 instances (18.4%), Belief Modeling for 4,394 (17.2%), Pattern Induction for 4,390 (17.2%), Temporal Dynamics for 4,311 (16.9%), Experience-Grounded Planning for 3,630 (14.2%), Cross-Modal Latent Inference for 2,286 (8.9%), Topic-Scoped Visual Prototypes for 1,050 (4.1%), and Personalized Visual Baselines for 794 (3.1%). In addition, 4,130 questions are derived from multimodal real-world conversations, ensuring that the benchmark tests not only text-based long-range abstraction but also image-grounded reflective reasoning. As also shown in Figure 2, this combination of scale, multimodality, real-world grounding, evidence anchoring, and full eight-dimension coverage distinguishes RefMem-Bench from existing dialogue resources and prior memory benchmarks.

D.7 Evaluation Protocols.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the evaluation set. We assess model performance with four metrics: answer accuracy, text-based memory recall, visual memory recall, and, for direct-answer questions, BLEU-1 and token-level F1. The overall answer accuracy is defined as

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} s_i, \tag{8}$$

where

$$s_i = \begin{cases} \mathbf{1}(P_i = G_i), & \text{for option-based tasks,} \\ J(\hat{y}_i, y_i), & \text{for direct-answer tasks,} \end{cases} \tag{9}$$

with P_i and G_i denoting the predicted and gold option sets, respectively, $\mathbf{1}(\cdot)$ the indicator function, and $J(\hat{y}_i, y_i) \in \{0, 1\}$ the LLM-as-a-judge verdict for open-ended answers. To quantify evidence recovery, we further define memory recall. For text-dominant dimensions, let E_i^g and E_i^p denote the gold and predicted evidence sets for sample i, let $\phi(\cdot)$ be the sentence encoder, and let $\cos(\cdot, \cdot)$ denote cosine similarity. The sample-level and overall text-based memory recall are defined as

$$\mathrm{MR}_i^{\text{text}} = \frac{1}{|E_i^g|}\sum_{e \in E_i^g} \max_{\hat{e} \in E_i^p} \cos\big(\phi(e), \phi(\hat{e})\big), \qquad \mathrm{MR}^{\text{text}} = \frac{1}{|\mathcal{V}|}\sum_{i \in \mathcal{V}} \mathrm{MR}_i^{\text{text}}. \tag{10}$$

where $\mathcal{V}$ denotes the set of valid samples whose gold and predicted evidence are both non-empty. For visually grounded dimensions, let G_i^v and P_i^v denote the gold and predicted image-index sets for sample i. Similarly, the sample-level and overall visual memory recall are defined as

$$\mathrm{MR}_i^{\text{vis}} = \frac{|G_i^v \cap P_i^v|}{|G_i^v|}, \qquad \mathrm{MR}^{\text{vis}} = \frac{1}{|\mathcal{V}_{\text{vis}}|}\sum_{i \in \mathcal{V}_{\text{vis}}} \mathrm{MR}_i^{\text{vis}}. \tag{11}$$

where $\mathcal{V}_{\text{vis}}$ is the set of valid visual samples with non-empty gold image indices. For direct-answer evaluation, we additionally report BLEU-1 and token-level F1. Let T_p and T_r denote the tokenized

Table 2: Overall statistics of RefMem-Bench. We treat all sources as a unified benchmark and report aggregate conversational and QA-level statistics.

Conversation Statistics	
Organized contexts	35
Sessions	1,341
Topic segments	5,604
Dialogue turns	71,062
Images	1,124
Words	958,572
Avg. sessions / context	38.31
Avg. turns / context	2,030.34
Avg. turns / session	52.99
Avg. words / context	27,387.77
Avg. words / turn	13.49
Avg. images / context	32.11
QA Benchmark Statistics	
Total questions	25,573
Multi-Choice (MC)	8,444
Single-Choice (SC)	6,180
Direct Answer (DA)	10,949
Avg. options / MC-SC item	4.00
Avg. answer cardinality	2.71
Avg. evidence anchors / item	4.08
Dimension Distribution	
Temporal Dynamics	4,311 (16.9%)
Pattern Induction	4,390 (17.2%)
Belief Modeling	4,394 (17.2%)
Experience-Grounded Planning	3,630 (14.2%)
Implicit Boundary Modeling	4,718 (18.4%)
Cross-Modal Latent Inference	2,286 (8.9%)
Personalized Visual Baselines	794 (3.1%)
Topic-Scoped Visual Prototypes	1,050 (4.1%)

prediction and reference, respectively, and let $c_p(w)$ and $c_r(w)$ denote the occurrence counts of token w. Defining unigram overlap as

$$\text{Overlap} = \sum_{w} \min\big(c_p(w), c_r(w)\big), \tag{12}$$

BLEU-1 is computed by

$$P_1 = \frac{\text{Overlap}}{|T_p|}, \qquad \text{BP} = \begin{cases} 1, & |T_p| \geq |T_r|, \\ \exp\Big(1 - \frac{|T_r|}{|T_p|}\Big), & |T_p| < |T_r|, \end{cases} \qquad \text{BLEU}_1 = \text{BP} \cdot P_1, \tag{13}$$

while token-level F1 is defined as

$$P = \frac{\text{Overlap}}{|T_p|}, \qquad R = \frac{\text{Overlap}}{|T_r|}, \qquad \text{F1} = \frac{2PR}{P+R} = \frac{2 \cdot \text{Overlap}}{|T_p| + |T_r|}. \tag{14}$$

D.8 Evaluation Metric Building

To accurately evaluate the diverse open-ended responses produced by different models, we use an expert-written prompt to instruct GPT-4o as the correctness judge. We present the full prompt in Figure 10. To select a reliable and cost-efficient evaluator, we conduct a meta-evaluation against expert human annotations. We consider two candidate judges, GPT-4o and Qwen3-8B-Instruct. We sample 30 questions for each dimension and report the judgment correctness by category.

As shown in Table 3, the prompt-engineered GPT-4o judge achieves stronger overall agreement with human annotations, with an average evaluation accuracy of 0.99 across the eight dimensions.

You are a semantic equivalence judge for a direct-answer question answering task.
Task: Given a question, a list of acceptable reference answers, and a model's predicted answer, determine whether the model's prediction is semantically equivalent to ANY ONE of the acceptable reference answers.
Judgment Criteria:
1. Semantic equivalence means the prediction and at least one reference answer refer to the same underlying concept, entity, action, or phenomenon, even if different words, synonyms, abbreviations, or paraphrases are used.
2. Minor grammatical differences, such as singular versus plural, gerund versus infinitive, or articles, should not cause rejection.
3. The prediction does not need to match any reference answer verbatim; conceptual alignment is sufficient.
4. If the prediction refers to a different concept, a broader or narrower scope that changes the meaning, or an unrelated idea, it must be rejected.
5. If the prediction is empty, nonsensical, or clearly irrelevant, reject it.
Output Format:
Return only a JSON object with a single boolean field "verdict".
{"verdict": true} if the prediction is semantically equivalent to any reference answer.
{"verdict": false} if the prediction does not match any reference answer.
No extra text, no explanation, and no markdown fences. Return only the JSON object.
Question: {question}
Acceptable Reference Answers: {acceptable reference answers}
Model Prediction: {model's predicted answer}

Figure 10: Prompt template used by the LLM-as-a-judge for direct-answer evaluation in RefMem-Bench. The judge determines whether a model prediction is semantically equivalent to any acceptable reference answer and returns a JSON verdict only.

In comparison, Qwen3-8B-Instruct obtains an average accuracy of 0.93, indicating that it is also reasonably reliable but still less aligned with expert judgments. These results suggest that GPT-4o provides a more stable and accurate judge for our direct-answer evaluation. We will release the full evaluation prompt with the benchmark to enable consistent comparisons in future work.

Table 3: Meta-evaluation results of the prompt-engineered GPT-4o judge, showing high evaluation accuracy across all dimensions.

Models	TD	PI	BM	EGP	IBM	CMLI	PVB	TSVP	Avg
GPT-4o	1.00 (30/30)	1.00 (30/30)	1.00 (30/30)	1.00 (30/30)	0.97 (29/30)	0.97 (29/30)	1.00 (30/30)	1.00 (30/30)	0.99
Qwen3-8B-Instruct	0.90 (27/30)	0.90 (27/30)	1.00 (30/30)	0.83 (25/30)	1.00 (30/30)	1.00 (30/30)	0.90 (27/30)	0.93 (28/30)	0.93

E EXPERIMENTS

E.1 BASELINES

We compare **REMIND** on our RefMem-Bench against several representative baselines of conversational memory ability. ① **Base settings**: Qwen3-VL-8B (Yang et al., 2025) and Qwen3-VL-8B with NaiveRAG (Lewis et al., 2020). ② **Training-free memory methods**: MemGPT (Packer et al., 2023), Mem0 (Chhikara et al., 2025), MemoryOS (Kang et al., 2025), LangMem (LangChain, 2025), A-Mem (Xu et al., 2025), LightMem (Fang et al., 2025), and GAM (Yan et al., 2025a). These approaches augment a frozen backbone model with explicit memory storage, retrieval, summarization, or update mechanisms, without task-specific training on RefMem-Bench. ③ **Trained methods**: MemAgent (Yu et al., 2025), MEM1 (Zhou et al., 2025), Qwen3-VL-30B (Yang et al., 2025), and DS-Distill-Qwen-14B (DeepSeek-AI et al., 2025). Compared with training-free methods, these baselines either learn memory-oriented agent behaviors or benefit from stronger post-trained reasoning capabilities. This setup allows us to compare explicit memory systems against powerful trained

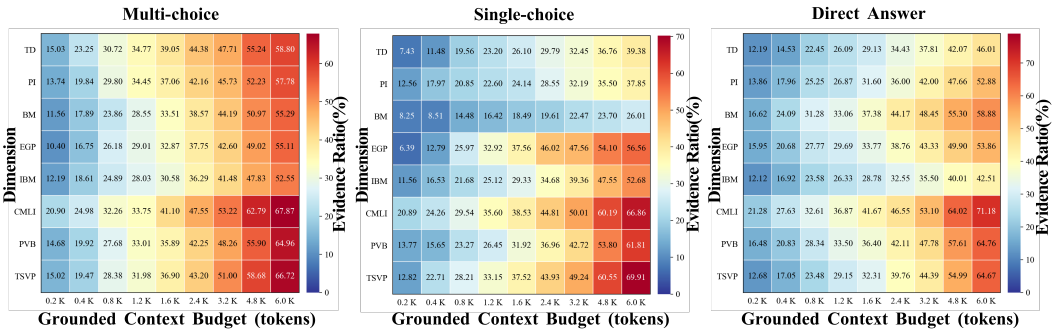


Figure 11: Evidence Ratio of salience-guided memory grounding across grounded-context budgets. The ratio measures the fraction of gold evidence covered by the strategy-salient context used to construct the attentional state.

LLMs under a unified reflective-memory evaluation protocol.For all memory baselines, we keep their memory retrieval and construction settings consistent with the original papers, including the parameter configurations and embedding models. During inference, we set the temperature to 0.0 for all models. The prompts are provided in Appendix E.6.

E.2 Implementation Details

We build REMIND on top of Qwen3-VL-8B. GPT-4o-mini is used as the auxiliary LLM to derive segment-level event cues and generate the structured strategy, while GPT-5-mini[4] serves as the frontier LLM and produces the reflective summaries used only during training. The text encoder $\phi(\cdot)$ underlying the relevance score η_n is instantiated with all-MiniLM-L6-v2. We retain the top-ranked 20% of segments as $\mathcal{C}_{\text{ret}}$ to form the factual state $x^{(0)}$, which preserves broad evidence coverage while filtering out redundant history.

For Top-K SAE training, we follow Ma et al. (2026). We set the latent dimension to $d_{\text{latent}} = 4096$ and the sparsity parameter to $k = 32$. The SAE is mounted on four representative layers of Qwen3-VL-8B. We pretrain it on DialogCC (Lee et al., 2024), whose multi-turn multimodal dialogues match our memory-grounded setting. DialogCC shares no dialogue content with our benchmark. This alignment helps the sparse dictionary capture dialogue-specific features, including cross-turn coreference and visual anchoring. We optimize the SAE with AdamW using a learning rate of 5×10^{-5}. For salience-guided grounding, we build on the learned Top-K SAE sparse space. We set the anchor boosting coefficient to $\alpha = 6$, the speaker-conditioned broadcast coefficient to $\beta = 4$, and the modality fusion weight to $\lambda = 0.3$. Each selected turn is expanded with a neighborhood window of $w = 1$. Figure 11 reports the Evidence Ratio under filtered-context budgets from 0.2K to 6.0K tokens. This metric measures how much gold evidence is covered by the strategy-salient context used to construct $x^{(1)}$. The consistently high ratios show that the SAE-based salience signal provides stable grounding, supporting our parameter choice and SAE training strategy. Moreover, the ratio increases rapidly up to 3.2K and then becomes nearly saturated. We therefore set the salience threshold δ to keep a 3.2K-token grounded context, which offers a practical trade-off between evidence coverage and context efficiency.

For progressive reflective alignment, we set the distillation temperature to $\tau = 2.0$. The forward and reverse KL terms are balanced equally by $\alpha_0 = \alpha_1 = 0.5$ for both pyramid transitions $x^{(0)} \to x^{(1)}$ and $x^{(1)} \to x^{(2)}$. The transition weights $\gamma_0 = \gamma_1 = 3.0$ give attentional grounding and reflective abstraction comparable influence over the factual pathway. The hard–soft balance coefficient is set to $\lambda_{\text{hard}} = 0.2$, so that cross-level distillation carries the main learning signal while $x^{(0)}$ remains anchored to the ground-truth sequence.We optimize the model with AdamW using a learning rate of 1×10^{-5}. All experiments are conducted on a single node with $2\times$ NVIDIA A100 GPUs (40 GB each), and the training batch size is set to 2. All reported results are obtained with decoding temperature set to 0.0. Detailed prompts are provided in Appendix E.6.

[4]The specific version is `gpt-5-mini-2025-08-07`.

E.3 Detailed Results on the Benchmark

In this section, we provide dimension-wise results on RefMem-Bench. Tables 4 and 6 report the main experimental results. Tables 5 and 7 report the corresponding open-source implementation results.

Table 4: Detailed results of the main experiments on RefMem-Bench. We report answer accuracy (Acc) and memory recall (MemR) for Multi-Choice, Single-Choice, and Direct-Answer settings. "–" indicates that the original paper setting does not support the corresponding evaluation protocol.

Models	TD		PI		BM		EGP		IBM		CMLI		PVB		TSVP	
	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR
Multi-Choice																
Base Model	36.88	48.78	47.44	50.86	42.31	47.43	24.38	48.62	25.55	46.54	24.38	39.56	27.85	39.24	36.64	38.66
- w/ NaiveRAG	37.47	50.45	47.71	51.32	40.44	49.93	23.80	51.89	27.20	47.54	25.21	37.84	28.48	38.34	38.17	39.70
MemGPT	44.64	51.44	36.32	44.16	31.20	42.77	26.12	42.50	33.79	39.13	37.12	35.05	46.20	39.56	51.91	36.31
Mem0	38.14	44.19	33.72	44.11	25.33	43.72	28.01	44.76	31.32	43.06	55.12	–	**63.29**	–	**64.89**	–
MemoryOS	42.78	53.87	39.37	53.59	37.42	50.15	28.30	53.95	32.14	51.81	40.44	38.23	51.90	40.29	57.25	49.40
LangMem	45.49	47.17	39.28	46.42	37.78	44.34	34.11	45.55	36.17	43.29	40.44	–	44.94	–	51.15	–
A-mem	52.41	52.42	46.46	54.72	43.29	**50.40**	36.28	52.33	37.64	45.60	48.48	48.65	53.16	35.51	52.67	48.00
Light-mem	55.19	50.55	53.54	50.04	50.76	49.53	33.67	51.24	34.43	48.38	48.20	46.73	53.16	42.17	57.25	48.77
GAM	53.33	49.93	59.73	53.07	44.36	50.27	36.14	51.18	37.64	47.41	43.77	39.42	48.73	38.53	51.91	39.56
MemAgent-7B	30.38	45.30	24.13	45.86	25.87	45.07	23.80	45.86	28.02	43.20	30.19	39.67	13.92	29.00	23.66	30.88
MEM1-7B	22.28	42.04	26.10	41.50	26.13	45.72	13.06	40.11	14.84	40.14	24.38	20.28	13.29	21.59	18.32	19.51
Qwen3-VL-30B	37.72	49.40	48.25	50.98	44.00	48.21	27.72	49.67	34.43	46.72	27.70	40.83	33.54	41.32	36.64	41.13
DS-Distill-Qwen-14B	31.98	45.48	33.09	47.10	21.60	43.98	20.03	46.06	24.73	42.80	18.28	46.28	16.46	39.89	30.53	38.67
Ours	**60.93**	**56.06**	**65.02**	**55.52**	**64.18**	50.31	**52.83**	**56.77**	**51.74**	**53.37**	**59.56**	**66.90**	60.13	**63.37**	60.31	**62.14**
Single-Choice																
Base Model	36.69	44.69	55.57	42.87	41.34	38.53	52.15	47.10	49.60	48.86	33.41	19.77	47.49	27.08	43.04	33.20
- w/ NaiveRAG	44.30	44.37	58.79	42.22	44.93	39.09	53.17	48.36	51.10	51.24	33.54	18.65	49.81	29.97	45.18	35.19
MemGPT	35.21	43.72	47.79	40.69	39.87	36.06	48.26	46.55	40.18	45.30	48.59	23.76	61.00	36.31	60.17	41.96
Mem0	43.80	38.83	55.70	36.22	48.20	33.80	59.51	39.68	63.63	40.60	34.88	–	32.43	–	49.68	–
MemoryOS	39.67	44.83	58.39	41.40	48.37	37.34	57.67	47.17	61.22	48.95	53.73	45.53	59.07	36.71	46.25	25.30
LangMem	50.58	44.66	63.09	39.69	45.92	37.87	60.94	47.99	63.23	44.52	53.61	–	63.32	–	61.46	–
A-mem	49.75	46.89	60.40	43.98	43.46	39.93	61.96	50.10	65.43	50.29	50.92	44.03	64.86	40.49	62.74	38.12
Light-mem	49.42	46.90	61.88	44.72	51.14	40.57	59.30	50.87	56.31	51.26	52.02	39.56	**64.86**	40.42	60.60	38.52
GAM	50.41	46.21	64.97	44.56	**59.48**	**46.55**	49.69	47.80	65.33	52.54	49.45	28.02	55.21	32.05	56.96	33.67
MemAgent-7B	48.60	43.84	53.96	42.07	45.92	40.37	52.56	42.16	57.82	43.88	51.90	50.20	56.37	45.38	60.60	42.98
MEM1-7B	48.60	43.97	55.03	42.44	43.30	38.93	45.40	40.30	55.91	43.00	45.17	23.53	50.58	28.12	49.46	35.10
Qwen3-VL-30B	51.57	43.25	63.49	42.08	50.49	38.78	59.92	45.95	62.63	47.50	55.32	30.37	62.16	36.69	60.39	39.29
DS-Distill-Qwen-14B	47.93	44.80	57.05	45.00	49.18	43.30	63.39	41.92	62.42	45.78	54.83	56.65	52.12	39.00	52.46	48.78
Ours	**60.99**	**50.68**	**68.32**	**46.55**	58.82	42.48	**69.33**	**53.17**	**68.64**	**55.02**	**65.36**	**62.70**	64.48	**52.68**	**73.66**	**54.87**
Direct-Answer																
Base Model	21.43	47.24	20.80	47.59	22.48	49.46	19.82	49.13	19.80	43.06	37.39	25.32	21.97	21.95	3.88	19.77
- w/ NaiveRAG	21.30	46.46	22.78	49.29	15.96	45.87	18.92	48.87	14.67	43.10	36.89	28.69	24.22	26.50	3.88	20.59
MemGPT	15.44	45.19	14.26	45.66	18.73	48.27	15.17	48.15	17.45	40.06	38.07	17.22	15.70	12.57	3.45	17.06
Mem0	23.65	48.42	21.76	48.63	24.61	51.20	20.66	50.00	20.61	41.05	38.41	–	19.28	–	5.60	–
MemoryOS	17.72	46.53	16.98	46.67	21.51	49.23	16.72	49.18	18.50	41.05	36.72	19.96	17.94	14.06	5.17	18.16
LangMem	22.08	48.42	16.71	48.65	23.51	50.93	18.21	51.51	18.25	42.05	38.07	–	26.46	–	5.17	–
A-mem	23.78	50.80	21.15	50.82	26.03	53.86	20.01	49.54	22.83	43.89	39.42	25.17	23.77	22.98	6.47	19.89
Light-mem	25.67	50.24	21.15	49.47	25.58	53.90	21.24	49.47	21.35	44.14	39.09	26.83	24.22	21.70	7.33	23.22
GAM	21.17	45.30	22.44	47.54	24.03	49.70	21.82	47.15	21.41	41.72	43.99	27.95	23.77	17.33	6.90	17.63
MemAgent-7B	10.16	31.46	11.12	41.94	13.05	33.91	11.36	41.23	15.22	25.25	23.52	23.18	11.21	23.66	4.74	17.20
MEM1-7B	11.21	22.47	15.35	44.50	10.08	21.18	10.59	41.49	13.80	17.89	24.03	20.88	12.56	12.99	3.45	15.44
Qwen3-VL-30B	21.30	46.92	23.87	49.69	26.42	51.35	21.17	50.64	21.47	45.19	44.67	29.85	25.56	28.69	3.88	21.01
DS-Distill-Qwen-14B	17.00	47.40	29.06	49.20	13.82	50.45	18.08	43.45	13.43	42.33	34.01	29.10	24.22	33.50	3.45	17.15
Ours	**27.23**	**53.73**	**31.51**	**52.84**	**35.66**	**57.24**	**26.79**	**53.91**	**28.59**	**48.96**	**59.05**	**54.15**	**43.50**	**63.28**	**10.78**	**58.68**

Table 5: Detailed results on RefMem-Bench are reported using **open-source** implementations of training-free memory methods. Qwen3-VL-8B is used as the LLM backbone. We separately report answer accuracy (Acc) and memory recall (MemR) for Multi-Choice, Single-Choice, and Direct-Answer settings. "–" indicates that the original paper setting does not support the corresponding evaluation protocol.

Models	TD		PI		BM		EGP		IBM		CMLI		PVB		TSVP	
	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR	Acc	MemR
Multi-Choice																
Base Model	36.88	48.78	47.44	50.86	42.31	47.43	24.38	48.62	25.55	46.54	24.38	39.56	27.85	39.24	36.64	38.66
- *w/ NaiveRAG*	37.47	50.45	47.71	51.32	40.44	49.93	23.80	51.89	27.20	47.54	25.21	37.84	28.48	38.34	38.17	39.70
MemGPT	38.82	44.92	29.33	45.01	26.40	43.55	20.32	45.56	25.55	42.59	31.02	30.61	31.65	34.66	37.40	31.27
Mem0	12.57	37.43	10.76	39.20	11.82	39.27	12.63	39.76	7.88	34.77	33.24	–	36.71	–	33.59	–
MemoryOS	41.94	46.44	33.18	48.33	32.62	46.01	27.14	50.58	27.29	46.75	29.09	42.52	36.08	39.43	53.44	44.22
LangMem	36.20	50.53	34.53	50.69	35.56	**50.53**	26.12	50.52	22.80	46.63	26.59	–	31.65	–	41.98	–
A-mem	48.61	49.88	42.24	51.90	39.82	48.90	33.67	52.16	35.99	47.89	39.34	27.89	48.73	29.60	46.56	26.75
Light-mem	45.65	50.44	42.33	51.17	39.56	49.10	23.22	52.92	26.01	46.67	38.23	30.25	49.37	30.80	51.91	35.56
GAM	45.65	49.52	42.96	50.35	36.53	48.52	33.53	52.32	32.14	46.44	41.55	35.68	38.61	28.07	41.22	39.43
MemAgent-7B	30.38	45.30	24.13	45.86	25.87	45.07	23.80	45.86	28.02	43.20	30.19	39.67	13.92	29.00	23.66	30.88
MEM1-7B	22.28	42.04	26.10	41.50	26.13	45.72	13.06	40.11	14.84	40.14	24.38	20.28	13.29	21.59	18.32	19.51
Qwen3-VL-30B	37.72	49.40	48.25	50.98	44.00	48.21	27.72	49.67	34.43	46.72	27.70	40.83	33.54	41.32	36.64	41.13
DS-Distill-Qwen-14B	31.98	45.48	33.09	47.10	21.60	43.98	20.03	46.06	24.73	42.80	18.28	46.28	16.46	39.89	30.53	38.67
Ours	**54.94**	**56.06**	**64.04**	**55.52**	**63.20**	50.31	**51.81**	**56.77**	**49.73**	**53.37**	**59.56**	**66.90**	**53.80**	**63.37**	**60.31**	**62.14**
Single-Choice																
Base Model	36.69	44.69	55.57	42.87	41.34	38.53	52.15	47.10	49.60	48.86	33.41	19.77	47.49	27.08	43.04	33.20
- *w/ NaiveRAG*	44.30	44.37	58.79	42.22	44.93	39.09	53.17	48.36	51.10	51.24	33.54	18.65	49.81	29.97	45.18	35.19
MemGPT	38.35	45.94	53.02	41.58	37.58	40.94	48.47	48.76	46.59	50.85	47.74	20.73	52.51	26.07	51.82	34.70
Mem0	44.13	37.10	56.91	33.60	38.73	31.60	55.01	48.55	59.72	48.28	49.82	–	50.97	–	52.68	–
MemoryOS	37.02	43.62	52.35	40.84	38.40	39.59	52.56	47.60	46.79	44.76	49.33	30.27	55.98	38.54	55.67	43.60
LangMem	42.64	44.37	61.34	41.10	47.06	36.91	62.78	47.69	46.99	45.08	48.23	–	51.74	–	53.53	–
A-mem	46.12	45.92	58.66	43.11	38.07	40.67	54.81	48.88	52.20	47.80	52.51	36.71	54.83	34.95	59.31	37.11
Light-mem	39.34	43.75	61.21	42.53	44.12	38.83	53.58	48.31	51.60	47.91	52.88	39.91	56.37	38.75	56.96	36.21
GAM	43.14	44.98	59.06	44.29	41.01	39.78	53.58	48.46	52.40	46.96	45.53	36.45	50.19	39.50	55.46	35.43
MemAgent-7B	48.60	43.84	53.96	42.07	45.92	40.37	52.56	42.16	57.82	43.88	51.90	50.20	56.37	45.38	60.60	42.98
MEM1-7B	48.60	43.97	55.03	42.44	43.30	38.93	45.40	40.30	55.91	43.00	45.17	23.53	50.58	28.12	49.46	35.10
Qwen3-VL-30B	51.57	43.25	63.49	42.08	50.49	38.78	59.92	45.95	62.63	47.50	55.32	30.37	62.16	36.69	60.39	39.29
DS-Distill-Qwen-14B	47.93	44.80	57.05	45.00	49.18	**43.30**	63.39	41.92	62.42	45.78	54.83	56.65	52.12	39.00	52.46	48.78
Ours	**60.99**	**50.68**	**66.31**	**46.55**	**58.82**	42.48	**69.33**	**53.17**	**68.64**	**55.02**	**65.36**	**62.70**	**64.48**	**52.68**	**73.66**	**54.87**
Direct-Answer																
Base Model	21.43	47.24	20.80	47.59	22.48	49.46	19.82	49.13	19.80	43.06	37.39	25.32	21.97	21.95	3.88	19.77
- *w/ NaiveRAG*	21.30	46.46	22.78	49.29	15.96	45.87	18.92	48.87	14.67	43.10	36.89	28.69	24.22	26.50	3.88	20.59
MemGPT	12.96	35.04	10.57	35.81	13.82	41.29	10.65	40.46	10.27	33.06	36.38	41.45	17.94	30.83	4.31	20.42
Mem0	7.75	35.42	11.05	35.62	13.05	37.26	12.72	41.11	11.82	40.05	38.41	–	26.01	–	4.74	–
MemoryOS	15.83	38.20	15.48	38.66	19.25	45.81	13.36	41.62	15.84	40.22	32.15	35.69	18.83	39.38	4.74	24.75
LangMem	20.65	48.42	19.30	48.65	19.96	50.31	19.95	48.34	18.19	43.46	38.58	–	22.42	–	3.45	–
A-mem	23.13	48.98	19.44	50.67	23.58	53.35	17.95	48.57	21.23	44.61	37.06	25.09	22.87	22.76	5.60	21.69
Light-mem	21.89	47.99	17.12	48.39	19.64	50.16	17.37	46.74	17.08	41.12	36.89	22.09	18.39	20.74	4.31	18.71
GAM	19.61	48.03	19.71	46.53	24.87	50.91	18.33	47.58	19.06	41.02	40.44	26.75	25.11	25.56	6.03	21.24
MemAgent-7B	10.16	31.46	11.12	41.94	13.05	33.91	11.36	41.23	15.22	25.25	23.52	23.18	11.21	23.66	4.74	17.20
MEM1-7B	11.21	22.47	15.35	44.50	10.08	21.18	10.59	41.49	13.80	17.89	24.03	20.88	12.56	12.99	3.45	15.44
Qwen3-VL-30B	21.30	46.92	23.87	49.69	26.42	51.35	21.17	50.64	21.47	45.19	44.67	29.85	25.56	28.69	3.88	21.01
DS-Distill-Qwen-14B	17.00	47.40	29.06	49.20	13.82	50.45	18.08	43.45	13.43	42.33	34.01	29.10	24.22	33.50	3.45	17.15
Ours	**27.23**	**53.73**	**31.51**	**52.84**	**35.66**	**57.24**	**25.82**	**53.91**	**28.59**	**48.96**	**59.05**	**54.15**	**43.50**	**63.28**	**10.78**	**58.68**

Table 6: Detailed Direct-Answer results for the main experiments on RefMem-Bench. We report BLEU-1 (B1) and token-level F1 (F1) for each evaluation dimension.

Models	TD		PI		BM		EGP		IBM		CMLI		PVB		TSVP	
	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1
Direct-Answer																
Base Model	17.24	20.66	16.19	19.13	15.00	18.12	22.12	24.04	17.58	21.03	37.38	41.92	9.74	18.02	3.92	4.23
- *w/ NaiveRAG*	19.25	16.28	18.06	19.45	15.31	17.71	18.71	20.20	15.81	19.60	35.67	40.40	10.09	18.89	3.03	5.63
MemGPT	14.80	16.85	12.64	14.17	14.39	16.42	18.60	20.54	17.70	20.61	39.89	42.77	9.27	16.85	4.77	5.94
Mem0	13.74	15.47	15.26	16.96	14.38	17.06	21.33	23.53	19.06	21.82	40.39	43.11	10.21	15.21	5.76	6.00
MemoryOS	17.47	19.68	17.25	19.42	20.40	22.27	20.41	22.97	18.36	21.78	39.66	42.30	10.02	14.63	5.06	6.18
LangMem	20.72	22.15	19.72	21.25	20.98	22.96	23.16	25.42	18.54	20.94	41.48	44.28	9.12	14.99	5.12	6.35
A-mem	16.95	19.14	14.55	16.60	18.84	21.62	18.91	21.20	16.40	19.58	30.74	33.18	9.77	10.74	5.22	6.22
Light-mem	21.93	25.68	16.56	18.78	19.81	22.76	22.61	25.03	19.28	22.79	42.67	46.53	11.37	16.95	6.29	7.34
GAM	17.70	20.36	18.00	21.10	20.21	24.16	24.39	27.09	17.92	21.76	43.33	45.29	11.01	16.54	6.95	7.54
MemAgent-7B	9.71	11.86	9.34	10.56	11.02	14.62	14.25	16.53	14.84	16.66	25.93	29.75	8.35	10.39	3.16	3.28
MEM1-7B	7.75	9.12	13.52	15.85	7.46	9.89	9.73	11.58	10.03	13.04	24.22	27.74	7.93	11.00	4.76	5.21
Qwen3-VL-30B	19.06	21.68	20.01	22.16	17.69	19.82	23.99	26.53	18.76	21.31	41.83	44.11	10.11	18.31	4.29	5.21
DS-Distill-Qwen-14B	15.65	18.32	20.83	25.03	13.92	16.51	19.49	21.70	15.39	17.50	25.12	28.60	8.14	17.06	4.52	5.60
Ours	**25.50**	**27.25**	**27.57**	**29.87**	**29.35**	**30.28**	**32.60**	**35.68**	**27.25**	**29.98**	**49.48**	**53.07**	**17.69**	**24.40**	**11.34**	**12.77**

Table 7: Additional Direct-Answer results on RefMem-Bench are reported using **open-source** implementations of training-free memory methods. Qwen3-VL-8B is used as the LLM backbone. We report BLEU-1 (B1) and token-level F1 (F1) for each evaluation dimension.

Models	TD		PI		BM		EGP		IBM		CMLI		PVB		TSVP	
	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1	B1	F1
Direct-Answer																
Base Model	17.24	20.66	16.19	19.13	15.00	18.12	22.12	24.04	17.58	21.03	37.38	41.92	9.74	18.02	3.92	4.23
- *w/ NaiveRAG*	19.25	16.28	18.06	19.45	15.31	17.71	18.71	20.20	15.81	19.60	35.67	40.40	10.09	18.89	3.03	5.63
MemGPT	13.03	14.49	12.07	12.90	12.00	12.98	14.06	15.24	13.79	15.42	36.87	39.77	9.07	13.82	3.50	5.17
Mem0	12.81	13.70	10.29	11.20	13.22	14.22	15.09	16.21	13.09	14.47	38.21	41.19	11.30	13.86	2.39	3.46
MemoryOS	15.02	16.54	14.55	15.76	14.32	15.64	15.51	16.68	15.84	17.60	37.28	38.78	10.49	12.37	2.75	3.96
LangMem	19.05	21.07	16.69	19.24	19.23	20.94	19.26	21.00	17.91	17.03	34.25	36.87	9.13	12.51	3.13	3.92
A-mem	15.56	17.34	11.84	12.83	12.21	13.51	14.29	15.51	14.01	15.75	24.44	26.44	10.73	12.65	2.77	3.44
Light-mem	18.06	20.28	12.22	13.61	10.41	11.78	15.71	17.21	14.42	16.56	36.25	39.36	10.61	12.18	2.51	3.03
GAM	17.97	20.14	16.28	18.14	18.38	20.50	20.26	22.08	18.06	20.60	38.45	42.23	11.90	15.86	2.03	3.93
MemAgent-7B	9.71	11.86	9.34	10.56	11.02	14.62	14.25	16.53	14.84	16.66	25.93	29.75	8.35	10.39	3.16	3.28
MEM1-7B	7.75	9.12	13.52	15.85	7.46	9.89	11.73	13.58	10.03	13.04	24.22	27.74	7.93	11.00	4.76	5.21
Qwen3-VL-30B	19.06	21.68	20.01	22.16	17.69	19.82	23.99	26.53	18.76	21.31	41.83	44.11	10.11	18.31	4.29	5.21
DS-Distill-Qwen-14B	15.65	18.32	20.83	25.03	13.92	16.51	19.49	21.70	15.39	17.50	25.12	28.60	8.14	17.06	4.52	5.60
Ours	**25.50**	**27.25**	**27.57**	**29.87**	**29.35**	**30.28**	**32.60**	**35.68**	**27.25**	**29.98**	**49.48**	**53.07**	**17.69**	**24.40**	**11.34**	**12.77**

E.4 COMPUTATION COMPLEXITY

Table 8: Complexity comparison between REMIND and other memory baselines.

System	(I) Segment & Retrieval	(II) Construction	(III) Inference
Baselines	$O(N)\cdot\mathcal{C}_{\text{seg}}$ $+O(N)\cdot\mathcal{C}_{\text{LLM}}$ $+O(NM_1)\cdot\mathcal{C}_{\text{emb}}$	$O(NM_1)\cdot\mathcal{C}_{\text{sim}}$ $+O(NM_1R_1)\cdot\mathcal{C}_{\text{LLM}}$	$O(N)\cdot\mathcal{C}_{\text{LLM}}$ $+O(NM_1R_1)\cdot\mathcal{C}_{\text{LLM}}$ $+O(1)\cdot\mathcal{C}_{\text{LLM}}$
REMIND	$O(1)\cdot\mathcal{C}_{\text{emb}}$ $+O(N)\cdot\mathcal{C}_{\text{sim}}$ $+O(N\log k_1)$	$O(k_1T\cdot k)\cdot\mathcal{C}_{\text{sae}}$ $+O(k_1T)\cdot\mathcal{C}_{\text{LLM}}$	$O(N)\cdot\mathcal{C}_{\text{sim}}$ $+O(1)\cdot\mathcal{C}_{\text{LLM}}$

As shown in Table 8, we analyze a long-horizon dialogue consisting of N turns, each containing an average of T tokens, and evaluate the cost of answering a reflective memory query q. In other memory systems (OMS), the dialogue is first segmented into N turns, and each turn triggers a summarization and indexing operation. Each summarization requires an LLM forward pass, and each segment is further encoded into M_1 memory entries for indexing:

$$\mathcal{C}^{\text{I}}_{\text{OMS}} = O(N)\cdot\mathcal{C}_{\text{seg}} \;+\; O(N)\cdot\mathcal{C}_{\text{LLM}} \;+\; O(NM_1)\cdot\mathcal{C}_{\text{emb}}. \tag{15}$$

Each summarization then produces M_1 memory entries. A fraction R_1 of these entries trigger an action (e.g., an add, update, or delete operation), each of which requires an additional LLM call:

$$\mathcal{C}^{\text{II}}_{\text{OMS}} = O(NM_1)\cdot\mathcal{C}_{\text{sim}} \;+\; O(NM_1R_1)\cdot\mathcal{C}_{\text{LLM}}. \tag{16}$$

This results in a total inference cost of:

$$\mathcal{C}^{\text{infer}}_{\text{OMS}} = O(N)\cdot\mathcal{C}_{\text{LLM}} \;+\; O(NM_1R_1)\cdot\mathcal{C}_{\text{LLM}} \;+\; O(1)\cdot\mathcal{C}_{\text{LLM}}. \tag{17}$$

In **REMIND**, the segment-level cues $\{\kappa_n\}_{n=1}^{N}$ and embeddings $\{\phi(\kappa_n)\}_{n=1}^{N}$ are precomputed offline and reused across queries. Each query requires one embedding computation followed by matching against all segments:

$$\mathcal{C}^{\text{I}}_{\text{REMIND}} = O(1)\cdot\mathcal{C}_{\text{emb}} \;+\; O(N)\cdot\mathcal{C}_{\text{sim}} \;+\; O(N\log k_1). \tag{18}$$

On top of this, the reflective layer performs an LLM call to generate a reflective summaries. Since the input consists of k_1 selected segments of average length T, this call incurs a cost of $O(k_1T)\cdot\mathcal{C}_{\text{LLM}}$, independent of N. Therefore, the total cost of Stage II in REMIND is:

$$\mathcal{C}^{\text{II}}_{\text{REMIND}} = O(k_1T\cdot k)\cdot\mathcal{C}_{\text{sae}} \;+\; O(k_1T)\cdot\mathcal{C}_{\text{LLM}}. \tag{19}$$

The LLM cost is thus decoupled from the dialogue length N. Combining retrieval, attention, reflection, and final response generation, REMIND requires:

$$\mathcal{C}^{\text{infer}}_{\text{REMIND}} = O(N)\cdot\mathcal{C}_{\text{sim}} \;+\; O(1)\cdot\mathcal{C}_{\text{LLM}}. \tag{20}$$

Following prior efficiency-oriented works, we report FLOPs measured by calflops (Ye, 2023). As shown in Figure 12, REMIND shifts the dominant online cost away from the expensive LLM forward pass and toward much cheaper sparse SAE computation. At the same time, through endogenous activation alignment within the cognitive pyramid, reflective reasoning is progressively absorbed into the model's internal pathway, allowing inference to rely on retrieval and the model's native capabilities rather than repeated LLM-mediated memory updates.

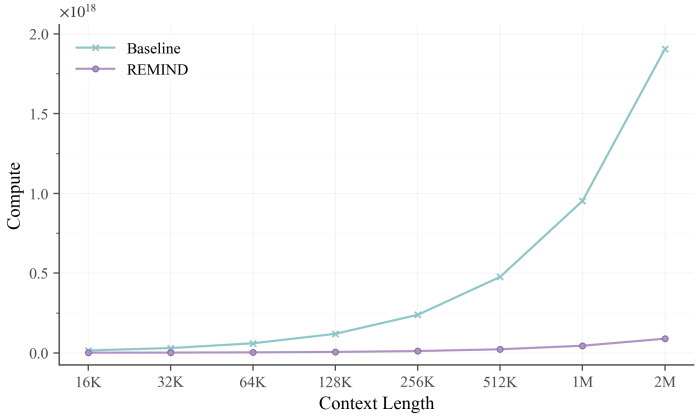


Figure 12: Floating point operations across context lengths from 16K to 2M.

Table 9: Notation used in the complexity analysis.

Symbol	Definition
N	Total number of dialogue turns (also the number of segments in the fact layer).
T	Average number of tokens per turn.
k_1	Number of segments retained after top-k_1 relevance selection in the fact layer.
k	Number of active dimensions retained in Top-K SAE (sparsity parameter).
M_1	Number of memory entries produced by each summarization in OMS.
R_1	Fraction of memory entries that successfully retrieve at least one relevant neighbor and trigger an update in OMS.
$\mathcal{C}_{\text{LLM}}$	Cost of a single-token forward pass of the backbone LLM (*dominant term*); can represent runtime, FLOPs, or other computational measures. An LLM call with input length L incurs cost $L \cdot \mathcal{C}_{\text{LLM}}$.
$\mathcal{C}_{\text{emb}}$	Cost of a single embedding forward pass.
$\mathcal{C}_{\text{sim}}$	Cost of a single vector similarity computation.
$\mathcal{C}_{\text{sae}}$	Cost of a single Top-K SAE forward pass.

E.5 Training Trajectory Analysis of REMIND

We present the training trajectory of REMIND in Figure 13. The figure combines two complementary views of training dynamics. The stacked bars show the normalized loss distribution over equal training intervals. Each color denotes the normalized contribution of one objective term to the total loss, including the hard supervision loss $\mathcal{L}_{\text{hard}}$, the lower-level distillation loss $\mathcal{L}_{\text{dist}}^{(0\to1)}$, and the higher-level distillation loss $\mathcal{L}_{\text{dist}}^{(1\to2)}$. The height of each segment therefore reflects which learning signal dominates a given training stage. The red curve reports the Jensen–Shannon divergence (JSD) between the student distribution and its direct teacher distribution. Following the Shannon-entropy formulation of Jensen–Shannon divergence introduced by Lin (1991), we use its equivalent KL form to measure the discrepancy between the student and teacher output distributions:

$$\text{JSD}(p_s, p_t) = \frac{1}{2} D_{\text{KL}}(p_s \| m) + \frac{1}{2} D_{\text{KL}}(p_t \| m), \qquad m = \frac{1}{2}(p_s + p_t), \tag{21}$$

where p_s and p_t denote the student and teacher output distributions, respectively. A lower JSD indicates stronger distribution-level alignment. In the early phase, covering roughly the first 30% of training, $\mathcal{L}_{\text{hard}}$ contributes the largest share of the objective. The JSD is also high and unstable. This suggests that the model still relies mainly on task-level supervision, while the student distribution has not yet aligned well with the teacher distribution.

After this initial stage, REMIND enters a middle-level alignment phase. From around 30% to 70% of training, $\mathcal{L}_{\text{dist}}^{(0\to1)}$ becomes the dominant signal. During the same period, the JSD decreases steadily from a high value to a near-zero level. This indicates that the attentional state begins to guide the factual pathway effectively. The model no longer only fits the hard answer label. Instead, it starts to absorb the intermediate semantic structure induced by memory grounding.

A second transition appears after around 70% of training. The contribution of $\mathcal{L}_{\text{dist}}^{(1\to2)}$ increases and becomes dominant, while the JSD remains low. This implies that the lower-level alignment has largely stabilized. The model can then allocate more capacity to the higher-level reflective signal. At this stage, REMIND shifts from local answer fitting to reflective abstraction transfer.

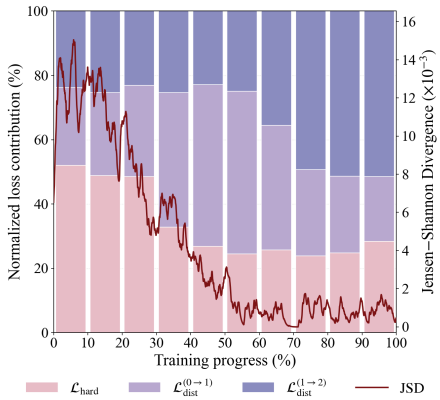


Figure 13: Training trajectory of REMIND across training progress.

Finally, in the last training phase, the loss composition becomes stable and the JSD stays close to zero. This shows that the student and teacher distributions remain well aligned after the dominant distillation signal changes. The trajectory suggests a bottom-up curriculum learned without an explicit schedule: REMIND first anchors task correctness, then absorbs attentional grounding, and finally consolidates reflective abstraction. This staged behavior supports the effectiveness of progressive reflective alignment in transferring multi-level memory reasoning into the factual inference pathway.

E.6 PROMPTS

Table 10: Prompt templates for text-based memory recall.

Task type	Prompt template
Multi-choice	You are an expert at long-context question answering over dialogue transcripts. Given the dialogue context and a multiple-choice question with Four options (A–D), select All correct option letters. (Note: One, two, three, or four options may be correct.) **Rules:** (1) Use Only information supported by the dialogue context. (2) Output the answer inside <answer>[A,D]</answer> – letters must be uppercase A–D, comma-separated, sorted alphabetically. (3) Output All supporting dialogue evidence inside <explanation>...</explanation>, where each line follows the format Speaker: text_content, copied exactly from the dialogue context. Include only relevant lines. (4) No extra text outside the tags. **Dialogue context:** {dialogue_context} **Question:** {question} **Options:** {options[0]} {options[1]} {options[2]} {options[3]}
Single-choice	You are an expert at long-context question answering over dialogue transcripts. Given the dialogue context and a multiple-choice question with Four options (A–D), select exactly One correct option letter. (Note: Only one option is correct.) **Rules:** (1) Use Only information supported by the dialogue context. (2) Output the answer inside <answer>A</answer> – the letter must be uppercase A–D, exactly one letter, no extra text outside the tags. (3) Output All supporting dialogue evidence inside <explanation>...</explanation>, where each line follows the format Speaker: text_content, copied exactly from the dialogue context. Include only relevant lines. (4) No extra text outside the tags. **Dialogue context:** {dialogue_context} **Question:** {question} **Options:** {options[0]} {options[1]} {options[2]} {options[3]}
Direct answer	You are an expert at long-context question answering over dialogue transcripts. Given the dialogue context and a direct-answer question, provide a concise answer as a short noun phrase or verb phrase of 1–4 words. **Rules:** (1) Use Only information supported by the dialogue context. (2) Output your answer inside <answer>your short phrase here</answer> – the answer Must be a noun or verb phrase of 1–4 words, no full sentences, no "Because ..." explanations, use vocabulary from the dialogue when possible. (3) Output All supporting dialogue evidence inside <explanation>...</explanation>, where each line follows the format Speaker: text_content, copied exactly from the dialogue context. Include only relevant lines. (4) No extra text outside the tags. **Dialogue context:** {dialogue_context} **Question:** {question}

Table 11: Prompt templates for visual memory recall.

Task type	Prompt template
Multi-choice	You are an expert at long-context question answering over dialogue transcripts. The dialogue context contains text turns interleaved with image captions; each caption is labeled with its 0-based global index in the format [Image #N Caption]. Given the dialogue context and a multiple-choice question with Four options (A–D), select All correct option letters. (Note: One, two, three, or four options may be correct.) **Rules:** (1) Use Only information supported by the dialogue context. (2) Output the answer inside <answer>[A,D]</answer> – letters must be uppercase A–D, comma-separated, sorted alphabetically. (3) Output the indices of the images you used inside <explanation>[0, 1]</explanation> as a JSON list of integers, each corresponding to the N in [Image #N Caption]; include only relevant indices, or [] if none. (4) No extra text outside the tags. **Dialogue context:** {dialogue_context} **Question:** {question} **Options:** {options[0]} {options[1]} {options[2]} {options[3]}
Single-choice	You are an expert at long-context question answering over dialogue transcripts. The dialogue context contains text turns interleaved with image captions; each caption is labeled with its 0-based global index in the format [Image #N Caption]. Given the dialogue context and a multiple-choice question with Four options (A–D), select exactly One correct option letter. (Note: Only one option is correct.) **Rules:** (1) Use Only information supported by the dialogue context. (2) Output the answer inside <answer>A</answer> – the letter must be uppercase A–D, exactly one letter, no extra text outside the tags. (3) Output the indices of the images you used inside <explanation>[0, 1]</explanation> as a JSON list of integers, each corresponding to the N in [Image #N Caption]; include only relevant indices, or [] if none. (4) No extra text outside the tags. **Dialogue context:** {dialogue_context} **Question:** {question} **Options:** {options[0]} {options[1]} {options[2]} {options[3]}
Direct answer	You are an expert at long-context question answering over dialogue transcripts. The dialogue context contains text turns interleaved with image captions; each caption is labeled with its 0-based global index in the format [Image #N Caption]. Given the dialogue context and a direct-answer question, provide a concise answer as a short noun phrase or verb phrase of 1–4 words. **Rules:** (1) Use Only information supported by the dialogue context. (2) Output your answer inside <answer>your short phrase here</answer> – the answer Must be a noun or verb phrase of 1–4 words, no full sentences, no "Because ..." explanations, use vocabulary from the dialogue when possible. (3) Output the indices of the images you used inside <explanation>[0, 1]</explanation> as a JSON list of integers, each corresponding to the N in [Image #N Caption]; include only relevant indices, or [] if none. (4) No extra text outside the tags. **Dialogue context:** {dialogue_context} **Question:** {question}

Table 12: Level-specific prompt components across the three cognitive pyramid levels. The factual level provides the plain reconstructed dialogue with a Reflective-Memory chain-of-thought guide, the attentional level wraps the dialogue with context boundaries and soft-attention markers, and the reflective level augments the dialogue with a multi-faceted guided-reasoning analysis that must be verified against the dialogue.

Pyramid level	**Level-specific prompt component in REMIND**
Factual level	Provides the plain reconstructed dialogue together with a Reflective-Memory chain-of-thought block that guides where to attend, but never overrides the dialogue itself. **Component:** The following is a structured chain-of-thought analysis produced by a prior reflective memory stage. It contains anchor information (target speaker, related entities, behavioral focus, temporal scope) and step-by-step reasoning cues that help you locate and interpret the most relevant evidence in the dialogue. Use this as an analytical Guide to direct your attention when reading the dialogue context. However, you must Always verify its claims against the actual dialogue — if the CoT contradicts what you observe in the dialogue, trust the dialogue. {reflective_memory_cot}
Attentional level	Wraps the dialogue with explicit context boundaries and injects soft-attention markers {attention_token} around turns identified as semantically critical, then reinforces a marker-priority instruction. **Component:** {marked_dialogue} **Critical reminder:** Dialogue turns enclosed by {attention_token} markers contain the Most relevant evidence for answering the question below. These marked sections were identified through semantic analysis as being directly pertinent to the question. You Must prioritize information within the {attention_token} markers when reasoning about the correct answer. Unmarked turns provide background context but are less likely to contain the key evidence needed for the answer.
Reflective level	Augments the dialogue with a multi-faceted guided-reasoning analysis (dimension understanding, option-grounded evidence, guiding conclusion), and explicitly demands cross-verification against the dialogue so the analysis acts as a reference rather than a verdict. **Component:** The following is a structured reasoning guide produced by a prior analytical stage. It contains dimension understanding, option-grounded evidence analysis, and a guiding conclusion. Use this as an analytical Reference to inform your reasoning process — it highlights which behavioral patterns to look for and how to interpret the evidence. However, you must Always verify its claims against the actual dialogue context above. If the guided reasoning contradicts what you observe in the dialogue, trust the dialogue. {guided_reasoning} **Note:** The guided reasoning above is a reference, not a definitive answer. Combine its analytical framework with the concrete evidence from the dialogue context to arrive at your final answer. Focus on verifiable behavioral patterns (repeated actions, speech patterns, avoidance moves) that you can confirm in the dialogue text.