MEMLENS: BENCHMARKING MULTIMODAL LONG-TERM MEMORY IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory is essential for large vision-language models (LVLMs) to handle long, multimodal interactions, with two method directions providing this capability: long-context LVLMs and memory-augmented agents. However, no existing benchmark conducts a systematic comparison of the two on questions that genuinely require multimodal evidence. To close this gap, we introduce MEMLENS, a comprehensive benchmark for memory in multimodal multi-session conversations, comprising 789 questions across five memory abilities (information extraction, multi-session reasoning, temporal reasoning, knowledge update, and answer refusal) at four standard context lengths (32K–256K tokens) under a cross-modal token-counting scheme. On 634 image-essential and image-supportive questions, controlled image ablation drops two frontier LVLMs below 2% accuracy. Across 27 LVLMs and 7 memory-augmented agents, the best 32K accuracies are 58.68% for LVLMs and 33.46% for memory agents. LVLMs degrade as conversations grow, whereas memory agents are comparatively length-stable but lose visual fidelity under storage-time compression. Multi-session reasoning caps most systems below 30%, and neither approach alone solves the task. Abstention results suggest memory-oriented post-training risks weakening correct refusal. These results motivate hybrid architectures that combine long-context attention with structured multimodal retrieval.

1 INTRODUCTION

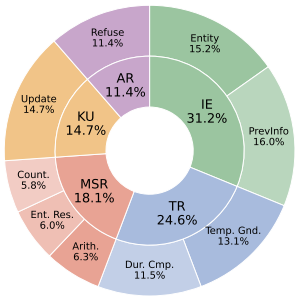
Memory is essential for enabling large vision-language models (LVLMs) (Seed; OpenAI, 2025) to maintain consistency across extended multimodal interactions and to continuously incorporate new information over time (OpenAI, 2023a; Anthropic, 2024; Team et al., 2024). When LVLMs are deployed as agents, the inputs they handle accumulate incrementally rather than arriving all at once. As interaction histories expand, the agent must recall past content and reason over the growing context to remain consistent with prior turns. At the same time, it must integrate new facts as they arrive and revise outdated information to keep its knowledge current. Two directions have emerged to provide this capability, namely long-context LVLMs and memory-augmented agents. Long-context LVLMs enlarge the native context window so that the complete dialogue history, including interleaved images, can be processed directly by the model (Team et al., 2024; Seed, 2026; Anthropic, 2025). Memory-augmented agents, building on retrieval-augmented generation, instead compress, index, and selectively retrieve past content from an external store (Packer et al., 2024; Yadav et al., 2025; Jin et al., 2024).

Despite progress along both directions, no current benchmark conducts a length-controlled comparison of long-context LVLMs and memory-augmented agents on questions that genuinely require visual evidence. Long-context multimodal benchmarks measure context-length scaling on long documents and retrieval tasks (Wang et al., 2025d; 2024b; Ma et al., 2024), but cover mainly LVLMs and do not place them alongside memory-augmented agents for direct comparison. Text-only conversational memory benchmarks such as LongMemEval (Wu et al., 2025) and MemoryAgentBench (Hu et al., 2026) overlook the visual modality entirely, treating memory as a single-modality problem. Multimodal conversational benchmarks such as LoCoMo (Maharana et al., 2024) and Mem-Gallery (Bei et al., 2026) retain both visual and text modalities, yet most of their questions admit a text-only shortcut, rendering the visual modality redundant. As Table 1 summarizes, no existing benchmark

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.

Table 1: **Left:** Comparison of MEMLENS with existing long-context and conversational memory benchmarks. *Type* — Dial (Dialogue) or Doc (Document); *Max L* — maximum context length reported ("—" if not applicable); *L Control* — standardized token length control; *Comp. Eval.* — paper benchmarks both LVLMs and memory-augmented agents on the same data; *Memory Tasks* — dot pattern showing coverage of the five evaluation tasks (● = supported, ○ = not supported) in fixed order: **IE** (Information Extraction), **MSR** (Multi-Session Reasoning), **TR** (Temporal Reasoning), **KU** (Knowledge Update), **AR** (Answer Refusal). **Right:** Distribution of the 789 evaluation questions in MEMLENS across the five major task types (inner ring) and nine reporting subtypes (outer ring).

Benchmark	Type	Max L	Multi-modal	Multi-Sess.	L Control	Comp. Eval.	Memory Tasks
MemoryBank	Dial	—	✗	✓	✗	✗	●○○○○
LoCoMo	Dial	10K	✓	✓	✗	✗	●●●○○
MM-NIAH	Doc	128K	✓	✗	✓	✗	●○○○○
MMLongBench-Doc	Doc	128K	✓	✗	✗	✗	●○○○○
MRAG-Bench	Doc	—	✓	✗	✗	✗	●○○○○
LongMemEval	Dial	1.5M	✗	✓	✓	✗	●●●●●
MMLongBench	Doc	128K	✓	✗	✓	✗	●○○○○
Multimodal NIAH	Doc	128K	✓	✗	✓	✗	●○○○○
MemAgentBench	Dial	115K	✗	✓	✗	✓	●○○○○
MEMLENS (Ours)	**Dial**	**256K**	✓	✓	✓	✓	●●●●●



requires visual evidence to answer its questions while supporting a length-controlled comparison of long-context LVLMs and memory-augmented agents.

To bridge the gap, we introduce MEMLENS, a comprehensive benchmark for assessing memory in multimodal multi-session conversations. MEMLENS consists of 789 questions covering five core memory abilities: information extraction, multi-session reasoning, temporal reasoning, knowledge update, and answer refusal. We design questions that demand cross-modal reasoning over both visual and textual evidence, requiring the retrieval of multimodal information hidden within one or more conversations between a user and an assistant. To verify that solving MEMLENS requires visual evidence, we conduct an image-ablation study: when these images are removed, the accuracy of two frontier LVLMs collapses below 2% (§3.4). Following the needle-in-a-haystack construction paradigm (Kamradt, 2023), we design a four-stage pipeline that builds a coherent multi-session chat history for each question, distributing the evidence across one or more user-assistant sessions alongside topically related distractor turns. Because distractor turns can be added independently of the evidence, the chat length is freely extensible. We release four standard context lengths (32K, 64K, 128K, and 256K tokens) under a cross-modal token-counting scheme (Wang et al., 2025d) that aligns text and vision tokens. MEMLENS thus enables a length-controlled comparison of long-context LVLMs and memory-augmented agents on multimodal questions spanning five comprehensive memory abilities.

Using MEMLENS, we evaluate 27 LVLMs and 7 memory-augmented agents at standardized context lengths up to 256K. Our evaluation yields three key findings. First, the five memory abilities are not interchangeable: strong information extraction does not predict multi-session reasoning, and multi-session reasoning caps most evaluated systems below 30%. Second, the two approaches exhibit complementary failure modes: long-context LVLMs deliver high short-context accuracy through direct visual grounding, but this advantage shrinks as conversations grow; memory agents, in contrast, are length-stable but lose visual fidelity under storage-time compression, and memory-oriented post-training of agent backbones can additionally weaken their abstention behavior. Third, neither approach comes close to solving the task of long-term memory. These results point to a clear next step: architectures that combine long-context attention with structured multimodal retrieval, rather than scaling either component in isolation.

2 RELATED WORK

Memory-Augmented LLM Agents. Recent surveys systematize memory representations and operations across LLM-based agents (Du et al., 2025a). Text-only memory agents span structured-fact stores (Zhong et al., 2023; Wang et al., 2025a), OS-inspired paging (Packer et al., 2024), tree-summarized retrieval (Sarthi et al., 2024), neurobiological graphs (Gutiérrez et al., 2025),

relational embeddings (Yadav et al., 2025), agentic self-organizing notes (Xu et al., 2025), intent-driven memory selection (Du et al., 2025c), RL-selected time-aware memory (Du et al., 2025b), layered memory tiers (Li et al., 2025), and sliding-window RL agents (Yu et al., 2025). Multimodal extensions add ColPali-style document retrieval (Cho et al., 2024; Faysse et al., 2025), multimodal embeddings (Meng et al., 2025), dual-layer semantic memory (Feng et al., 2026), entity-centric audio-visual memory (Long et al., 2025), sparse video memory (Song et al., 2024b), LoRA-tuned session retrieval (Jang et al., 2025), and intent-guided multimodal response retrieval over multi-session conversations (Wang et al., 2025b). Recent work probes the intersection of long-context LLMs and retrieval-augmented generation (Jin et al., 2024; Jiang et al., 2024; Asai et al., 2023), suggesting retrieval and long attention are complementary rather than competing.

Long-Context and Conversational Memory Benchmarks. Most current long-context benchmarks established core protocols for retrieval and length scaling (Bai et al., 2024a;b; Hsieh et al., 2024; Yen et al., 2024; An et al., 2024; Zhang et al., 2024b; Li et al., 2024), but neglect the visual modality. Multimodal extensions, such as MMLongBench (Wang et al., 2025d), scale context length over documents (Wang et al., 2025d; Ma et al., 2024), needle-style retrieval (Wang et al., 2024b; 2025c), long-form video (Zhou et al., 2025; Wu et al., 2024a), super-long documents (Deng et al., 2024; Chia et al., 2024), and multi-image inputs (Wang et al., 2024f; Song et al., 2024a; Wang et al., 2024c; Ye et al., 2024). Yet their inputs are documents or videos rather than multi-session conversations. Memory under conversational interaction therefore remains unexercised. Conversational memory benchmarks restore the multi-session structure but rarely require visual perception: LongMemEval (Wu et al., 2025) (500 questions over five memory abilities, up to 1.5M tokens), PerLTQA (Du et al., 2024), and MemoryAgentBench (Hu et al., 2026) (retrieval, test-time learning, long-range understanding, selective forgetting) discard images entirely, while LoCoMo (Maharana et al., 2024) and Mem-Gallery (Bei et al., 2026) embed images in persona-grounded dialogue but allow most questions to be answered from text alone, so neither stress-tests multimodal memory (Table 1).

3 The MemLens Benchmark

We propose MemLens, a multimodal long-term conversational memory benchmark that comprises 789 questions instantiated at four standardized input lengths (32K/64K/128K/256K tokens). In contrast to prior benchmarks restricted to text-only conversations and questions, MemLens provides multimodal conversation sessions in which text and images are interleaved, together with questions that require cross-modal reasoning over evidence images and the surrounding textual context. §3.1 defines the five memory abilities, §3.2 describes the construction pipeline (Figure 1), §3.3 presents the quality control, and §3.4 describes the cross-modality validation.

3.1 Memory Abilities

To comprehensively reflect the retrieval, reasoning, and update that a conversational assistant must perform over a long multimodal history, MemLens formulates five core memory abilities, each broken into subtypes that target specific reasoning operations. *Information Extraction (IE)* tests the recall of a specific fact from a single evidence session, either by identifying an abstracted entity in the evidence image and then retrieving the associated information from the surrounding text (*Entity*), or by recalling a visual detail from an image shared in an earlier session (*PrevInfo*); the visual identification hop draws on fine-grained recognition, image–text alignment, object counting, spatial reasoning, and numerical reasoning. *Multi-Session Reasoning (MSR)* evaluates aggregation across three to eight sessions, through counting unique items identified only by their evidence images, summing values stated in text or embedded in visual artifacts, and resolving co-reference by comparing images across sessions. *Temporal Reasoning (TR)* assesses joint reasoning over temporal references and visual content, comparing two intervals derived from textual or visual cues and grounding events chronologically or by date, with temporal expressions themselves replaced by visual artifacts such as clock faces and calendar pages. *Knowledge Update (KU)* tracks an evolving user attribute across a chain of four successive updates (Xu et al., 2024) (e.g., "I used to like apples" $\rightarrow$ "now I prefer kiwi"), asking for either the current state or a specified historical one. *Answer Refusal (AR)* removes all supporting evidence from an otherwise answerable question, so a correct model must decline rather than hallucinate (Zhang et al., 2024a); AR serves as a calibration check for hallucination detection, not a core memory retrieval task. A single cross-modal principle unifies the four answerable types:

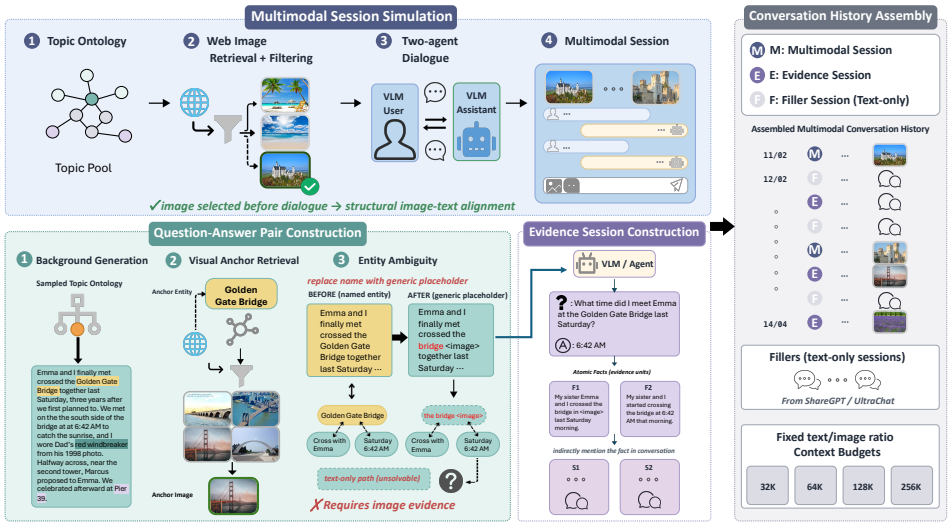


Figure 1: MEMLENS construction pipeline.

the evidence image carries information that the text deliberately withholds, so correct answers require joint visual–textual reasoning (§3.2); §3.4 validates this dependency empirically. The formal problem specification appears in Appendix C.1. Appendix I grounds this taxonomy in established memory research, and Appendix C.3 provides the complete subtype breakdown with representative examples.

3.2 DATA CURATION

The construction proceeds in four components (Figure 1): (i) first, multimodal session simulation generates topic-grounded multimodal dialogue sessions that form the conversation background; (ii) next, question construction produces evaluation questions whose answers require visual content through entity abstraction; (iii) evidence session construction wraps each evidence fact into a complete session that matches its topical and stylistic profile; and (iv) conversation history assembly interleaves evidence, haystack, and text-only filler sessions in timestamp order and compiles each question into four standardized input lengths.

Multimodal session construction. Each session begins with topic sampling from a hierarchical ontology. For each topic, we generate an image query to retrieve a batch of candidate images from the web[1]. We then filter these candidates for visual relevance. From this filtered set, we select the final image. Given this image, we generate sessions through a dual-model simulation with GPT-5.1 (OpenAI, 2025) as the user and Gemini-3-Pro (Google DeepMind, 2026b) as the assistant, producing multi-turn dialogues that incorporate the selected images into interleaved image-text sequences (Laurençon et al., 2023). Because image selection precedes dialogue generation, image–text alignment is a structural property of the data rather than a post-hoc filter. The image retrieval pipeline (§C.5), full topic ontology (§C.2), and image diversity statistics (§C.6) are detailed in the Appendix.

Question-Answer Pair. Question generation follows a four-step pipeline. (i) A topic is sampled from the hierarchical ontology, and a LVLM generates a background paragraph containing salient named-entities, for instance, a paragraph about San Francisco landmarks that mentions the Golden Gate Bridge. (ii) One entity is selected as the visual anchor for the question. A text query derived from the entity is issued to the same web-crawling and multi-model scoring pipeline used for haystack images (§C.5), yielding a high-relevance image (Hu et al., 2023); the anchor may range from a famous landmark to a specific product model. (iii) Cross-modal dependency is enforced through *entity abstraction* (Wang et al., 2024e;d; Lerner et al., 2024): the entity in the background paragraph is replaced with a higher level concept drawn from a dictionary spanning 55 entity

[1] `https://github.com/hellock/icrawler`; sourcing, licensing, privacy, and release detailed in Appendix C.7.

categories (e.g., "Golden Gate Bridge" → "the bridge shown in <image>"). After replacement, the paragraph no longer names the entity, and only the evidence image can resolve the reference. (iv) The abstracted paragraph, evidence image, and original entity name are provided to Gemini-3-Pro (Google DeepMind, 2026b), which generates a (question, answer) pair together with atomic evidence facts. The generation prompt constrains the question to require both the image and the surrounding text, closing the text-only shortcut documented in prior work (Maharana et al., 2024). The full pipeline and per-subtype generation routes appear in Appendix C.4.

Evidence session. Directly inserting evidence statements into the conversation history creates abrupt semantic shifts that make the evidence trivially locatable by similarity-based retrieval. Prior work has shown that increasing contextual similarity between evidence and distractors raises retrieval difficulty, a design principle also adopted by LongMemEval (Wu et al., 2025). Accordingly, each evidence fact is wrapped in a complete evidence session generated using the same pipeline as haystack sessions but grounded on the evidence fact rather than a sampled topic. Consequently, the evidence facts match the topical and stylistic profile of the surrounding haystack, so that evidence cannot be located by surface-level shortcuts. To further increase the difficulty, we also prompt the user model to mention the facts indirectly without stressing them: To embed the fact "I started a new job last month," the user turn might open by asking about updating the tax withholding and mention the job change incidentally later. When an evidence fact carries an evidence image, the image is placed adjacent to the corresponding textual mention within the session, preserving unambiguous image-text co-reference after entity abstraction.

Conversation history assembly. For each question, evidence sessions are inserted into a timestamp-ordered history of haystack sessions, with positions chosen uniformly at random (for KU, we preserve the evidence-session order; Appendix C.8). Haystack sessions are curated to be contextually related but uninformative for the question, and never include answer-relevant details. We vary the number of haystack sessions to produce four standardized context lengths (32K/64K/128K/256K tokens) using the cross-modal counting scheme of MMLongBench (Wang et al., 2025d) (Table 2). To avoid revealing evidence positions via image clustering, we keep a fixed text-per-image ratio across the history, padding with text-only filler sessions from ShareGPT and UltraChat (Ding et al., 2023). A post-hoc classifier achieves only marginally above-chance accuracy when separating evidence from haystack text (Appendix G.8), and the generator scales beyond 256K.

Table 2: Dataset statistics.

Statistic	Value
Questions	789
Types / Subtypes	5 / 9
Evidence sessions	2,145
Avg. turns/session	∼10
Avg. images/session	∼1.5
Tokens/image	∼2,000
	32K → 256K
Sessions/instance	14 → 93
Images/instance	20 → 138

3.3 Quality Control

Automated filtering. Each candidate question is screened by two automated checks. A rule-based pre-filter removes images from the question and evidence and drops cases where the remaining text already determines the answer. An LLM judge (GPT-5.1 (OpenAI, 2025)) then sees only the question text (without evidence or conversation history) and removes items solvable from parametric knowledge alone (Mallen et al., 2023). Remaining questions therefore require using the evidence image, which is a prerequisite for the multimodal analyses below.

Human review. Three rounds of human review operate as a cumulative quality gate on the filtered questions. Round 1 verifies that the evidence image carries answer-critical information; for AR questions, which are constructed by removing the evidence facts, this round instead confirms that the removed facts were answer-critical. Round 2 checks that evidence facts are naturally embedded and recoverable from their sessions, and that each session reads as a plausible user–assistant exchange with natural dialogue flow; sessions that fail the naturalness criterion are subsequently refined by humans. Round 3 reviews the haystack sessions to assess both the quality of the images and the naturalness of the dialogues. Together, the automated filters and three-round review reduce the initial pool of 20k candidates to the final 789 questions. The filler sessions drawn from ShareGPT and

Table 3: Cross-modality ablation on the image-essential and image-supportive questions. With evidence: question + evidence facts + evidence images, no haystack ($n = 634$). W/o evidence: question + evidence facts, no evidence images.

Model	Input	Overall	IE	MSR	TR	KU	Δ
GPT-5.4	With evidence image	92.43	94.31	100.00	96.91	75.86	—
	W/o evidence image	1.74	0.41	0.00	5.15	0.00	−90.69
Gemini-3.1-Pro	With evidence image	90.06	89.02	89.74	96.39	81.90	—
	W/o evidence image	1.89	0.00	0.00	6.19	0.00	−88.17

UltraChat (§3.2) further ground the surrounding context in authentic user–AI conversational patterns. Annotation guidelines and inter-annotator agreement appear in Appendix E.1.

3.4 Cross-modality Validation

The construction pipeline targets image-necessary questions: across MemLens, 65.7% of questions are image-essential (the answer is unrecoverable without the evidence image), 14.7% are image-supportive (the image confirms or disambiguates a textual fact), and 19.6% are text-sufficient (all AR questions plus a subset of MSR items that test cross-session reasoning over purely textual evidence). We verify this cross-modal dependency with two empirical checks (Table 3). An answerability test supplies each image-essential and image-supportive question ($n = 634$) with its full evidence (textual facts and evidence images) and confirms that the questions are answerable: GPT-5.4 reaches 92.43% overall and Gemini-3.1-Pro 90.06%. A multimodal ablation then removes all evidence images: overall accuracy collapses below 2% for both models. Two frontier proprietary LVLMs converge on near-identical collapses, showing our questions are highly multimodal.

4 Evaluation and Analysis

4.1 Experimental Setup

We evaluate 27 LVLMs and seven memory-augmented agents on MemLens. The LVLMs span closed-source systems including GPT-5.4 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), and Gemini-3.1-Pro (Google DeepMind, 2026a), alongside major open-source families such as Kimi-K2.5 (Team, 2026a), Qwen3.5 (Qwen Team, 2026), GLM-4.6V (Zhipu AI, 2025), and Gemma3 (Gemma Team, 2025). The memory-augmented agents comprise three multimodal pipelines (M3-Agent (Long et al., 2025), M2A (Feng et al., 2026), and M3C (Jang et al., 2025)) and four text-only pipelines (Mem0 (Yadav et al., 2025), MemOS (Li et al., 2025), MemAgent-7B (Yu et al., 2025), and Memory-T1 (Du et al., 2025b)). Since the four text-only frameworks lack image-processing interfaces, we follow standard practice and replace the image inputs with captions generated by BLIP-2 (Li et al., 2023); the three multimodal agents use their native multimodal input interfaces. We report LLM-as-Judge accuracy (Zheng et al., 2023). The complete model roster, backbones, input adapters, context-length coverage, decoding budgets, and judge protocol are detailed in Appendix B.

4.2 Main Results

Overall performance. Figure 2 reports per-type accuracy across context lengths for a representative 13-LVLM cohort and all seven memory agents; the full 27-LVLM roster is in Appendix G.1. At 32K, the top eight LVLMs fall within a 6.34-point band, with the best reaching 58.68%, so frontier systems cluster closely in short-context accuracy. At 128K, Gemini-3.1-Pro and Kimi-K2.5 reach 50.65% and 50.95%, down 3.45 and 3.93 points; Qwen3.5-122B-A10B falls by 14.65 points to 44.03%. AR has the largest mean decline in the fixed LVLM cohort, motivating closer examination of abstention at long contexts. The memory agents span 15.38–33.46% Overall accuracy at 32K: MemOS leads at 33.46%, followed by Memory-T1 at 28.72% and Mem0 at 27.12%. MemAgent-7B, M3-Agent, M3C, and M2A reach 25.86%, 23.57%, 20.91%, and 15.38%. These pipelines trail leading LVLMs on IE and KU, while their AR scores vary widely.

	IE			MSR			TR			KU			AR			Overall		
	32K	64K	128K	32K	64K	128K	32K	64K	128K	32K	64K	128K	32K	64K	128K	32K	64K	128K
Claude Sonnet 4.5	28.1	26.4	19.5	13.3	9.1	2.8	40.2	35.0	33.0	29.3	21.6	16.4	97.8	94.4	93.3	36.5	32.5	27.8
Gemini-3.1-Pro	63.6	58.9	55.8	32.2	31.5	29.4	40.9	42.3	41.2	49.1	48.6	46.0	97.8	95.6	96.7	54.1	52.5	50.6
GPT-5.4	63.7	63.0	60.2	28.2	27.3	21.7	39.2	42.3	39.7	47.4	43.1	43.1	97.8	96.7	94.4	52.7	52.3	49.6
Kimi-K2.5	51.2	52.4	51.6	39.9	35.7	28.7	53.1	52.1	48.2	50.9	49.1	48.6	97.8	97.8	93.3	54.9	54.0	51.0
Qwen3.5-122B-A10B	74.4	67.1	43.1	30.1	27.3	22.4	51.5	51.5	41.8	49.1	48.6	46.0	88.9	84.4	83.3	58.7	55.3	44.0
Qwen3.5-27B	70.3	62.2	54.1	29.4	21.0	14.0	39.2	35.6	34.0	42.2	44.0	38.8	86.7	67.8	63.3	53.0	46.1	40.7
Nemotron-Nano-12B	63.8	56.1	—	16.8	21.0	—	44.9	42.8	—	39.7	34.5	—	45.6	35.6	—	45.0	40.9	—
Cosmos-Reason2-8B	55.7	46.3	37.0	23.8	20.3	20.3	48.5	45.9	43.8	31.9	24.1	12.9	68.9	54.4	35.6	46.1	39.2	31.9
Phi4-Multimodal	31.7	28.1	22.8	20.3	21.0	22.4	43.8	39.2	30.9	19.8	24.1	14.7	43.3	21.1	15.6	32.2	28.1	22.7
GLM-4.6V	59.4	61.8	52.9	20.3	23.8	23.1	53.1	53.1	43.8	44.0	34.5	28.4	93.3	64.4	30.0	52.3	49.0	39.0
Gemma3-27B	35.0	26.4	19.1	23.8	22.4	18.2	41.2	44.3	37.1	30.2	27.6	13.8	68.9	61.1	48.9	37.6	34.2	26.0
Qwen3-VL-235B (T)	51.6	38.6	29.3	25.2	18.2	15.4	49.5	42.8	43.3	36.2	35.3	30.2	88.9	75.6	67.8	48.3	39.7	34.7
Qwen3-VL-30B (T)	41.5	39.0	33.7	11.9	11.9	10.5	35.6	37.1	36.6	30.2	29.3	20.7	71.1	65.6	60.0	36.4	35.2	31.3
Mem0	8.5	10.2	8.9	17.5	16.1	17.5	29.4	27.8	27.8	22.4	15.5	15.5	94.4	92.2	90.0	27.1	25.7	25.3
MemOS	19.9	19.1	18.7	19.6	20.3	19.6	43.3	41.2	41.2	25.9	23.3	25.9	81.1	78.9	84.4	33.5	32.2	33.0
MemAgent-7B	18.7	16.3	13.8	21.0	20.3	17.5	46.9	47.4	44.8	28.4	21.6	17.2	4.4	7.8	6.7	25.9	24.5	21.8
M3-Agent	23.6	21.1	22.4	16.8	18.2	16.1	34.0	28.9	32.0	11.2	5.2	4.3	27.8	16.7	25.6	23.6	19.6	21.3
M3C	17.1	17.1	14.2	14.7	16.1	17.5	33.0	29.4	26.3	12.1	6.9	5.2	26.7	25.6	15.6	20.9	19.4	16.6
Memory-T1†	18.0	19.7	21.3	25.7	22.9	22.9	62.5	56.2	56.2	13.8	17.2	27.6	9.1	18.2	13.6	28.7	28.7	30.3
M2A†	34.4	18.0	21.3	8.6	14.3	11.4	2.1	25.0	27.1	0.0	0.0	0.0	22.7	13.6	13.6	15.4	15.9	16.9

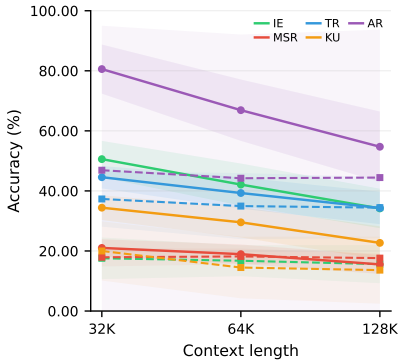
Figure 2: Per-type accuracy (%) by context length for 13 representative LVLMs and all seven memory agents. Each panel shows one question type; dashes indicate unsupported lengths. Dashed and dotted rules separate LVLMs from memory agents; † marks Memory-T1 and M2A; see Appendix B. Full rosters and protocols appear in Tables 13 and 14 and Appendix B.

Type-specific difficulty. Among LVLMs at 32K, per-type accuracy ceilings range from 97.78% on AR down to 39.90% on MSR (Kimi-K2.5). AR serves as a calibration check for hallucination detection, although its apparent ease erodes substantially at long contexts for open-weight LVLMs (e.g., GLM-4.6V drops from 93.33% at 32K to 30.00% at 128K). TR reaches a 60.82% ceiling; timestamps and dates in session metadata provide explicit retrieval anchors. IE (74.39% ceiling) requires visual grounding, with Entity questions demanding two-hop reasoning through the abstracted image reference. KU (50.86%) requires identifying the queried state within a four-fact update chain; missing the relevant image anchor can change the answer. MSR remains difficult because it requires aggregation over three to eight sessions: Kimi-K2.5 reaches 39.90%, Gemini-3.1-Pro 32.17%, and Qwen3.5-122B-A10B 30.07%; all other LVLMs and every memory agent stay below 30%. Within types, IE-Entity is harder than single-hop IE-PrevInfo, and MSR-Arithmetic is particularly difficult in the subtype diagnostic (Appendix C.3).

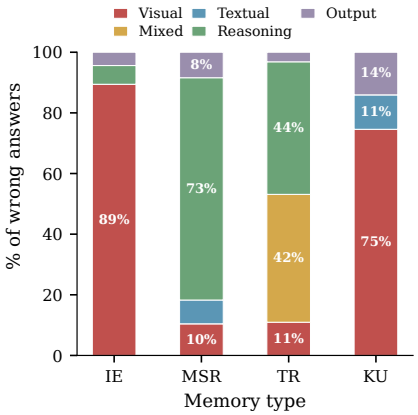
	IE	MSR	TR	KU	AR
GPT-5.4	63.7	28.2	39.2	47.4	97.8
Gemini-3.1-Pro	63.6	32.2	40.9	49.1	97.8
Kimi-K2.5	51.2	39.9	53.1	50.9	97.8
Qwen3.5 (2)	72.4	29.7	45.4	45.7	87.8
GLM-4.6V	59.4	20.3	53.1	44.0	93.3
Mem0	8.5	17.5	29.4	22.4	94.4
MemAgent-7B	18.7	21.0	46.9	28.4	4.4
Memory-T1†	18.0	25.7	62.5	13.8	9.1
M3-Agent	23.6	16.8	34.0	11.2	27.8

Figure 3: Memory-ability profiles at 32K. Qwen3.5 averages two models; † as in Figure 2.

No model dominates all memory abilities. No single model family dominates across all types (Figure 3). GLM-4.6V (Zhipu AI, 2025) reaches 53.09% on TR and 43.97% on KU, while Qwen3.5 (Qwen Team, 2026) inverts the pattern. Kimi-K2.5 (Team, 2026a) is relatively strongest on MSR at 32K, though this advantage fades at longer contexts. Gemini-3.1-Pro is among the models simultaneously competitive on IE, KU, and MSR at 128K. Similar Overall scores can mask substan-



(a) Per-type accuracy vs. input length. Solid: the same 25 LVLMs at all lengths; dashed: memory agents. Bands: 95% t intervals across model scores, not question-level intervals.



(b) Wrong-answer error decomposition at 128K context across the four answerable memory types, split by failure modality (visual, textual, mixed, reasoning, output). Category definitions appear in Table 19.

tial differences in individual abilities: at 32K, Kimi-K2.5 and Gemini-3.1-Pro differ by only 0.78 percentage points overall, but by 7.73 points on MSR. This contrast reinforces the need to compare ability profiles alongside aggregate scores. Memory agents exhibit an inverted profile: Memory-T1 reaches high TR accuracy (62.50%) through BM25 date matching but falls well below direct LVLMs on IE (18.03%), substituting keyword retrieval for the visual grounding that IE demands.

4.3 Analysis

Current memory pipelines lose faithfulness to original visual evidence. The gap between agents and LVLMs is pronounced on the visually grounded types (IE and KU). Despite their different input formats (§4.1), both text-only and multimodal pipelines compress evidence visual information into a fixed memory representation at storage time, leaving the original image pixels inaccessible at query time. Retrieval over these compressed encodings is also less reliable than retrieval over the raw conversation text. The loss is sharpest on IE and KU because captions and embeddings retain only the gist of an image and discard fine-grained visual cues, such as counts, attributes, and spatial relations, that these two types specifically probe. The same gap appears when blip-captions replace raw images for text-only agents, indicating that the bottleneck lies in lossy cross-modality storage rather than in the answering model. Closing this gap requires memory architectures that preserve image-level evidence rather than caption-based compression.

Memory-oriented post-training risks weakening abstention. At 32K, Mem0 and MemOS use frozen Qwen3-VL-8B backbones and achieve 94.44% and 81.11% AR. In contrast, the post-trained MemAgent-7B and M3-Agent reach only 4.44% and 27.78%; Memory-T1 reaches 9.09%. The distinction is how memory capability is acquired: Mem0 and MemOS add external memory without updating the answer backbone, whereas MemAgent-7B, Memory-T1, and M3-Agent use RL-trained models for memory-based answering (Yu et al., 2025; Du et al., 2025b; Long et al., 2025). Their reported reward functions emphasize answer correctness, with Memory-T1 additionally rewarding evidence grounding and temporal consistency, but include no dedicated objective for abstaining when evidence is absent. This contrast suggests that memory-oriented post-training can weaken refusal behavior when optimization emphasizes successful answering without explicitly preserving evidence-sensitive abstention. Future agent designs should therefore optimize memory access, answer accuracy, and correct refusal jointly (Zhai et al., 2026), rather than treating memory management as independent of hallucination control.

LVLMs and memory agents show complementary scaling trade-offs. Figure 4a plots accuracy across five abilities from 32K to 128K for 25 LVLMs and the memory agents. The two approaches respond to context scaling in structurally different ways. LVLMs degrade substantially on retrieval-heavy types, with IE and KU losing 16.42 and 11.72 percentage points respectively as evidence images become harder to locate among growing filler content. MSR shows a smaller absolute decline,

from 21.02% to 15.52%, reflecting a floor effect rather than genuine robustness. AR shows the steepest LVLM decline (80.58% at 32K to 54.71% at 128K), suggesting that growing filler content erodes abstention and pushes LVLMs to hallucinate on unanswerable questions rather than refuse them. The memory agents, by contrast, change by no more than 4.81 percentage points in Overall accuracy from 32K to 256K (Table 14) because their retrieve-then-reason pipeline largely decouples answering from input length. The two failure modes are orthogonal: LVLMs lose to context length, while memory agents lose to lossy multimodal compression at storage time. Because each architecture covers only one axis, scaling along that axis leaves the other failure mode untreated, motivating hybrid designs that span both axes.

Memory ability correlations reveal overlapping task demands. We analyze pairwise Spearman correlations among the five question types at 32K (Figure 5). Correlations vary substantially across type pairs, indicating that model rankings differ across abilities. IE and KU form the most consistent retrieval-oriented pair: they have the strongest correlation at 32K ($\rho = 0.79$), consistent with their shared need to first locate the relevant evidence image. KU additionally requires selecting the queried state across updates, and also correlates with AR at 32K ($\rho = 0.64$). In contrast, MSR shows weak correlations with IE and AR at 32K ($\rho = 0.21$ and 0.17), suggesting that single-session retrieval and refusal do not fully capture cross-session aggregation. These findings suggest overlapping difficulty axes, as MSR also correlates with KU ($\rho = 0.50$), rather than a clean separation between retrieval and aggregation. They also illustrate how a single aggregate score can obscure differences among long-context abilities, motivating per-type evaluation alongside overall performance (Hsieh et al., 2024; Yen et al., 2024).

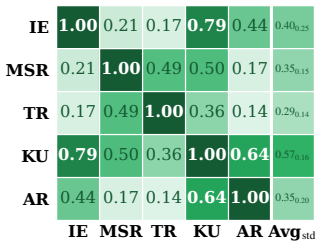


Figure 5: Spearman rank correlation (ρ) at 32K across all 34 evaluated LVLMs and memory agents.

Error analysis. Figure 4b decomposes wrong answers at 128K across the four answerable types, with the category definitions provided in Appendix G.4. On IE and KU, 89% and 75% of errors, respectively, fall in the Visual category, showing that the model fails to locate or read the evidence image; once the image is reached, the answer is usually extracted correctly. TR errors split between Mixed and Reasoning, reflecting both image-supportive grounding and small closed-set selection. MSR is the only type whose errors are dominated by the Reasoning category (73%). We further conduct an oracle-retrieval diagnostic in Appendix G.7, which restores accuracy on MSR to 89.74–100% when only the required textual facts and evidence images are supplied. This suggests a reasoning error at long context is largely induced by upstream retrieval failure, since the model computes correctly over the wrong subset of evidence. LVLMs and memory agents fail differently. LVLMs lose visual evidence as filler content accumulates in the context window, whereas memory agents lose it through lossy cross-modal compression at storage time. As memory grows, near-misses at 32K shift toward total misses at 128K (Appendix G.4) (Levy et al., 2024), suggesting that growth weakens evidence retrieval rather than reasoning. Future memory designs should therefore focus more on cross-modal evidence-retrieval fidelity instead of solely reasoning improvements.

5 Conclusion

We introduced MemLens, a benchmark for multimodal conversational memory that evaluates long-context LVLMs and memory-augmented agents across five memory abilities and four controlled context lengths. The task remains challenging: the strongest LVLM reaches 58.68% at 32K, while the best memory agent reaches 33.46%. LVLMs degrade sharply as input contexts grow, whereas memory agents remain more length-stable but trail on visually grounded retrieval. Memory-oriented post-training also risks weakening abstention, highlighting the need to optimize memory access, answer accuracy, and correct refusal jointly. Distinct ability profiles and cross-type correlations motivate per-type evaluation beyond a single Overall score. Error analysis and the facts-plus-images diagnostic point to visual-evidence retention and retrieval as central targets for improvement, alongside cross-session aggregation. These findings motivate hybrid architectures that combine long-context attention with structured multimodal retrieval to preserve the fidelity of visual evidence as conversations grow.

AI USE STATEMENT

We used generative AI for manuscript writing and editing, synthetic data construction, automated evaluation, and preparation of the released artifacts, as detailed in Appendix A. The research ideas and experimental design originated with the authors, who conducted the experiments and take responsibility for the manuscript and released artifacts.

ETHICS STATEMENT

MEMLENS is designed exclusively for evaluating multimodal long-term conversational memory in large vision-language models (LVLMs) and memory-augmented agents under a unified, length-controlled protocol. The benchmark is not intended as a training dataset for developing memory-equipped systems but rather as a controlled diagnostic for fair comparison across architectures. We release every artifact under frozen version tags so that any reported leaderboard cell remains traceable to the exact item set on which it was scored.

The 4,695 source images underlying MEMLENS are retrieved from public web image search via iCrawler against a non-person-centric topic ontology; candidates carrying watermarks, stock-photo logos, or copyright overlays are excluded at retrieval time, and the construction prompts never request the model to identify, name, or describe any depicted person. Author-produced artifacts (questions, evidence facts, prompts, and human-annotation records) are released under CC-BY-4.0 and the evaluation harness under MIT, while third-party images retain their original source-site licenses; per-image provenance metadata (source URL, retrieval timestamp, perceptual hash) accompanies the release, and a takedown contact in the project repository allows seven-day removal of any flagged image.

Human review is conducted by project members rather than crowd-workers, so no third-party data collection or human-subjects protocol is involved. Three sequential rounds (Appendix E.1) audit cross-modal necessity at the question level, naturalness and recoverability at the session level, and a stratified sample of haystack sessions before any item reaches the released set.

We caution that MEMLENS is not intended as a training dataset: exposure of the 789 evaluation items as supervised data would compromise their diagnostic value, and the datasheet explicitly discourages such use. While the benchmark is grounded in multi-session multimodal conversation, the five memory abilities it isolates (information extraction, multi-session reasoning, temporal reasoning, knowledge update, and answer refusal) describe general functional requirements of any long-horizon conversational assistant. We therefore expect MEMLENS to serve as a useful indicator of memory robustness in adjacent settings, such as document-grounded or voice-grounded assistants, where retrieval, integration, and calibrated abstention play comparable roles.

REPRODUCIBILITY STATEMENT

All closed-source LVLM evaluations and synthetic data construction were performed via the public APIs of OpenAI, Anthropic, and Google, with a total experimental cost of approximately 4,500 USD covering both pipeline construction and benchmarking. Open-weight LVLMs are served locally with vLLM v0.17–0.18 on 8$\times$A100-80GB nodes with tensor parallelism for context windows at and above 128K. Per-model decoding budgets, judge protocol, retrieval depth for memory agents, and adapter details distinguishing caption inputs for text-only agents and native visual interfaces for multimodal agents are documented in Appendix B. The four-stage construction pipeline (multimodal session simulation, question construction, evidence session construction, conversation history assembly) is described in Section 3, with per-subtype generation routes in Appendix C.4 and full prompt templates in Appendix F. Quality control combines automated filtering with a three-round human review (Section 3.3, Appendix E.1) and an LLM-as-Judge protocol cross-validated by an independent judge family and a three-annotator human consensus; agreement statistics and per-question-type breakdowns appear in Appendix E.2. The 789-question benchmark, 4,695 source images with provenance metadata, evaluation harness, and prompt templates are publicly released under frozen version tags at the anonymous dataset repository (dataset) and in the supplementary material (code) so that every reported leaderboard cell can be reproduced end-to-end.

REFERENCES

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, 2024.

Anthropic. Claude 3 model card. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf, 2024. Accessed: 2026-04-30.

Anthropic. Claude Sonnet 4.5 System Card. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024a.

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. *arXiv preprint arXiv:2412.15204*, 2024b.

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, 2026. URL https://arxiv.org/abs/2601.03515.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16495–16504, 2022.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023a. URL https://arxiv.org/abs/2306.15595.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions?, 2023b. URL https://arxiv.org/abs/2302.11713.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, CHAOQUN LIU, Maojia Song, Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. 2024.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding, 2024. URL https://arxiv.org/abs/2411.04952.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*, 2024.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, 2023.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending LLM context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering, 2024. URL https://arxiv.org/abs/2402.16288.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z. Pan. Rethinking memory in llm based agents: Representations, operations, and emerging topics, 2025a. URL https://arxiv.org/abs/2505.00675.

Yiming Du, Baojun Wang, Yifan Xiang, Zhaowei Wang, Wenyu Huang, Boyang Xue, Bin Liang, Xingshan Zeng, Fei Mi, Haoli Bai, Lifeng Shang, Jeff Z. Pan, Yuxin Jiang, and Kam-Fai Wong. Memory-t1: Reinforcement learning for temporal reasoning in multi-session agents, 2025b. URL https://arxiv.org/abs/2512.20092.

Yiming Du, Bingbing Wang, Yang He, Bin Liang, Baojun Wang, Zhongyang Li, Lin Gui, Jeff Z. Pan, Ruifeng Xu, and Kam-Fai Wong. Memguide: Intent-driven memory selection for goal-oriented multi-session llm agents, 2025c. URL https://arxiv.org/abs/2505.20231.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2025. URL https://arxiv.org/abs/2407.01449.

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Google DeepMind. Gemini 3.1 pro, February 2026a. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf. Model card.

Google DeepMind. Gemini 3 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-pro/, 2026b.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models, 2025. URL https://arxiv.org/abs/2405.14831.

Rongpei Hong, Jian Lang, Ting Zhong, Yong Wang, and Fan Zhou. TAMEing long contexts in personalization: Towards training-free and state-aware MLLM personalized assistant. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 452–463. ACM, April 2026. doi: 10.1145/3770854.3780214. URL https://doi.org/10.1145/3770854.3780214.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12065–12075, 2023.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating knowledge acquisition from multi-discipline professional videos, 2025a. URL https://arxiv.org/abs/2501.13826.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. MRAG-Bench: Vision-centric evaluation for retrieval-augmented multimodal models, 2025b. URL https://arxiv.org/abs/2410.08182.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions, 2026. URL https://arxiv.org/abs/2507.05257.

Jihyoung Jang, Minwook Bae, Minji Kim, Dilek Hakkani-Tur, and Hyounghun Kim. Enabling chatbots with eyes and ears: An immersive multimodal conversation system for dynamic interactions, 2025. URL https://arxiv.org/abs/2506.00421.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*, 2024.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*, 2024.

Greg Kamradt. Needle in a haystack — pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023. GitHub repository.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-5729. URL https://doi.org/10.52202/085713-5729.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Cross-modal retrieval for knowledge-based visual question answering, 2024. URL https://arxiv.org/abs/2401.05736.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*, 2024.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, Junpeng Ren, Zehao Lin, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhiqiang Yin, Qingchen Yu, Bo Tang, Hongkang Yang, Zhi-Qin John Xu, and Feiyu Xiong. Memos: An operating system for memory-augmented generation (mag) in large language models, 2025. URL https://arxiv.org/abs/2505.22101.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory, 2025. URL https://arxiv.org/abs/2508.09736.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations, 2024. URL https://arxiv.org/abs/2407.01523.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents, 2024. URL https://arxiv.org/abs/2402.17753.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents, 2025. URL https://arxiv.org/abs/2507.04590.

Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

NVIDIA. Cosmos-Reason2-8B Model Card. https://huggingface.co/nvidia/Cosmos-Reason2-8B, 2025a.

NVIDIA. Nvidia nemotron nano v2 vl, 2025b. URL https://arxiv.org/abs/2511.03929.

OpenAI. Gpt-4 technical report, 2023a. URL https://arxiv.org/abs/2303.08774.

OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023b. Accessed: 2026-04-30.

OpenAI. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023. URL https://arxiv.org/abs/2309.00071.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. https://www.alibabacloud.com/blog/qwen3-5-towards-native-multimodal-agents_602894, 2026.

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

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval, 2024.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. Technical report, Technical report, Bytedance, 2025. URL https://lf3-static. bytednsdoc. com ....

Bytedance Seed. Seed1. 8 model card: Towards generalized real-world agency. *arXiv preprint arXiv:2603.20633*, 2026.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024a.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024b.

Linxin Song, Taiwei Shi, and Jieyu Zhao. The hallucination tax of reinforcement finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2105–2120, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.112. URL https://aclanthology.org/2025.findings-emnlp.112/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2024. URL https://arxiv.org/abs/2104.09864.

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.

Kimi Team. Kimi k2.5: Visual agentic intelligence, 2026a. URL https://arxiv.org/abs/2602.02276.

V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026b. URL https://arxiv.org/abs/2507.01006.

Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. SCM: Enhancing large language model with self-controlled memory framework, 2025a. URL https://arxiv.org/abs/2304.13343.

Bingbing Wang, Yiming Du, Bin Liang, Zhixin Bai, Min Yang, Baojun Wang, Kam-Fai Wong, and Ruifeng Xu. A new formula for sticker retrieval: Reply with stickers in multi-modal and multi-session conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24): 25327–25335, Apr. 2025b. doi: 10.1609/aaai.v39i24.34720. URL https://ojs.aaai.org/index.php/AAAI/article/view/34720.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models, 2025c. URL https://arxiv.org/abs/2406.11230.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. Needle in a multimodal haystack, 2024b. URL https://arxiv.org/abs/2406.07230.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024c.

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. Subeventwriter: Iterative sub-event sequence generation with coherence controller. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1590–1604, 2022.

Zhaowei Wang, Quyet V Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. Cola: Contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5253–5271, 2023.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Wong, and Simon See. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 973–994, 2024d.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3991–4010, 2024e.

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Ye Tian, Yue Yang, Kaixin Ma, Xiaoman Pan, Yangqiu Song, and Dong Yu. Divscene: Towards open-vocabulary object navigation with large vision language models in diverse scenes. *arXiv preprint arXiv:2410.02730*, 2024f.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. MMLongBench: Benchmarking long-context vision-language models effectively and thoroughly, 2025d. URL https://arxiv.org/abs/2505.10610.

Shaohang Wei, Wei Li, Feifan Song, Wen Luo, Tianyi Zhuang, Haochen Tan, Zhijiang Guo, and Houfeng Wang. TIME: A multi-level benchmark for temporal reasoning of LLMs in real-world scenarios, 2025.

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

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024a. URL https://arxiv.org/abs/2407.15754.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024b. URL https://arxiv.org/abs/2412.10302.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents, 2025. URL https://arxiv.org/abs/2502.12110.

Deshraj Yadav, Taranjeet Singh, and Prashant Srivastava. Mem0: Building production-ready AI agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6569–6587, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.333. URL https://aclanthology.org/2025.emnlp-main.333/.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *CoRR*, 2024.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

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, 2025. URL https://arxiv.org/abs/2507.02259.

Skylar Zhai, Jingcheng Liang, and Dongyeop Kang. Abstain-R1: Calibrated abstention and post-refusal clarification via verifiable RL. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19674–19695, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.985. URL https://aclanthology.org/2026.findings-acl.985/.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say "I don't know". 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. 7113–7139, 2024a.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. ∞Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024b.

Jianjie Zheng, Zhichen Liu, Zhanyu Shen, Jingxiang Qu, Guanhua Chen, Yile Wang, Yang Xu, Yang Liu, and Sijie Cheng. Evaluating memory capability in continuous lifelog scenario, 2026. URL https://arxiv.org/abs/2604.11182. Findings of ACL 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhipu AI. GLM-4.6V Model Card. https://huggingface.co/zai-org/GLM-4.6V, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, 2023. URL https://arxiv.org/abs/2305.10250.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. URL https://arxiv.org/abs/2406.04264.

A USE OF LARGE LANGUAGE MODELS

We used ChatGPT and Claude to assist with manuscript writing and language editing. The research ideas, methodology, and experimental design originated with the authors, who conducted the experiments. AI writing assistance does not imply AI authorship of these research contributions.

As part of the benchmark construction pipeline, GPT-5.1 and Gemini-3-Pro generated synthetic dialogues, and Gemini-3-Pro generated questions and answers (Section 3.2). Automated filtering and human quality control are described in Section 3.3. Qwen3-VL-235B-A22B-Instruct scores the retained LVLM, Memory-T1, and M2A panels, while GPT-5.4-mini scores the five full-benchmark agents. Independent model and human agreement checks are described in Appendix E.2. These uses are distinct from evaluating models as benchmark subjects.

An AI coding assistant also helped prepare the anonymous release, including code maintenance, tests, documentation, and anonymization of data paths. Release checks covered dataset integrity and offline regression tests; they do not constitute a rerun of the reported model experiments. The authors retain responsibility for the final text, claims, experimental results, and released artifacts.

B EVALUATION SETUP

B.1 MODELS

We evaluate 27 LVLMs and seven memory-augmented agent systems. Because compute is limited and not all LVLMs support a 256K-token context, we evaluate the LVLMs at 32K, 64K, and 128K for a fair comparison, and the memory-augmented agents at all four lengths.

Closed-source API models. GPT-5.4 (OpenAI, 2025) (OpenAI), Claude Sonnet 4.5 (Anthropic, 2025) (Anthropic), and Gemini-3.1-Pro (Google DeepMind, 2026a) (Google).

Open-source LVLMs. Kimi-K2.5 (Team, 2026a) (Moonshot), Qwen3-VL family (Bai et al., 2025) (235B-A22B and 30B-A3B MoE; 8B, 4B, 2B dense) in both Instruct and Thinking modes, GLM-4.6V (Zhipu AI, 2025), GLM-4.5V (Team, 2026b), Gemma3 (Gemma Team, 2025) (27B, 12B, 4B), Phi4-Multimodal (Microsoft, 2025), Cosmos-Reason2-8B (NVIDIA, 2025a), Nemotron-Nano-12B (NVIDIA, 2025b), and the Qwen3.5 family (Qwen Team, 2026) (122B-A10B MoE; 27B, 9B, 4B, 2B dense).

Memory-augmented agents. We evaluate seven memory-augmented agents: Mem0 (Yadav et al., 2025), MemOS (Li et al., 2025), MemAgent-7B (Yu et al., 2025), Memory-T1 (Du et al., 2025b), M3-Agent (Long et al., 2025), M3C (Jang et al., 2025), and M2A (Feng et al., 2026). Mem0 and MemOS use Qwen3-VL-8B (Bai et al., 2025) with text-only inputs for memory processing and answering. Mem0 uses `mem0ai` 2.0.4 with `text-embedding-3-small` and top-20 vector retrieval; MemOS uses Qwen3-Embedding-0.6B, Neo4j/Qdrant storage, and top-20 retrieval in fast mode. MemAgent-7B uses its released RL-MemoryAgent-7B checkpoint with recurrent 5K-token chunks. M3-Agent uses its memorization and control checkpoints, `text-embedding-3-large`, and iterative graph retrieval (threshold 0.5, at most five rounds); complete entries preceding a truncated memory JSON response are recovered and consecutive duplicates removed. M3C combines its released retrieval model with a base Qwen2-VL-2B-Instruct answerer, retrieves three sessions, and caps the answer context at 12K characters; its retriever encodes the first four turns of each session, up to 500 characters per turn. Memory-T1 and M2A retain their reported Qwen2.5-3B and Qwen3-VL-8B configurations, respectively. Table 5 summarizes the input interfaces; Table 21 reports the separate backbone ablation.

Agent evaluation protocol. Mem0, MemOS, MemAgent-7B, M3-Agent, and M3C are evaluated on all 789 questions at each of 32K, 64K, 128K, and 256K. For Memory-T1 and M2A, we retain the 195-question canonical-subset results (Appendix G.2). These two pipelines are computationally expensive, and the M2A full-benchmark run does not cover all four lengths. Direct LVLMs are evaluated at 32K, 64K, and 128K. The four text-only agents (Mem0, MemOS, MemAgent-7B, and Memory-T1) receive BLIP-2 (Li et al., 2023) image captions as text. The three multimodal agents use

their native visual interfaces: M3-Agent reads composite session screenshots during memorization and retrieves textual graph entries for answering, while M3C and M2A receive the original images. Table 4 lists the direct-LVLM specifications.

Table 4: Model specifications for all 27 evaluated LVLMs. Length means the training length (default) or claimed context window (denoted by †). All models are instruction-tuned. "Image Proc." stands for Image Processing, which is mainly Dynamic Resolution ViT (Wang et al., 2024a) or Dynamic Tiling (Wu et al., 2024b). Most models extend context length via RoPE (Su et al., 2024) with position interpolation techniques (Chen et al., 2023a; Ding et al., 2024; Peng et al., 2023).

Name	Length	Image Proc.	# Params
Proprietary			
Claude Sonnet 4.5	200,000†	?	?
Gemini-3.1-Pro	1,048,576†	?	?
GPT-5.4	1,000,000†	?	?
Qwen3.5			
Qwen3.5-122B-A10B	262,144	Dynamic-Resolution ViT	122B
Qwen3.5-27B	262,144	Dynamic-Resolution ViT	27B
Qwen3.5-9B	262,144	Dynamic-Resolution ViT	9B
Qwen3.5-4B	262,144	Dynamic-Resolution ViT	4B
Qwen3.5-2B	262,144	Dynamic-Resolution ViT	2B
Other Open-Source			
Nemotron-Nano-12B	131,072	Dynamic Tiling	12B
Cosmos-Reason2-8B	262,144	Dynamic-Resolution ViT	8B
Phi4-Multimodal	131,072	Dynamic Tiling	5.6B
Kimi-K2.5	262,144†	Dynamic-Resolution ViT	1T
GLM			
GLM-4.6V	131,072	Dynamic Resolution ViT	106B
GLM-4.5V	65,536	Dynamic Resolution ViT	106B
Gemma3			
Gemma3-27B	131,072†	Dynamic Tiling	27B
Gemma3-12B	131,072†	Dynamic Tiling	12B
Gemma3-4B	131,072†	Dynamic Tiling	4B
Qwen3-VL			
Qwen3-VL-235B (T)	262,144	Dynamic-Resolution ViT	235B
Qwen3-VL-235B (I)	262,144	Dynamic-Resolution ViT	235B
Qwen3-VL-30B (T)	262,144	Dynamic-Resolution ViT	30B
Qwen3-VL-30B (I)	262,144	Dynamic-Resolution ViT	30B
Qwen3-VL-8B (T)	262,144	Dynamic-Resolution ViT	8B
Qwen3-VL-8B (I)	262,144	Dynamic-Resolution ViT	8B
Qwen3-VL-4B (T)	262,144	Dynamic-Resolution ViT	4B
Qwen3-VL-4B (I)	262,144	Dynamic-Resolution ViT	4B
Qwen3-VL-2B (T)	262,144	Dynamic-Resolution ViT	2B
Qwen3-VL-2B (I)	262,144	Dynamic-Resolution ViT	2B

Agent input-format adapters. Table 5 lists the input interfaces used in these evaluations. Mem0, MemOS, MemAgent-7B, and Memory-T1 receive BLIP-2 captions, while M3-Agent, M3C, and M2A use their native visual interfaces. M3-Agent extracts memory from rendered session screenshots, and its answer model reads retrieved text graph entries. M3C and M2A receive the original images through their multimodal memory interfaces. Direct LVLMs attend to the original interleaved conversation within their context window. These interfaces determine which visual evidence is available to each pipeline; cross-system accuracy therefore reflects both the input representation and retrieval and answering quality.

Table 5: Input interfaces of the evaluated memory agents. The two visual columns describe how images enter the memory store and what is available when answering. Direct LVLMs attend to the original interleaved conversation within their context window.

Agent	Answer backbone	Write-time visual	Answer-time visual
Mem0	Qwen3-VL-8B	BLIP-2 captions	Retrieved text memories
MemOS	Qwen3-VL-8B	BLIP-2 captions	Captions only
MemAgent-7B	RL-MemoryAgent-7B	BLIP-2 captions	Captions only
M3-Agent	M3-Agent-Control	Session screenshots	Retrieved text graph entries
M3C	Qwen2-VL-2B-Instruct (base)	Original images	Stored embeddings
Memory-T1	RL-trained Qwen2.5-3B	BLIP-2 captions	Captions only
M2A	Qwen3-VL-8B	Original images	Stored embeddings

B.2 METRICS

LLM-as-Judge (J). LLM-as-Judge accuracy (Zheng et al., 2023) is our primary metric. GPT-5.4-mini scores the five full-benchmark agents in Table 14, routing each question to its exact dataset-provided question-subtype template. The retained direct-LVLM, Memory-T1, and M2A accuracy panels use Qwen3-VL-235B-A22B-Instruct with thinking disabled. Separate diagnostics follow their stated protocols; error categorization, for example, also uses GPT-4o-mini (Appendix G.4). String-match metrics such as substring exact match fail across MEMLENS because answers span heterogeneous formats—binary choice, counts, currency and date values, ranked orderings, short fill-ins, and explicit refusals—and LVLMs frequently wrap the correct answer in a multi-sentence rationale or a thinking trace. The judge reads the question, reference answer, and raw model output and emits a binary correct/incorrect verdict under task-specific criteria (Appendix B). To prevent degenerate outputs from contaminating scores, the pipeline tail-truncates outputs beyond 6,000 characters, auto-zeros responses exceeding 500 parsed words, and instructs the judge to reject circular reasoning. Judge reliability is quantified in Appendix E.2: the Qwen3-VL-235B judge agrees with GPT-5.4-mini on 96.40% of a stratified 800-item re-judge sample (Cohen's $\kappa = 0.93$; Spearman $\rho = 0.97$ at the model level) and with a three-annotator human consensus on 93.60% of 484 items ($\kappa = 0.86$). These checks support agreement within the tested panels but do not guarantee every rank or calibrate cross-judge differences between the retained VLMs and the five full-benchmark agents. A format-dependent leniency on very short (1–3 word) answers is documented and corrected in Appendix E.2.

Substring Exact Match (SE). The SE metric checks whether the normalized reference answer appears as a substring of the model output. This rule is generous to verbose models but inflates scores when the output contains reasoning traces that happen to mention the reference string.

Coverage and Per-Answer Accuracy. Beyond overall accuracy, we decompose performance into Coverage (Cov; fraction of the 699 answerable questions the model attempts) and Per-Answer Accuracy (PA; accuracy on attempted answers only). The two components recover overall accuracy via $J \approx (\text{Cov} \times \text{PA} \times 699 + \text{AR}_{\text{correct}}) \,/\, 789$, where 699 is the answerable subset and 789 is the full benchmark; the decomposition exposes the coverage–accuracy trade-off discussed in Appendix G.3.

B.3 INFRASTRUCTURE

Local models are served via vLLM (v0.17–0.18) with FlashAttention-2 (Dao, 2023) on 8×A100-80GB nodes with tensor parallelism for 128K inputs. API models use provider endpoints with concurrent requests (4–8 threads). Generation length is set to 2,048 tokens for direct models and 16,384 for thinking models to accommodate reasoning traces.

C Dataset Construction Details

C.1 Problem Formulation

An evaluation instance in MemLens is the 4-tuple (S, q, I, a), where $S = [(t_1, M_1), \ldots, (t_N, M_N)]$ is a sequence of N time-stamped multi-turn sessions ($t_1 < \cdots < t_N$), each interleaving text and images; $\mathcal{V}(S)$ denotes the set of all images appearing in S, and $I \subseteq \mathcal{V}(S)$ is the subset that carries answer-critical visual information not recoverable from the surrounding text; q is a query targeting one of five memory abilities; and a is the gold answer, or the literal string `NOT_MENTIONED` for answer-refusal items. A correct system must (i) localize the relevant evidence sessions within a long, distractor-heavy history and (ii) ground its answer in cross-modal reasoning over I.

C.2 Topic Ontology

Each of the four answerable memory abilities (IE, MSR, TR, KU) is backed by a dedicated ontology of $\sim$100 topics organized into three complementary tracks: *identification* (recognizing real-world entities such as products, landmarks, and dishes), *experience* (everyday activities and lifestyle moments), and *document* (text-rich artifacts such as receipts, menus, and schedules). Each topic is further expanded into $\sim$30 fine-grained subtopics, yielding more than 12,000 subtopics across the four answerable types. Table 6 groups the de-duplicated topic titles into 19 sub-attributes, each illustrated by representative leaves.

Table 6: MemLens topic ontology. Three tracks decompose into 19 sub-attributes; each row lists a small sample of the de-duplicated topic titles drawn from the underlying $\sim$400-title pool.

Sub-attribute	**Representative topic titles**
Track 1: Identification — recognizing real-world entities ($\sim$40%)	
1.1 Retail & commerce	supermarket shelves; bakery counters; pharmacy aisles; hardware/tool shops; music and sports stores; thrift stores; street food carts; fish and flower markets; secondhand electronics stalls; warehouse club bulk aisles
1.2 Home objects & belongings	kitchen appliances; knife and utensil drawers; pantry; spice jars; bookshelves; CD/vinyl collections; gaming desk; TV streaming setup; closet; jewelry/watch layout; makeup vanity; perfume shelf; skincare bottles; pet accessories; board games
1.3 Vehicles & mobility	personal cars; scooters; motorcycles; bicycle helmets and locks; bike-sharing docks; parked bicycles; license plates; parking lot vehicles; public transit vehicles; e-scooter models on sidewalks
1.4 Urban environment & landmarks	street signs; shopfront names; building numbers; mailboxes and doorbells; landmarks and tourist attractions; street murals and sculptures; residential houses; recycling bins and trash trucks; playgrounds; basketball courts; public swimming pools
1.5 Workspace & institutions	office kitchen; supply closet; lobby directories; vending machines; public library shelves; gym equipment; pharmacy windows; temple/church exteriors
Track 2: Experience — everyday activities and lifestyle moments ($\sim$40%)	
2.1 Dining & food activities	casual restaurants; fast-food counters; food courts; cafes; street food walks; late-night noodle shops; home cooking; weekend brunch; batch cooking; baking experiments; cooking classes; asian grocery hunts; farmer market pickups
2.2 Fitness & outdoor recreation	indoor climbing; badminton; trampoline park; futsal; table tennis; jogging routes; soccer at local field; pickup basketball; public swimming; sunrise yoga; weekend cycling; camping; hiking; beach outings; ski trips

Continued on next page

Table 6: MEMLENS topic ontology. Three tracks decompose into 19 sub-attributes; each row lists a small sample of the de-duplicated topic titles drawn from the underlying ∼400-title pool. (Continued)

Sub-attribute	Representative topic titles
2.3 Social & community gatherings	house parties; potluck dinners; game and movie nights; karaoke rooms; casual bar meetups; community festivals; charity runs; language exchanges; quiz nights; board game cafés; dance socials; volunteer shifts; sports stadium visits
2.4 Cultural & entertainment outings	art museum visits; science museum visits; gallery visits; historic house tours; concerts and small gigs; movie theater; outdoor summer cinema; theme park rides; indoor arcades; cosplay meetups; comic store browsing; photography walks
2.5 Routines & errands	morning bathroom and kitchen routines; commuting by bus or train; weekend laundry folding; apartment cleaning; grocery top-up trips; salon and barber visits; pet grooming; ride-hailing pickups; fuel stops; DIY home repair; content drafting on phone
2.6 Travel & mobility outings	weekend road trips; vacation rental stays; hotel check-ins; airport and station transit; riverside walks; coastal promenades; neighborhood evening jogs; weekend tram rides; day-trip walks
2.7 Home life & indoor hobbies	morning balcony coffee; evening living-room TV; evening reading corner; family breakfast; hobby painting sessions; sewing and knitting; pottery class; home garden balcony; video recording for reels
Track 3: Document — text-rich artifacts (∼20%)	
3.1 Receipts & bills	grocery self-checkout receipts; restaurant takeout receipts; monthly utility bills; taxi and ride receipts; parking garage receipts; car service invoices; veterinarian receipts
3.2 Tickets, passes & itineraries	event tickets and seating plans; museum day passes; transit monthly passes; intercity bus tickets; hotel booking confirmations; train reservations; travel itineraries; gym class punch cards; university club ID cards
3.3 Schedules & calendars	school and work weekly schedules; printed school lunch menus; cinema film schedules; fitness class timetables; chore rotation charts; carpool rotation sheets; weekly meal plans; fitness challenge calendars
3.4 Contracts & application forms	rental leases; job offer letters; gym membership contracts; international visa packets; school field trip permission slips; gym class registration forms; university club membership forms; workplace sign-in sheets
3.5 Personal records & logs	personal budget sheets and bank statements; fitness tracker logs; meal tracking notebooks; handwritten to-do lists; travel packing checklists; recipe cards; phone contact lists; household grocery expense notebooks
3.6 Subscriptions & digital records	streaming subscription billing emails; digital movie rental history; online shopping order screenshots; streaming watch history lists; social media profile/settings screens; social media content calendars
3.7 Health & service records	medical prescriptions and pharmacy labels; clinic appointment cards; home appliance warranty cards; smartphone repair invoices; household appliance repair job cards; product warranty manuals

C.3 SUBTYPE DETAIL

Table 7 provides the complete subtype taxonomy of MEMLENS, listing the 8 answerable subtypes plus Answer Refusal with per-subtype question counts, the visual skill or reasoning operation each subtype isolates, and a representative example. Across the benchmark, 65.7% of questions are image-essential (the evidence image is required to recover the answer), 14.7% are image-supportive (the image confirms or disambiguates a textual fact that a strong text-only model could otherwise guess), and 19.6% are text-sufficient (all AR questions plus a subset of MSR items retained by design). The image-essential share is highest in IE and MSR, substantial in KU, and lower in TR, where a portion of items renders the temporal cue as explicit textual dates or session-boundary timestamps.

The cross-subtype correlation structure (Figure 6) confirms that the nine subtypes do not consistently correlate with each other, supporting per-type evaluation rather than a single aggregate.

Table 7: Complete subtype taxonomy of MEMLENS ($n = 789$). Each subtype isolates a distinct visual skill or reasoning operation. Example questions are shortened for space; full versions appear in the released dataset.

Subtype	Type	n	Skill Tested / Representative Example
Entity	IE	120	Identify a visually grounded entity via two-hop reasoning (disambiguation, alignment, counting, spatial, arithmetic). *"What phrase does that message spell out?"*
PrevInfo	IE	126	Recall a visual detail from an earlier session's image (screenshot, app interface, photo). *"What color is the engine bay painted?"*
Arithmetic	MSR	50	Sum or compute over prices/quantities scattered across 3–8 sessions. *"How much total have I spent on weights?"*
Counting	MSR	46	Count entities matching a criterion across sessions. *"How many cats do I have?"*
Entity	MSR	47	Resolve whether two cross-session references denote the same entity, either by counting distinct entities or by Y/N identity matching. *"Is the new bird the same species as Rio?"*
Duration Cmp	TR	91	Compare two durations derived from session timestamps and visual cues (clocks, calendars). *"Which duration is longer?"*
Temporal Grounding	TR	103	Locate when an event occurred—either by sorting events chronologically or extracting a specific date—using session timestamps, textual dates, and clock/calendar images. *"Sort these facts in chronological order." / "When did I complete X? (YYYY/MM/DD)"*
Update	KU	116	Identify the current or a specified historical state within a 4-fact preference chain. *"What do I prefer now?"*
Refusal	AR	90	Decline to answer when the evidence image has been deliberately removed from the context. *"How many cables are visible?" (image absent)*
Total		**789**	

Table 8 presents representative examples from each question type, illustrating the cross-modal reasoning chain required to reach the correct answer.

C.4 QUESTION CONSTRUCTION PIPELINE

The question construction pipeline produces cross-modal evaluation questions from a topic ontology through four steps: background generation, entity selection with image retrieval, entity abstraction, and question generation. This pipeline is shared across all four answerable question types (IE, MSR, TR, KU), with type-specific adaptations described in the per-subtype routes at the end of this section.

Background generation. For each question, a topic is sampled from the hierarchical ontology (§C.2), and Gemini-3-Pro generates a background paragraph of three to five sentences grounded in that topic. The paragraph is required to contain at least two named entities drawn from recognizable real-world referents, such as landmarks, commercial products, biological species, or cultural institutions, whose visual appearance is distinctive and web-searchable. For example, a paragraph about San Francisco landmarks might mention the Golden Gate Bridge, Alcatraz Island, and Fisherman's

Table 8: Representative evaluation questions from MEMLENS, one per question type. Each question requires joint visual and textual reasoning: the evidence image carries the discriminative information that the surrounding text deliberately withholds. Fact sketches are abbreviated; full evidence sessions appear in the released dataset.

Type	Question	Answer	Why Multimodal
IE	What shape is the Mount Fuji detail in the artwork my Tokyo friend was studying?	Triangular	Resolve "the artwork" from `<image>` of *The Great Wave off Kanagawa*, then extract visual shape.
MSR	How much total have I spent on Arkham Horror expansions recently?	$124.94	Sum $59.99 + $64.95 from two sessions; one price is visible only on the box `<image>`.
TR	Which duration is longer: my time living in London vs. my stint as a barista?	A	London start date is on a boarding pass `<image>`; barista dates are textual. Compare two durations.
KU	Back when I liked it mainly for quick, filling meals, what was my favorite?	Farro	Locate the historical stage specified by the question and identify the grain in its associated `<image>`.
AR	How many vertical cables are visible on one side of the bridge?	`NOT_MENTIONED`	The bridge photograph has been deliberately removed; a correct model must refuse.

Wharf, providing both visual anchors and textual context from which questions can be derived. The background paragraph serves a dual purpose: it supplies the factual context that the question will target, and it introduces the named entity that will undergo abstraction in the next stage.

Entity selection and image retrieval. From the background paragraph, one named entity is selected as the visual anchor for the question. The selection requires that the entity be visually identifiable in a photograph, sufficiently specific that a web image search returns relevant results, and critical to at least one fact stated in the paragraph. A text query derived from the entity name is then issued to the web-crawling pipeline, which retrieves a batch of candidate photographs using iCrawler. Each candidate is scored by the multi-model filter described in §C.5, combining CLIP (ViT-L/14), SigLIP (ViT-SO400M), and a text–text cosine similarity channel between the query and a BLIP-2 caption of the candidate image. The highest-scoring candidate that passes negative-content filtering and global evidence-image deduplication is selected as the evidence image. This retrieval pipeline is identical to the one used for haystack images (§3.2), ensuring uniform visual quality across evidence and background content. The selected entity may range from a famous landmark to a specific product model or a biological species.

Entity abstraction. The surface form of the selected entity is then masked in the background paragraph and replaced with a natural anaphor that references the evidence image (Lerner et al., 2024; Hu et al., 2025b; Chen et al., 2023b). The replacement proceeds in two stages. The entity is first classified into one of 55 semantic categories spanning places (museums, parks, restaurants, bridges, temples), organizations (companies, foundations, institutes), objects (books, paintings, vehicles, instruments), and generic fallbacks. A replacement phrase is then sampled from a type-aware dictionary of $\sim$170 entries, with each category providing three to four paraphrase variants to ensure lexical diversity across questions. For instance, "Golden Gate Bridge" may be replaced with "the bridge shown in `<image>`," while "Portland Art Museum" might become "the gallery I visited, shown in `<image>`." When the entity type cannot be matched to any dictionary category, a generic fallback such as "the place shown in `<image>`" is used.

After replacement, the background paragraph no longer contains the entity name. The anaphor is deliberately under-specified: "the bridge" could refer to any of thousands of bridges worldwide, and only the accompanying evidence image allows the reader to identify the specific referent. This design enforces a cross-modal dependency in which the text provides the factual context (e.g., year of construction, architectural style) while the image provides the entity identity, following the visual information-seeking paradigm established in prior work (Chen et al., 2023b; Lerner et al., 2024).

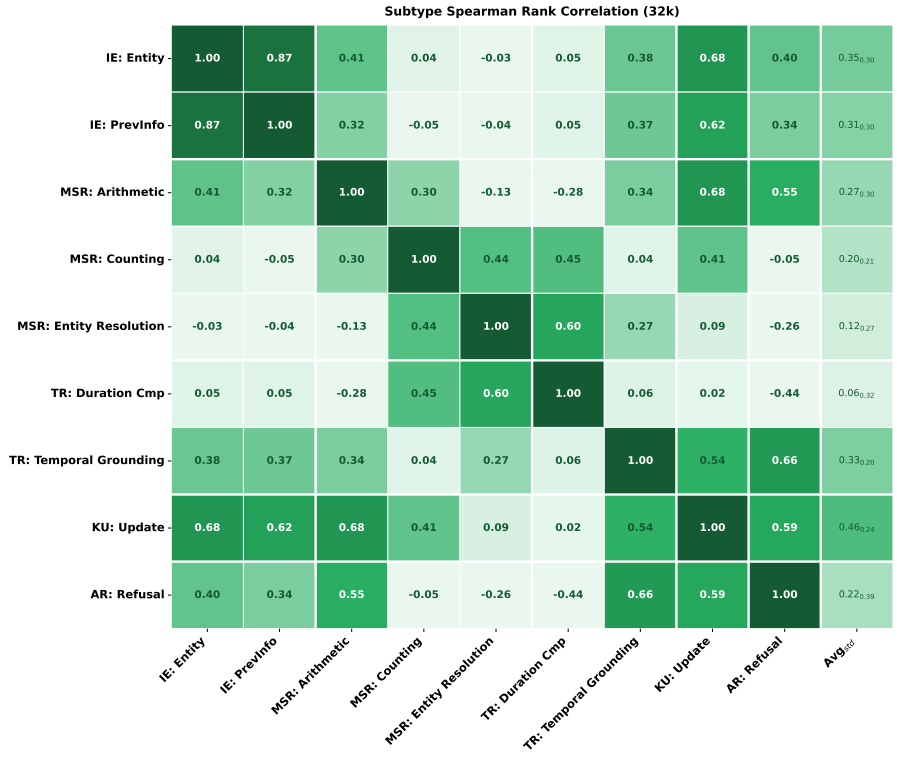


Figure 6: Cross-subtype Spearman rank correlation across the evaluated models ($n = 789$, 9 reporting subtypes). IE Entity and IE PrevInfo form the only near-ceiling pair ($\rho = 0.87$ at 32K, 0.94 at 128K), consistent with related task demands without isolating a shared internal skill. MSR-internal correlation is weak at 32K (mean pairwise $\rho = 0.20$) and rises to 0.38 at 128K while MSR scores remain low; the correlation alone does not establish a shared-failure mechanism; TR-internal correlation stays near zero at both contexts ($\rho = 0.06$). The heterogeneity supports evaluating all five major types separately rather than reporting a single aggregate (§4.3).

For KU questions, dictionary-based replacement alone is insufficient because the evolving attribute chain tracks concrete items within a single category (e.g., successive favorite fruits). In this case, an LLM rewrites each evidence fact with a short sensory or visual descriptor of at most five words that does not name the category. For instance, "blood orange" becomes "this tangy round thing `<image>`" and "blueberries" becomes "these tiny purple spheres `<image>`." The descriptor is constrained to be plausible for multiple items within the same category, preserving the ambiguity needed for cross-modal dependency.

Question generation and quality verification. The abstracted background paragraph, the evidence image, and the original entity name are provided to Gemini-3-Pro, which generates a (question, answer) pair together with one or more atomic evidence facts. The generation prompt enforces two constraints: the question must be answerable only when both the image and the surrounding text are available, and the answer must be derivable from the evidence facts without requiring external knowledge beyond the provided context. For MSR questions, a three-layer text-hackability defense additionally verifies that no textual cue leaks the entity identity: anti-leakage prompt rules prevent the generation model from naming the entity, a rule-based pre-filter rejects facts containing the entity name or close synonyms, and an LLM text-only judge confirms that the answer cannot be derived from the textual evidence alone. All generated questions then pass through the automated filtering and human review pipeline described in §3.3.

Per-subtype generation routes. The five question types follow different generation paths, each tailored to the visual skill under test:

- **IE – Entity two-hop** (5 subtypes, $n = 120$). An LLM generates a two-hop chain (Chang et al., 2022): the first hop resolves the entity from the evidence image, and the second hop retrieves a property of that entity from the surrounding text. The five subtypes vary the visual skill required for the first hop: disambiguation (distinguishing similar-looking entities), alignment (matching an image to a textual description), counting (enumerating items in the image), spatial reasoning (locating objects relative to each other), and arithmetic (computing from visually presented numbers).
- **IE – PrevInfo** (3 subtypes, $n = 126$). The question asks about a visual detail from an image shared in an earlier conversation session. The three subtypes correspond to the image source: a screenshot of a chat interface, an app or web interface, or a natural photograph. Entity abstraction is applied to the session reference rather than the entity itself.
- **KU – Update** ($n = 116$). A 4-fact atomic chain is generated: each fact updates a user attribute (e.g., favorite drink changes from tea $\rightarrow$ coffee $\rightarrow$ matcha $\rightarrow$ espresso martini). Questions ask for either the current state or a specified historical state within the chain. The four-fact construction does not imply that every question requires integrating all four facts. Entity abstraction masks the attribute's anchoring entity so the evidence image is needed to identify which preference chain is being queried.
- **MSR** (3 subtypes, $n = 143$). Facts are distributed across 3–8 sessions. Arithmetic ($n = 50$) requires summing prices or quantities; Counting ($n = 46$) requires enumerating entities matching a criterion; Entity ($n = 47$) requires determining whether two cross-session references denote the same entity, either by counting distinct entities or by Y/N identity matching. A 3-layer text-hackability defense (anti-leakage prompt rules, rule-based pre-filter, LLM text-only judge) ensures that the answer cannot be derived without the evidence images.
- **TR** (2 subtypes, $n = 194$). Duration Comparison ($n = 91$) derives two durations from session timestamps and visual cues (clocks, calendars) and asks which is longer. Temporal Grounding ($n = 103$) bundles two operations: order ranking ($n = 24$, sort events chronologically) and date extraction ($n = 79$, answer *"When did X happen?"* in `YYYY/MM/DD`). Three generation modes cross with these operations—Mode B renders the temporal cue itself as a visual artifact (clock, calendar, receipt), Mode C pairs an entity image with explicit textual dates, and Mode D pairs an entity image with session-level timestamps that serve as implicit temporal anchors—providing comprehensive coverage of temporal–visual integration.

C.5 Image Filtering

A unified cross-modal image filter is applied to every image in MemLens—both needle images in evidence sessions and filler images in haystack sessions—to maintain consistent visual quality across the benchmark. The filter operates in two stages.

Stage 1: Multi-channel relevance scoring. Each candidate image is scored against its associated textual query on three independent channels: CLIP (Radford et al., 2021) (ViT-L/14), SigLIP (Zhai et al., 2023) (ViT-SO400M), and a text–text cosine between the query and a BLIP-2 (Li et al., 2023) caption of the image. The CLIP threshold is set to 0.30. A candidate must exceed the CLIP threshold and at least one of the two secondary channels to pass.

Stage 2: Negative-content filtering. Candidates that pass relevance scoring undergo a negative-keyword filter that rejects images containing watermarks, stock-photo logos, copyright overlays, or resolution artifacts. For DocVQA-style images (Mathew et al., 2021) (receipts, menus, forms), a separate multimodal judge (GPT-4V) inspects each image for watermark presence and rejects any image flagged as containing artifacts that could distract from the document content.

Deduplication. A persistent URL registry tracks every image URL used across all generation batches, enforcing global uniqueness of evidence images: no two questions in MemLens share the same evidence image. Haystack filler images are drawn from the same filtered pool and may recur across questions. Within each question, duplicate detection uses perceptual hashing (pHash) with a

Hamming distance threshold of 6 to reject near-duplicate candidates from the same web-retrieval session.

C.6 IMAGE DIVERSITY

Unlike prior multimodal long-context benchmarks that categorize visual content by overlapping semantic topics (mixing scenes, media types, and tasks on a single axis), MEMLENS partitions images by their dominant visual format, with categories selected to cover distinct perceptual regimes (Table 9).

The first and largest category, natural photographs, covers real-world scenes, objects, food, and lifestyle items retrieved through a CLIP-filtered web pipeline. This category tests open-world object and scene recognition, spatial understanding, and visual attribute extraction in unstructured environments, and supplies the visual evidence for most IE, MSR, and KU questions. The second category, text-rich documents, includes receipts, menus, invoices, posters, price tags, and forms in which dense text and spatial layout carry the answer; reasoning over these images requires OCR together with layout grounding and text–visual alignment, the regime targeted by IE Entity alignment questions. The third category, digital and symbolic interfaces, covers app and web screenshots, chat interfaces, and synthetic clock and calendar renderings. These test interface-layout parsing and reading of synthetic symbols (clock hands, calendar grids, UI widgets) rather than natural-image recognition: an analog clock face or a settings panel exercises different perceptual regimes than a photograph. IE PrevInfo questions draw on screenshots, while TR Duration Comparison relies on synthetic clocks and calendars.

This distribution emerges from the CLIP-filtered web-retrieval pipeline rather than from manual balancing. Each category is anchored by its own topic subset in the ontology (§C.2), and images are retrieved per topic rather than per target proportion. As a result, MEMLENS evaluates multimodal evidence understanding across complementary perceptual regimes, and avoids the narrow object-centric or OCR-only focus of prior benchmarks.

Table 9: Image categories in MEMLENS, partitioned along a single capability-driven axis (dominant visual format). Each row corresponds to a distinct perceptual regime that the benchmark exercises.

Category	Description	Primary Types
Natural Photographs	Scenes, places, objects, food, and lifestyle items.	IE, MSR, KU
Text-Rich Documents	Receipts, menus, invoices, posters, price tags, and forms where dense text and spatial layout carry the answer.	IE Entity (alignment, receipt), MSR
Digital & Symbolic Interfaces	App and web screenshots, chat interfaces, and synthetic clock and calendar renderings.	IE Assistant PrevInfo, TR Duration Comparison

C.7 IMAGE SOURCING, LICENSING, AND RELEASE

Sourcing. Every image in MEMLENS originates from a public web image search issued through iCrawler, with queries derived from the topic ontology (§C.2) or the entity-abstraction slot for evidence images. iCrawler queries are issued against general-purpose image search rather than commercial stock aggregators, and the relevance filter described in Appendix C.5 is applied uniformly to evidence and haystack images.

Negative-content filtering. The negative-content filter rejects candidates carrying watermarks, stock-photo logos, copyright overlays, or resolution artifacts before any image is admitted into the benchmark, so that images exhibiting commercial-source markers are excluded at retrieval time rather than redistributed downstream. For DocVQA-style images (receipts, menus, posters, forms),

an additional multimodal judge (OpenAI, 2023b) inspects each image and rejects any flagged as containing watermarks or layout-distorting artifacts.

Privacy and identifiability. The topic ontology that drives image retrieval (§C.2) covers objects, places, products, text-rich documents, screenshots, and synthetic interfaces; no topic and no entity-abstraction slot is person-centric. iCrawler queries are derived from these topic and entity names rather than from any individual's name, and the retrieval pipeline performs no face- or identity-based search. Natural-photograph queries (e.g., *"Times Square at dusk"*) may include incidental human figures in the background, but the benchmark never targets identifiable individuals, and the construction prompts do not request the model to identify, name, or describe any depicted person. A takedown contact in the project repository allows seven-day removal of any image that, on user report, is found to reveal an identifiable individual.

Per-image metadata and datasheet. For each admitted image, the construction pipeline records its source URL, retrieval query, retrieval timestamp, BLIP-2 caption (Li et al., 2023), CLIP and SigLIP relevance scores, and perceptual hash, and a persistent URL registry enforces global uniqueness of evidence images across generation batches, so that no two questions in MEMLENS share the same evidence image. A datasheet for datasets (Gebru et al., 2021) accompanies the release at the same URL, documenting motivation, collection process, intended uses (evaluation only; see Appendix H), and licensing.

Release, license, and takedown. The 4,695 unique images referenced across the four context-length datasets are distributed alongside the dataset files at the anonymous dataset repository, together with the per-image metadata listed above (source URL, retrieval query, retrieval timestamp, and perceptual hash). The release is versioned with frozen tags so that any specific evaluation run remains reproducible. The author-produced artefacts of MEMLENS—dataset annotations (questions, evidence facts, abstracted paragraphs, consensus labels), per-image metadata, prompt templates, and human-annotation records—are released under CC-BY-4.0; the evaluation harness and supporting code in the supplementary material are released under MIT. Third-party images retrieved from public web search are *not* relicensed by the authors and remain governed by their original source-site licenses; we redistribute them solely to support reproducibility of the evaluation, and the per-image provenance metadata enables downstream users to independently verify or re-fetch the original source. A takedown contact is provided in the project repository, and any flagged image is removed within seven days.

C.8 CONVERSATION HISTORY ASSEMBLY

Evidence sessions are inserted into the full conversation history with positions chosen uniformly at random, except for KU questions where, while the position remains random, the relative order of the evidence sessions is kept the same since it is the user preference update order and critical to the answer.

D DATA EXAMPLES

D.1 INFORMATION EXTRACTION

Example Reasoning Chain — IE - Entity



Main Question. Examine the readout and provide the temperature value.
Answer with a short phrase only, no explanation.

Final Answer. 24

Evidence Facts:

1.(Image) I'm trying to make sense of which thermostat details actually matter before I start changing settings, especially since I don't want comfort fixes that quietly raise my bill. I've been looking at the thermostat setting on the left side of the display to make sure the AC isn't working harder than it needs to. ***Here's a photo of me checking smart home impact on my electricity bill in Singapore <image>.*** Based on that kind of setup, what would you pay closest attention to on the screen or in the menu to tell whether the fan and cooling behavior are set efficiently?

Figure 7: Sampled IE-Entity questions. The visually grounded entity is abstracted in the question text, so the agent must first identify the entity from the evidence image before retrieving the relevant fact.

Example Reasoning Chain — IE - PrevInfo



Main Question. What does the large text on the green wall of the supermarket say?
Answer with a short phrase only, no explanation.

Final Answer. Produce

Evidence Facts:

1. (Image) Okay, I'm staring at this wall of fruits and veggies, but what I actually want to understand is the labels at the vegetable counter <image>. Can you walk me through how those stickers usually break down price per pound, weight, and total cost so I can check that the register charge matches?

Figure 8: Sampled IE-PrevInfo questions. The answer is a visual detail (color, count, layout, on-screen text) of an image shared in an earlier session, requiring multi-session image recall.

D.2 MULTI-SESSION REASONING

Example Reasoning Chain — MSR - Arithmetic



Main Question. How much total have I spent on German language board games this month?
Answer with the exact amount (e.g., "$45.00") only.

Final Answer. $80

Evidence Facts:

1. (Text) Those examples are exactly the kind of thing I'm trying to internalize so I don't freeze mid-sentence during a heated turn. I'm especially interested in getting the nuances right between friendly teasing and something that sounds too harsh or formal, since tone feels tricky to judge in a second language. ***Speaking of which, I was organizing my games on my little Vienna kitchen table and noticed I'd dropped about 35 dollars on a German copy of Die Siedler von Catan***, but it actually feels worth it just to have everything written in the original German. Could you walk me through some core vocabulary for components and actions in both German and Spanish (like pieces, tiles, draw a card, roll the dice, shuffle, discard)?

2. (Image) I'd definitely like to dig into the sustainable materials angle, especially where there's tension between "luxury feel" and "eco" cues in what people keep. It's interesting because I catch myself doing this with games: if the box art feels special or a bit different, I'm way more inclined to keep that version. I just added another board game to the "stay" stack while deciding which copies to sell, and the one I'm keeping has slightly different cover art that feels more collectible to me. ***I actually have a picture of the stack with <image>that game on top here***, and I'm curious how much subtle artwork variants like this play into perceived long-term value or "display-worthiness."

3. (Text) The level of detail you're giving on what to look/listen for is exactly the kind of thing I worry I'm missing when I'm tempted by a "deal." Speaking of which, I've noticed how easy it is for me to rationalize prices when I'm in "collector brain" mode — ***like that game sitting on top of the stack in the picture I showed you earlier. I grabbed it earlier this month for about $45,*** which felt steep, but I convinced myself the German edition was worth it to round out the set alongside my Catan copy. How would you go about building amental cheat sheet for common repair ranges on the gear side so I can factor that into my buying decisions in real time?

Figure 9: Sampled MSR-Arithmetic questions. The agent sums or computes over prices, durations, or quantities scattered across sessions; at least one operand is visible only in an image.

Example Reasoning Chain — MSR - Counting



Main Question. How many educational board games do I currently have for the kids?
Answer with a single number only.

Final Answer. 4

Evidence Facts:

1. (Text) I'm really liking this whole idea of turning classic logic puzzles into things the kids can literally walk around inside; it feels way more likely to keep them engaged after a long school day. A life-sized Rush Hour setup sounds like exactly that kind of thing, and I'm curious how you'd structure the "cars" and rules so it stays strategic rather than devolving into chaos-tag. Speaking of which, ***I recently bought a Funskool Brainvita peg-jump puzzle for my oldest to stretch his logic muscles a bit***, and it's already sitting on the playroom shelf waiting for its turn in the rotation. How would you adapt the Rush Hour-style puzzle for, say, two kids plus an adult in a small living room space?

2. (Text) I really like the idea of turning the older kids into "teachers" so they're stretched instead of just racing to answer first; that feels like it would shift the mood from competition to collaboration. I'm especially interested in game formats that we can set up quickly after dinner without a lot of prep or printing. By the way, with my youngest ***I've been using a homemade "sight word bingo" board pretty often, where she has to read and cover simple English words***, and it's been one of the few things she'll willingly stick with. Could you suggest a few concrete word lists or themes that would work well for these mixed-level games—maybe by age band or difficulty level?

3. (Text) I like the idea of badges and a ship's log because my kids really respond when something feels "official" and ongoing, not just a one-off game. I'm curious how a reward store might fit into this without making it feel like they're only doing the missions for prizes instead of the story. Speaking of which, ***my sister gave them a board game called Sum Swamp for their birthday, and it's been a surprisingly big hit for practicing adding and subtracting*** without them noticing it's "math." How could I borrow the feel of that kind of game—lots of turns, simple decisions, a bit of luck—and blend it into a points or reward system for these campaigns?

4. (Image) That comparison between legacy-style campaigns and something like Hogwarts Battle is really helpful, especially the part about kids worrying less if "failure" is framed as part of the story. I'm still trying to balance that desire for meaningful consequences with not turning game night into a source of stress, especially for my more anxious kid. ***I just added another game to our collection on the playroom mat today***, and it seems like a solid fit for their age range—***I actually have a picture of it here on the mat with the other games <image>***. Based on what you can see in the photo and what we've been talking about, how would you ease kids into campaign/legacy elements so they stay excited about the ongoing story without fixating on "messing up" permanently?

Figure 10: Sampled MSR-Counting questions. The agent counts how many sessions or items match a given criterion across the conversation history.

Example Reasoning Chain — MSR - Entity Resolution



Main Question. Is the new creepy plant the same type of plant as Monty?
Answer with "Yes" or "No" only.

Final Answer. Yes

Evidence Facts:

1. (Text) I love these archetypes; they make it so much easier to "read" what a plant is asking for without turning it into a chore chart in my head. The idea of the respectful nod for snake plants especially makes me laugh because it's exactly how I already treat them without really thinking about it. Speaking of dramatic plants, ***I've got a big Monstera Deliciosa in my NYC place that I've been growing for years called Monty***, and he's just started putting out those fenestrated leaves and feels like this slowly emerging prima donna in the corner. Could you go a bit deeper into example names/rituals for a few other common types—like ZZ plants, philodendrons, or prayer plants—and then maybe circle back to what you mentioned about handling the "end of life" phase with named plants?

2. (Text) I'm trying to design a really immersive, horror-themed apartment setup for a classic scary movie night, and I want it to feel more like an experience than just people sitting on the couch. For example, I'm turning the living room into a mini Overlook Hotel for our Halloween screening of "The Shining," and ***I even picked up this oversized, ominous-looking plant*** with big holes in its leaves that matches all the Spirit Halloween props I put around the room — ***I'll show you a photo of it in a sec so you can see what I mean.*** I'd love help thinking through how to tie all the elements together so it feels cohesive and not just like a pile of random decorations. Can you suggest ways to structure the space, lighting, and small details so the whole evening feels like stepping into the movie's world?

3. (Image) The way certain textures feel like they "don't belong" in a supposedly cozy space is exactly what I'm trying to play with for this same gathering. I'm especially curious how a single object with a strange texture or silhouette can tip a room's mood toward uncanny even if everything else is pretty normal. Here's that ***creepy Shining-room plant I was just describing <image>*** — you can seethe slightly sinister, otherworldly vibe I was talking about. I'd love your take on how something like this might read in a photo depending on lighting and background textures around it.

4. (Text) I'm really into this idea of using the white walls almost like a projection screen for shadows instead of just blasting the room with overhead light. I'd love to get more specific about bulb choices and color temperatures, especially for keeping things cozy but still bright enough to read or work at night. Speaking of which, ***that same ominous Shining-room plant from the photo has earned a permanent spot in the center of the living room now,*** because when I dim the lights its outline on the wall turns into this really dramatic, sculptural shadow— those big fenestrated leaves throw such a recognizable silhouette. How would you mix warm vs. neutral bulbs in a small space like that so the silhouettes stay strong but the room doesn't feel gloomy or yellowed?

Figure 11: Sampled MSR-Entity Resolution questions. The agent decides whether two cross-session references denote the same entity, either via Yes/No identity matching or by counting distinct entities.

D.3 TEMPORAL REASONING

Example Reasoning Chain — TR - Duration (Comparison Explicit Temporal Information)



Main Question. Which of the following two durations is longer?
Duration 1: The time I lived in New York City
Duration 2: The time I spent working on my freelance writing project
A. Duration 1 is longer.
B. Duration 2 is longer.
Answer with "A" or "B" only.

Final Answer. A

Evidence Facts:

1. (Text) That distinction between restless avoidance and actual fatigue really lands for me, because my evenings tend to flip between those two and I don't always know which one I'm in until I've already lost a couple hours. I also like the idea of a five-minute test, but I think I'd need something even more concrete in the moment when my brain is foggy. Speaking of which, ***I came to New York City on June 9, 2009,*** feeling like I was stepping into an entirely new phase of my life, so this whole "unsettled after work" feeling is hitting something familiar for me. Could you give me maybe 3 really simple self-check questions I can ask myself at the end of the workday to decide whether I need rest, grounding, or a little structure?

2. (Image) I like the idea of smaller intentional moments more than one big send-off, because I think I'd get overwhelmed by a formal farewell party. I'm especially drawn to something that leaves people feeling connected instead of emotional in a stiff or awkward way. My own big departure felt very real when ***my long city adventure finally ended*** and I went to the airport for my flight back to London, and ***I actually have a image of the boarding pass from <image> that day***. How would you turn something like that into a simple farewell ritual or keepsake without it feeling too sentimental or forced?

3. (Text) That makes sense, especially the part about shifting from urgency to impact, because I tend to blur those together when I'm looking at a crowded list. I can usually tell what's important in theory, but in the moment every task starts making its own case and I end up second-guessing myself. Speaking of which, during a stay away from home ***I took on a freelance writing assignment and formally kicked it off on March 6, 2013***, and I remember having the same problem of not knowing whether to protect the writing time or clear the smaller obligations first. How would you set up a simple decision rule for days when paid work, personal creative goals, and admin all feel like they need attention right now?

4. (Text) I think the part I struggle with most is whether I need a little more time for the pattern to become obvious, or whether that's just me rationalizing extra effort. I'd like something even more concrete, maybe a very short scoring rule or threshold I can use at the end so I'm not relying on mood in the moment. Come to think of it, ***I had a small project once that didn't last long, and I officially closed it out on March 22, 2014***. Could you turn your checklist into a tiny yes/no decision tree I could reuse for habit tests and workflow experiments?

Figure 12: Sampled TR-Duration Comparison questions. The agent compares two time spans whose endpoints come from a mixture of textual dates, session timestamps, and visual cues.

Example Reasoning Chain — TR – Temporal Grounding (Implicit Temporal Information)



Main Question. When did I start living near Lake Zurich?
Answer with a date in YYYY/MM/DD format only.

Final Answer. 2019/03/01

Reasoning Chain:

1. (Image) That ridge-above-the-water idea is exactly the kind of tradeoff I'm trying to picture, because I want the views without ending up on a route that floods or gets jammed every holiday weekend. I'm also wondering how much difference there is between a town where the lake is kind of beside your commute versus one where the main road hugs the shoreline the whole way. Starting that spring, I really loved living near the water because it made my drive to work feel a lot more scenic, and ***I have a picture I took attached here*** <***image***>. When you're comparing places like that, what signs usually tell you the scenery is worth it without the day-to-day hassle becoming a pain?
2. (System Timestamp)
Session Date: 2019/03/01

Figure 13: Sampled TR-Temporal Grounding questions, including chronological ordering and absolute date extraction. The temporal cue is sometimes available only as an image of a clock face or a calendar page.

D.4 KNOWLEDGE UPDATE

Example Reasoning Chain — KU - Update









Main Question. What is my favorite hat now?
Answer with only the name or phrase, no explanation.

Final Answer. baseball cap

Evidence Facts:

1. (Image) I think the hard part for me is keeping the system short enough that I'll actually look at it when I'm rushing out the door. Since I switch bags so much, I'm realizing I probably need one visual cue for the pouch and a separate reminder for the weather-dependent stuff. ***At first I was really into this kind of hat that my son wore one time. Here's a photo of it <image>***—would you put something like that in the "add fresh each night" category, or is there a smarter way to build it into the checklist without making the list longer?

2. (Image) I've been thinking about the small accessories that anchor a daily ritual, like how one object can make the start of the day feel more settled and intentional. For me it's less about fashion and more about having a familiar thing that signals I'm in my usual rhythm. At one point, ***I began carrying a felt hat with a brim around with me, and it oddly became part of the routine I liked the most. Here's a photo of this kind of hat<image>***,and I'm curious what kinds of everyday accessories tend to create that same ritual feeling for other people.

3. (Image) I've been thinking a lot about how my clothes kind of mirror my mood, and I'm curious whether that's actually a common thing or if I'm just reading too much into it. ***Here's a photo of the hat <image> that I currently like the most, since accessories seem to be a big part of this for me.*** Lately, that big hat of mine has become one of those pieces I pick out more intentionally depending on how I'm feeling. Can you help me understand why fashion choices can feel so tied to emotional state?

4. (Image) I've been realizing that certain comfort items make me feel way more put-together and confident when I'm out, even if they seem small. For me, it's mostly hats and a couple of familiar accessories, and ***here's a photo of my favorite type of hat now <image> so you can see the style I mean***. Honestly, that hat has become such a go-to that I feel a little off leaving the house without it. Can you help me understand why something like that can feel so reassuring and confidence-boosting?

Figure 14: Sampled KU-Update questions. A four-step preference chain is anchored by a different image at each step; in this current-state example, the gold answer is the most recent state.

D.5 Answer Refusal

Example Reasoning Chain — AR - Refusal

Main Question. Which fast food brand has its logo displayed on the black sign directly above the store where I bought my jacket?

Final Answer. Insufficient information

Figure 15: Sampled AR-Refusal questions. The supporting evidence has been deliberately removed from the conversation history, so the gold answer is a refusal phrase rather than a content answer.

E Quality Assurance

E.1 Annotation Guidelines and Quality Assurance

Human review is organized in three sequential rounds, each targeting a different granularity of the benchmark construction pipeline.

Round 1: Question-level review. Four annotators participate in Round 1; each of the $n = 20k$ generated candidates is independently reviewed by two of the four. For each question, both reviewers verify:

1. The evidence image carries information vital to the answer—the question cannot be answered from text alone.
2. The question text does not leak the answer (e.g., the entity name does not appear in the question after abstraction).
3. The answer is unambiguous given the evidence.
4. The difficulty is calibrated: trivially easy items (e.g., "What color is the sky?") and impossibly hard items (requiring domain expertise beyond the conversation) are flagged for revision or removal.

Each of the two reviewers marks the item as *accept*, *revise*, or *reject*. Items marked *revise* by either of the two reviewers are rewritten to address the flagged issue and re-reviewed. Items marked *reject* by both reviewers are removed. Inter-annotator agreement on the accept/reject decision is $\kappa = 0.78$ (Cohen's κ (Cohen, 1960); $n = 200$ items sampled from the double-coded pool), indicating substantial agreement.

Round 2: Session-level review. Two annotators read every evidence session end-to-end ($n = 2{,}145$ sessions after Round 1 filtering, an average of 2.7 evidence sessions per question; see Table 2) and verifies:

1. All needle facts are present in the session dialogue and recoverable from it.
2. Needle facts are distributed across diverse conversational positions—not concentrated in a single turn.
3. The session reads naturally as a plausible user–assistant conversation.
4. The evidence image is placed adjacent to the corresponding textual mention, ensuring unambiguous image–text co-reference.

Sessions that fail any criterion are returned to the generation pipeline for regeneration with adjusted constraints.

Round 3: Haystack auditing. Two annotators inspect a stratified random sample of 500/689 multimodal haystack sessions (stratified by topic track: identification, experience, document) and verifies:

1. The haystack image is relevant to the topic of the conversation and the surrounding textual context.
2. The haystack session does not accidentally contain information that could serve as a spurious answer to any needle question.
3. Image quality meets a minimum standard (no broken images, no extreme blur, no primarily text-on-white stock images).
4. The dialogue progresses in a natural way and closely resembles real user–assistant interactions.

Multimodal sessions that fail quality checks are further rewritten and refined by human annotators manually.

Conversational naturalness. Naturalness is maintained through two complementary mechanisms. Round 2 review directly targets dialogue quality: evidence sessions that read as stilted or overly formal in register are returned to the generation pipeline for revision until the annotator is satisfied that the exchange is a plausible user–AI interaction. This criterion covers colloquial phrasing, turn-taking coherence, and the indirect embedding of factual content, following practices established in prior conversational memory benchmarks (Wu et al., 2025). Separately, the filler sessions that constitute the majority of each assembled context are drawn from ShareGPT and UltraChat (Ding et al., 2023), providing real user–AI conversations as the surrounding conversational frame rather than additional synthetic material. A post-hoc classifier separates evidence from haystack sessions with 57.92% accuracy (DeBERTa; Appendix G.8), detecting a small text-distribution difference. This control measures separability under the tested classifier, rather than conversational naturalness or the usefulness of these cues for retrieval.

Overall Quality Control. The three rounds collectively reduce the candidate pool from 20k to the final 789 questions. The primary reasons for removal are: answer leakage in the question text (23% of rejections), evidence image not carrying answer-critical information (31%), ambiguous or multi-interpretable answers (18%), and difficulty calibration failures (28%).

E.2 Judge Validation Details

A natural concern with LLM-as-Judge evaluation is systematic bias, particularly when the judge model belongs to the same family as evaluated models. We address this through cross-family validation and format-dependent bias correction.

Cross-family agreement is high. We re-evaluate a stratified sample of 800 items, drawn from 73,784 total judge calls (≈1.08% of the population), with GPT-5.4-mini as an independent judge; the sample combines 200 random, 250 targeted, and 350 GPT-only extended items to ensure coverage across model family, context length, question type, and judge score. Item-level agreement reaches 96.40% (Cohen's $\kappa = 0.93$), and the model-level ranking correlation is Spearman $\rho = 0.97$ ($p < 10^{-6}$), with a mean per-model accuracy delta of 3.70%. The high rank correlation indicates broad agreement in this validation panel, not identical rankings or a bound on cross-judge differences in the current full-agent comparison.

The judge agrees with human consensus on 93.60% of items. Three annotators independently labeled 484 items; disagreements were resolved to consensus. Against these consensus labels, our Qwen3-VL-235B judge reaches 93.60% raw agreement with Cohen's $\kappa = 0.86$. The errors are leniency-biased: 29 false positives versus 2 false negatives, meaning the judge credits borderline answers more often than rejecting correct ones. The per-question-type breakdown appears in Table 10 below.

The tested panel shows little additional Qwen-family leniency. Because the Qwen3-VL-235B-A22B-Instruct judge also appears as a benchmarked model (Table 13, row Qwen3-VL-235B (I)), we guard against self-favoritism with two independent oracles. First, the pre-specified cross-family test compares the Qwen-vs-GPT leniency gap on Qwen-family outputs (+3.00%) against non-Qwen

outputs (+2.70%), giving a difference of +0.33%—an order of magnitude below the 3% practical-significance threshold. Second, on the 484 human-annotated items, the judge's false-positive pattern is type-dependent (IE partial matches, AR hedge phrases) rather than family-dependent, suggesting a format- and type-related component to leniency without ruling out family effects outside the tested sample. Beyond family bias, we identify a format-dependent bias: the judge evaluates very short answers (1–3 words) more leniently. The retained evaluation applies the reported false-positive correction to six affected models. This validation and correction should not be read as a re-judging of the five full-benchmark agent runs.

Sampling design and human annotation protocol. This independent LLM-as-Judge validation uses a two-tier sample over its stored population of 73,784 judge calls, spanning roughly 92 model×context runs (each LVLM run on the full 789-question benchmark and each agent run on the 195-question canonical subset, plus auxiliary backbone-ablation runs; some LVLM runs do not support the 128K context). A core of 450 items (200 random + 250 targeted on hard cells) is annotated by both human raters and GPT-5.4-mini; an extended 350 items is re-judged by GPT-5.4-mini only, yielding 800 cross-judge items in total. Stratification spans model family (Qwen vs non-Qwen), context length (32K / 64K / 128K), question type, and judge score, so that rare cells—MSR arithmetic, KU stale-retrieval, and thinking-mode degenerate outputs—are represented at a minimum floor rather than by chance. All 484 items released for human verification received a consensus verdict; a further 120 are tagged for a future dedicated inter-annotator reliability study (item IDs in `double_annotation_item_ids.json`). The human annotation protocol used three annotators labeling each item across three rounds; whenever the three labels disagreed, annotators discussed the case and converged on a single consensus verdict, which is what we release as the human reference.

Cross-judge agreement by question type. Table 10 reports Qwen3-VL-235B judge vs GPT-5.4-mini accuracy on the 585 cross-judged items (after excluding 200 auto-zero cases and 15 empty outputs). The Qwen3-VL-235B judge is equally or slightly more lenient across every question type; the gap is largest on IE (-6.50%), driven by partial-match acceptances, and zero on MSR. The sign of the aggregate judge-score difference is consistent across these types; this does not establish unchanged model ranks.

Table 10: Cross-judge agreement by question type on the 585 cross-judged items. Negative Δ indicates the Qwen3-VL-235B judge is more lenient than GPT-5.4-mini; each Δ is computed from item counts before rounding. IE: Information Extraction; MSR: Multi-Session Reasoning; TR: Temporal Reasoning; KU: Knowledge Update; AR: Answer Refusal.

Type	Qwen acc (%)	GPT acc (%)	Δ (%)	n
IE	45.60	39.10	-6.50	169
MSR	34.40	34.40	$+0.00$	128
TR	47.70	46.30	-1.30	149
KU	30.30	27.60	-2.60	76
AR	63.50	60.30	-3.20	63

Human-vs-judge false-positive diagnosis. Table 11 breaks down the 29 false positives (judge=1, human consensus=0) from Appendix E.2 by question type and disagreement pattern. Two observations matter. First, 9 of the 10 AR false positives live in pre-retest runs that no longer feed any production leaderboard; on canonical retest runs, a deterministic substring rule for the canonical refusal phrase reaches 95.90% agreement with the human consensus (consistent with the AR row of the deterministic typed-accuracy audit; Table 12). Second, the 11 IE false positives reflect the same partial-match leniency observed in the aggregate cross-judge analysis—the Qwen3-VL-235B judge accepts short factual answers that GPT and the human consensus both reject—which identifies an answer-format association without isolating its cause or excluding family-specific effects.

Scope of the agreement estimates. The reported $\kappa = 0.86$ and 93.60% raw agreement quantify concordance with the adjudicated human-consensus labels on the sampled items. They do not measure agreement between individual annotators or establish a universal reliability threshold. The

Table 11: False-positive diagnosis on the 484-item human-annotated subset. FP = judge scored 1 where human consensus scored 0. Only 2 FN cases exist, so the asymmetry is strongly leniency-biased.

Type	FP count	Disagreement pattern
Information Extraction (entity + previnfo)	11	Partial match on short factual answers
MSR / TR	3	Edge cases
Knowledge Update	5	Verbose-correct vs literal mismatch
Answer Refusal	10	Hedge phrases credited as refusal

per-type breakdown in Table 10 and the false-positive diagnosis in Table 11 identify residual leniency, especially on partial-match IE answers. These audits support the use of the reported metric within their tested scope while leaving uncertainty in individual decisions, close rankings, and comparisons scored by different judges.

Deterministic typed-accuracy audit. We complement the 484-item human-consensus audit with a large-scale rule-based rescoring of every judge call whose reference answer is closed-form. Six of the nine reporting subtypes admit deterministic normalization, yielding seven rule families because TR Temporal Grounding splits into its order-ranking and date-extraction operations (§C.4): MSR Counting (integer match), MSR Arithmetic (currency-normalized scalar), MSR Yes/No (Yes/No identity match), TR Order Ranking (tuple of session indices), TR Duration Comparison (A/B label), TR Date Extraction (multi-format date canonicalization), and AR Answer Refusal (canonical refusal-phrase set). IE Entity, IE PrevInfo, and KU produce free-form short answers and remain LLM-judged. Rule families are assigned by answer form rather than by the subtype labels of Table 7, so MSR Entity items answered with a distinct-entity count fall under the integer-match rule and those answered by Yes/No under the identity rule; the per-family n therefore does not reproduce the subtype counts of Table 7. The stored diagnostic applies these rules to 34 evaluation rosters at 32K (27 LVLMs and seven agent configurations; 12,234 deterministic items), comparing each rule outcome with its LLM-judge label. This is a separate diagnostic panel, not deterministic rescoring of the five full-benchmark agent runs. Item-level agreement ranges from 87.4% on TR Order Ranking to 98.5% on MSR Arithmetic (Table 12); the count-weighted mean is 93.6%, in line with the 93.6% raw agreement on the human-consensus subset. The disagreement is systematically leniency-biased: the LLM judge credits a deterministically-wrong answer (*J-FP*) on a weighted 5.4% of items, against only 1.0% deterministically-correct answers rejected (*J-FN*). The largest leniency channels are TR Order Ranking (12.6% J-FP, partial credit on near-correct tuples), TR Date Extraction (8.6% J-FP, format-flexible date matching), and AR (5.7% J-FP, hedge phrases credited as refusal)—the same partial-match and hedge-phrase patterns identified on the 484-item human subset, but estimated here at $25\times$ the sample size. At the model-leaderboard level, the Spearman rank correlation between the LLM- and the deterministic-aggregated per-model accuracy across the 34 rosters is $\rho = 0.78$, the top-10 sets overlap on 7 of 10 entries, and within the LLM-top-10 the rank correlation is $\rho = 0.82$. The top-5 (Kimi-K2.5, Qwen3-VL-30B-Instruct, Qwen3.5-122B, Qwen3-VL-235B-Instruct, Qwen3-VL-8B-Instruct) is preserved under both metrics. The three rosters that drop out of the top-10 under deterministic rescoring—GLM-4.6V, Qwen3-VL-235B-Thinking, and Mem0 (GPT-4.1-mini backbone)—all produce verbose justifications that the LLM judge credits but the rule-based check rejects, consistent with the format-dependent leniency already corrected for short outputs in this section. The 5.4% over-credit and 1.0% under-credit rates imply an approximately 4.4-point net score difference on this rule-scored item pool. Preserved top-five membership coexists with changes elsewhere in the ranking; this diagnostic does not establish unchanged ranks or an error bound for the current full-agent panel.

F Prompt Templates

This appendix documents the prompt templates that drive the construction and evaluation of MemLens. The selection covers one canonical template per pipeline stage: the user prompt and judge rubric used at evaluation time, the persona-driven dialogue prompt used for haystack sessions, the question-generation prompt for each of the five major question types, the assistant template that builds evidence sessions, and the text-only judge that gates anti-shortcut filtering. Helper prompts for

Table 12: Deterministic typed-accuracy audit across 34 evaluation rosters at 32K. Rows are rule families keyed to answer form, not to the reporting subtypes of Table 7. *Agree* is the per-item agreement between the LLM judge and the rule-based label. *J-FP* marks LLM-judge over-credits (judge=1, deterministic=0); *J-FN* marks LLM-judge under-credits (judge=0, deterministic=1). n counts (model, item) pairs in this stored diagnostic; its agent configurations use the canonical subset and are distinct from the full-benchmark agent panel.

Rule family	n	**LLM (%)**	**Det (%)**	**Agree (%)**	**J-FP (%)**	**J-FN (%)**
MSR Counting	1,551	11.7	10.6	97.6	1.7	0.7
MSR Arithmetic	1,440	6.8	5.5	98.5	1.4	0.1
MSR Yes/No	1,082	47.3	42.2	94.7	5.2	0.1
TR Order Ranking	661	27.5	15.0	87.4	12.6	0.0
TR Duration Comparison	2,632	36.8	31.8	95.0	5.0	0.0
TR Date Extraction	2,285	53.0	47.3	88.3	8.6	3.1
AR Answer Refusal	2,583	78.1	73.8	92.9	5.7	1.4
Aggregate (weighted)	**12,234**	**42.3**	**37.8**	**93.6**	**5.4**	**1.0**

image search, and intermediate validators share the same skeleton as the canonical templates and are omitted here; the full set is available in the code release. Within each template, identifiers in curly braces (for example `{context}`, `{question}`, `{theme}`) are runtime placeholders that the pipeline substitutes before the call is issued.

F.1 Evaluation Prompts

This subsection lists the two prompts that drive the evaluation pipeline of MemLens. The first prompt is the user message sent to every Vision-Language Model in the main evaluation. The conversation history, with images already inlined as `<image>` tokens, is substituted into the `{context}` placeholder, and the instruction line is fixed to `Directly output the answer with no extra output` for every result reported in the paper (the chain-of-thought and structured-reasoning variants present in the codebase are not part of the reported numbers). The second prompt is the rubric followed by the LLM-as-Judge that produces the per-question accuracy scores in the main results table. The system block encodes a final-answer-extraction policy that is robust to long thinking traces and to circular reasoning, after which a task-specific criterion is appended at runtime; the Information Extraction criterion is reproduced below as a representative example, while the remaining seven criteria (MSR Yes/No, MSR Counting, MSR Arithmetic, TR Duration Comparison, TR Order Ranking, TR Date Extraction, KU Knowledge Update, AR Answer Refusal) follow the same format. Each judge call is cached in a SQLite store keyed by question, reference, prediction, and judge identifier, so that re-runs over the same outputs remain deterministic.

LVLM Evaluator User Prompt

```
Provide answers based on the given conversation history. If the question
cannot be answered based on the given conversation, respond with
"Insufficient information".
Conversation:
{context}

Directly output the answer with no extra output.
Question Date: {question_date}
Question: {question}
```

For answer-refusal items, the dataset stores the literal string `NOT_MENTIONED` as the gold answer (Appendix F.3), while the eval prompt above instructs the model to output `Insufficient information` as its surface refusal phrase. The judge maps any prediction in a canonical refusal-phrase set to a successful abstention, so the gold token in the dataset and the natural-language phrase emitted by the model are intentionally distinct strings.

LLM-as-Judge Prompt (system block + IE task criterion)

```
Now your role is a grading teacher. Your task is to review and score student
answers based on reference standard answers for a question-answering
benchmark. You need to notice the following key points:
- First, extract the final answer from the student's solution, then analyze
  and judge whether the answer is correct.
- Scoring should only refer to the final answer obtained by the student;
  there is no need to examine whether the intermediate problem-solving
  steps are correct.
- If the response contains both hesitation and a clear answer, judge the
  answer itself.
- If the response contains both a refusal and a guessed answer, judge the
  final committed answer.
- If the response gives multiple inconsistent answers, assign 0 points.
- If the student's response shows circular reasoning - repeatedly revisiting
  the same evidence, flip-flopping between answers (e.g., "yes... wait
  no... actually yes... let me reconsider... no"), with no clear final
  commitment - assign 0 points. A correct answer mentioned during reasoning
  does NOT count unless it is the clearly stated final conclusion.
- Only the student's LAST clearly stated answer counts. If the student
  initially says "yes" but later concludes "no" (or vice versa), score
  based on the LAST position only. Intermediate answers during reasoning
  are not final answers.
- If the student's output is entirely a reasoning trace (e.g., "let me
  check... scanning evidence... re-reading session 5...") with no clear
  concluding answer statement, assign 0 points - even if the correct answer
  appears somewhere within the reasoning.
- If the student's output begins with "[...truncated earlier reasoning...]",
  this means only the final portion of a longer response is shown. Focus on
  extracting the answer from this final portion.

Below are examples of problematic student outputs that apply to ALL task
types. Study these before grading.

[Universal Example A - Circular Reasoning -> 0 points]
<Question>: Is the cat in Mark and Jenny's 5-year anniversary post named
Mittens?
<Standard Answer>: Yes
<Student Answer>: let me check session 5 mark has cat named mittens session
7 mark and jenny anniversary post there is no link wait let me re-read
session 5 says mittens session 7 says anniversary actually looking at this
again wait is it possible let me reconsider mark has mittens in session 5
the post is about couple therefore answer is likely no wait let me double
check session 5 mark has cat mittens actually maybe yes but text doesnt say
[Scoring Rationale]: The output flip-flops between "yes" and "no" multiple
times with no clear final commitment. The last position is ambiguous
("maybe yes but text doesnt say"). This is circular reasoning.
In summary, the student's answer deserves 0 points.
[JSON]: {"answer_score": 0}

[Universal Example B - Redundant But Committed -> Score Normally]
<Question>: Is the trailing plant in the cafe safe for my kitten?
<Standard Answer>: No
<Student Answer>: the plant looks like pothos pothos is toxic to cats
therefore answer is no let me double check is pothos safe for cats no is
answer no yes wait let me verify image 3 is pothos image 9 is pothos pothos
is toxic therefore answer must be no
[Scoring Rationale]: Despite excessive self-verification, the student
consistently commits to "no" throughout and never wavers to a different
answer. The final answer is clearly "No", which matches the standard
answer.
In summary, the student's answer deserves 1 point.
[JSON]: {"answer_score": 1}

[Universal Example C - Reasoning Trace With No Answer -> 0 points]
<Question>: How many sessions mentioned the coffee shop?
<Standard Answer>: 3
<Student Answer>: scanning session 1 yes coffee mentioned session 2 no
session 3 yes coffee again session 4 no session 5 maybe let me re-read
session 5 it mentions cafe is that same as coffee shop need to check
session 6 no session 7 unclear let me look at session 5 again
[Scoring Rationale]: The student walks through sessions but never states a
final count. The output is entirely a reasoning trace with no concluding
answer.
In summary, the student's answer deserves 0 points.
[JSON]: {"answer_score": 0}

Now proceed with grading. Remember: the universal examples above apply
```

```
regardless of task type.

- When analyzing and judging whether the answer is correct, you need to
  write down the scoring rationale, organize it into clear statements that
  follow the logical flow. The summary of the scoring rationale should be
  placed at the end, using the following format: "In summary, the student's
  answer deserves x points" (where x represents the student's specific
  score).
- Keep the whole process concise, within 150 words.
- Provide the score based on your analysis and display it in a code block
  in "JSON" format.
- An item is covered if it is strictly mentioned or unambiguously implied
  by a semantic equivalence. This includes numerical equivalence (e.g.,
  10% and 0.1), synonyms (e.g., UK and United Kingdom), plural/singular
  forms (e.g., "apple" and "apples"), and equivalent date formats (e.g.,
  2024-01-15 and January 15, 2024). However, do not accept loosely related
  concepts.
- Ignore minor formatting differences, capitalization, punctuation, and
  equivalent wording when meaning is unchanged.

Your output format is:
[Scoring Rationale]:
[Score]: x points
[JSON]:
{"answer_score": <integer_value>}

Below is the grading rubric:
[Scores]:
The scoring scale consists of 2 levels in total, from highest to lowest:
1 point, 0 points (the minimum is 0 points).
[Tier Details]:
1 point: Assign 1 point if the student's final answer matches the standard
answer under the task-specific criteria below.
0 points: Assign 0 points if the student's final answer does not match the
standard answer, the student refuses to answer, claims insufficient
information (except for AR tasks), or does not clearly answer.

[Task-Specific Criteria]
[IE - Information Extraction]
This is an information extraction question.
- Assign 1 point if the student's response contains the core information
  from the standard answer. Minor wording differences are acceptable, but
  the essential information must be present and correct.
- Assign 0 points if the core information is missing, contradicted, too
  vague, refused, or incorrect.

[Remaining seven task-specific criteria (MSR Yes/No, MSR Counting,
MSR Arithmetic, TR Duration Comparison, TR Order Ranking, TR Date
Extraction, KU Knowledge Update, AR Answer Refusal) follow the same
format and are bundled with two worked examples per type in the
codebase release.]
```

F.2 HAYSTACK GENERATION

The haystack pipeline turns a persona profile, the conversation summary up to that point, and a list of recent life events into the next message of the conversation. The box below shows the canonical prompt used at this step. The same template generates both user and assistant turns by swapping the persona slot; image sharing is encouraged in the body of the message itself, so that the resulting haystack sessions remain authentically multimodal rather than text-only sessions with images appended as side annotations. Persona descriptions, event tables, and the previous-session digest are filled into the slots marked with `%s`, in the order listed at the bottom of the prompt.

Haystack: Persona-Driven Dialogue Turn

```
Use a given PERSONALITY to write the next message in this ongoing,
friendly chat.

PERSONALITY: %s

STYLE:
- No hard word limit - write fluidly but keep it conversational, not like
  a letter.
- Focus on real emotions, relationships, regrets, or inspirations tied to
  life events.
```

```
- Reference real people and times ("this morning", "last week", "when I
  turned ten").
- Ask thoughtful follow-up questions when feels natural.
- IMPORTANT: Share photos when discussing experiences, memories, or current
  activities.
  Good examples:
  - "The view was incredible! [shares a photo from the overlook]"
  - "This is what I've been dealing with all week. [shares a photo of the
     messy project]"
  - "Look what I found in the attic! [shares an old family photo]"
- Make the photo sharing feel conversational and relevant to what you're
  saying.
- Avoid talking about outdoor activities or sports.

EVENT MARKING RULES:
If you mention an event, place **EVENT:ID** at the end of the sentence,
after punctuation.
Multiple events: **EVENT:EA1** **EVENT:EA2**
Never in mid-sentence or as a heading.

%s last chatted with %s on %s. Today is %s. You are %s.

Conversation summary:
%s

Recent life events (with IDs):
%s

Background known to both:
%s

%s
Write a heartfelt, realistic continuation to %s.
Discuss how the relevant EVENTS have influenced you - show joy, sadness,
or frustration honestly.
```

F.3 Question Generation

This subsection collects one canonical generator prompt per major question type. Each prompt receives topic context, image metadata, and event-table information, and returns a strict JSON object containing a question, an answer, and supporting facts. Generated outputs then pass through a downstream rule-based pre-filter and the text-only judge described in subsection F.5 before being admitted to the candidate pool that goes to human review. Variants of these prompts (one per subtype within a major type) share the same skeleton with only minor changes to the rule list and the JSON schema; only the canonical entry is reproduced here.

Information Extraction. The two-hop alignment prompt below is the canonical IE generator. The first hop is a textual cue that identifies which visual element to inspect, and the second hop is a visual extraction from that element. The same template handles OCR, attribute, counting, and spatial variants by varying the `visual_type` field; an analogous template (omitted) handles the PrevInfo subtype that grounds the visual hop in screenshots from earlier sessions.

Question Generation: IE (Two-Hop Alignment)

```
[SYSTEM]
You create two-hop visual questions testing alignment (knowing WHERE to
look).

Structure:
- Hop1: A first-person fact directs attention to a specific element/region
  in the image
- Hop2: The answer comes ONLY from visually inspecting that element

Rules:
- Without the fact, it's ambiguous which element to examine
- Vary the visual skill: read text, check color/attribute, count items,
  spatial relations
- Never reference photos ("in the photo", "in the image I shared")

<example>
{
  "question": "Since what year does it state the brand has been trusted?",
```

```
  "answer": "1967",
  "target_element": "Circular seal in top right corner",
  "alignment_cue": "checking the seal on the package",
  "hop1_reasoning": "Fact directs attention to the circular seal in the
                     top right corner.",
  "hop2_reasoning": "The seal reads 'Trusted since 1967'.",
  "visual_type": "ocr",
  "fact": {"fact_id": "F1",
           "text": "I'm checking the circular seal in the top right corner
                    to see how long the brand's been around."},
  "rationale": "1) Fact points to circular seal. 2) Seal displays 1967.",
}
</example>

[USER]
<image_description>{image_description}</image_description>
<visual_elements>{visual_elements}</visual_elements>
<context>{context}</context>

<task>
Generate a two-hop alignment question:
1. Write a first-person fact that identifies which element to examine
   (no <image> token in the fact)
2. Ask a question whose answer requires visually inspecting that element
3. The answer must be impossible to determine without examining the image

Phrasing: ask naturally - never say "in the photo".
Vary structure - don't always start with "What".

Return JSON with the same schema as the example above.
</task>
```

Temporal Reasoning. The order-ranking prompt asks the model to produce eight first-person facts with one timestamp each, of which one is a needle whose timestamp is encoded by a clock image (Mode B) rather than by text. Modes B-date, C, and D follow the same skeleton with the timestamp source replaced by a date image, an entity image with implicit text dates, and an entity image with explicit text dates respectively.

Question Generation: TR (Order Ranking, Mode B)

```
Return ONLY valid JSON (no markdown/comments). Do NOT output an 'answer'
field.

TASK (ORDER_RANKING / Mode B clock): Generate exactly 8 facts (F1-F8),
each with ONE unique timestamp in time_points[0]. ONE needle contains
'<image>' once with source='image_clock', granularity='minute'. Needle
bound to {bound_event_id}, clock_label={clock_label} (fixed).

TIME: Needle value='YYYY/MM/DD HH:MM'; others='YYYY/MM/DD' or
'YYYY/MM/DD HH:MM'. question_date strictly later than all timestamps.

STYLE: First-person, conversational, distinct events. NO formatted
dates/times in text (use "late March", "that summer").

OUTPUT JSON:
{
  "question_date":"YYYY/MM/DD",
  "rationale":"Brief chronological reasoning.",
  "facts":[
    {"fact_id":"F1","text":"...","is_needle":true/false,
     "time_points":[
       {"role":"occurred_at|started_at|ended_at|arrived_at",
        "value":"...",
        "source":"text|image_clock",
        "granularity":"date|minute"}],
     "event_id":"...","event_type":"point"}
  ],
  "images":[
    {"image_id":"IMG1","file_path":"","bound_fact_id":"F?",
     "grounding_type":"temporal","clock_label":"..."}
  ]
}

INPUTS
BACKGROUND: {paragraph}
EVENTS: {events_table}
```

```
CLOCK: bound_event_id={bound_event_id}, clock_label={clock_label}
```

Knowledge Update. The chain-generation prompt produces a four-fact preference-evolution chain. Each fact is at the category level; the specific entity at each step is later substituted by an image, so that text alone can identify the time step but not the value at that step.

Question Generation: KU (Atomic Evolution Chain)

```
You generate a 4-step atomic fact chain showing how MY preferences evolve
over time.

Evolution theme: {evolution_theme}
Use these categories IN ORDER (one per fact, no repeats): {categories}

TASK
Write exactly 4 short, natural-sounding facts (1-2 sentences each) that
describe realistic preference changes across time.

HARD RULES
1) CATEGORY-LEVEL ONLY: use only the provided category names (no specific
   items/brands/subtypes).
2) FIRST-PERSON: each fact must be written from my perspective using
   "I / my / me".
3) PREFERENCE CHANGE: include one simple preference-change marker per fact:
   - Fact 1 marker: used to / before / at first
   - Fact 2 marker: then / later / after that
   - Fact 3 marker: lately / nowadays / recently
   - Fact 4 marker: now / currently / these days
4) DISTINCT CATEGORIES: Fact i must use category i from {categories}.
5) COMPLETE SENTENCES: each fact must be a grammatical sentence (not
   keywords), with a clear subject + verb.
6) LENGTH: each fact must be <= 150 characters total.

OUTPUT (STRICT JSON ONLY; no extra text)
{
  "facts": [
    {"text": "...", "category": "<category_1>",
     "temporal_position": 1, "temporal_marker": "<marker_1>"},
    {"text": "...", "category": "<category_2>",
     "temporal_position": 2, "temporal_marker": "<marker_2>"},
    {"text": "...", "category": "<category_3>",
     "temporal_position": 3, "temporal_marker": "<marker_3>"},
    {"text": "...", "category": "<category_4>",
     "temporal_position": 4, "temporal_marker": "<marker_4>"}
  ]
}

Generate the chain for theme "{evolution_theme}" using categories
(in order): {categories}.
```

Multi-Session Reasoning. The visual identity-match prompt is the canonical MSR generator (sub-pattern D1). It produces a scenario where one named entity and one vague reference are scattered across temporally distinct contexts, and an image attached to the vague reference is the only signal that resolves identity. Five anti-leakage rules (temporal isolation, no ownership disambiguation, no descriptive leakage, no text-based identity resolution, no cultural narrowing) are enforced inside the prompt and re-checked by the text-only judge of subsection F.5.

Question Generation: MSR (Entity Resolution / Visual Identity Match)

```
TASK: Generate a multi-session entity resolution question where the model
must determine whether an ambiguously referenced entity is the SAME AS a
specifically named entity.

Text facts reference an entity using VAGUE terms (e.g., "the dog I
adopted", "that plant from the market"). One fact has an image that
reveals the entity's visual identity. The model must match the image to a
named entity from another fact to answer.

THEME: {theme} ({theme_description})
EXAMPLE ITEMS: {example_entities}
NUM_FACTS: {num_facts}

RULES:
```

```
- At least 1 TEXT NEEDLE names a specific entity (e.g., "I got a golden
  retriever puppy named Max")
- At least 1 TEXT NEEDLE references an entity VAGUELY (e.g., "I adopted a
  dog from the shelter last week")
- EXACTLY 1 IMAGE NEEDLE with <image> token: shows the vaguely referenced
  entity visually. Text uses only vague terms (e.g., "Here's the dog I
  brought home <image>"). The image reveals whether it matches the named
  entity.
- Question: "Is the [vague reference] the same as [named entity]?"
  Answer: "Yes" or "No"
- {num_facts} facts as first-person chat messages ("I", "my"),
  2-4 sentences each
- CRITICAL: Every fact MUST be a first-person declarative statement.
  NEVER phrase facts as questions.
- EVERY fact must be vital to resolving the entity identity. Only use
  fact_type "text_needle" or "image_needle".
- Cross-modality: text alone cannot determine if the vague reference
  matches the named entity (the name is never stated for the vague
  reference). The image is the only way to confirm.
- Each fact should reference the entity with "entity_referenced" field
  tracking which entity it refers to.

ANTI-LEAKAGE RULES (violations cause rejection):
1. TEMPORAL ISOLATION: The named entity and the vague reference MUST
   appear in unrelated temporal contexts. NEVER place both in the same
   event, occasion, trip, or timeframe.
2. NO OWNERSHIP DISAMBIGUATION: Don't assign entities to different named
   owners if ownership alone reveals identity. Multiple entities should
   plausibly belong to the same person.
3. NO DESCRIPTIVE LEAKAGE: The vague reference text must NOT include
   adjectives that confirm or contradict the named entity's known
   characteristics.
4. NO TEXT-BASED IDENTITY RESOLUTION: No text fact may state what the
   vague reference IS. Phrases like "matches", "is the same as", "turns
   out to be", "which is actually" must NEVER appear in any fact.
5. NO CULTURAL NARROWING: The context must allow multiple plausible
   entity types.

NEGATIVE EXAMPLES (DO NOT generate scenarios like these):
- BAD (temporal co-occurrence): F1: "I named my new puppy Max, a golden
  retriever." F2: "I also got a dog from the shelter last week."
  Both acquired recently -> reader infers same event.
- BAD (ownership leakage): F1: "My roommate's cat is named Whiskers."
  F2: "The cat in my bedroom..." Different owners disambiguate.
- BAD (descriptive contradiction): F1: "I got a huge Great Dane named
  Duke." F2: "Here's the tiny puppy I found <image>" - "tiny" rules out
  Great Dane.
- BAD (explicit resolution): F1: "The plant I bought turned out to be a
  monstera." Text resolves identity directly.

OUTPUT (JSON only, no markdown):
{
  "theme": "{theme}",
  "named_entity": "<the specifically named entity>",
  "vague_reference": "<the vague reference used>",
  "same_entity": true/false,
  "question": "Is [vague_reference] the same [category] as [named_entity]?",
  "answer": "Yes" or "No",
  "explanation": "Step-by-step: F1 names [entity]. F2 refers vaguely to
                  [reference]. F3's image shows [what], which is/isn't
                  [named_entity]. Therefore Yes/No.",
  "facts": [
    {"fact_id":"F1","text":"...","fact_type":"text_needle",
     "has_image":false,"entity_referenced":"..."},
    ...,
    {"fact_id":"F3","text":"... <image> ...",
     "fact_type":"image_needle","has_image":true,
     "entity_referenced":"...","image_description":"...",
     "image_search_query":"...","image_search_object":"...",
     "image_provides":"..."}
  ]
}

Generate a creative, realistic scenario now:
```

Answer Refusal. The few-shot prompt below produces a question that is on-topic for the given context but whose answer is not present in the context. The fixed gold answer is the literal string `NOT_MENTIONED`, which the judge rubric in subsection F.1 treats as a successful abstention.

Question Generation: AR (Abstention Few-Shot)

```
You are generating Abstention QA for a multimodal benchmark.

Goal: craft a plausible, on-topic question that appears answerable from
the given info (paragraph + image caption if provided) but is actually
unanswerable. The gold answer must be the literal string "NOT_MENTIONED".

Rules
1) Use ONLY the provided context:
   * CONTEXT: the paragraph
   * VISUAL: the image or its caption (if provided)
2) The question must be topically related to entities/events/attributes
   present in the context.
   - Good: asks for a missing detail about a mentioned item.
   - Bad : asks about something totally unrelated.
3) The correct answer must be truly unknowable from the given info.
4) Output one compact JSON object with EXACTLY these keys:
   {
     "question": "one interrogative sentence",
     "answer": "NOT_MENTIONED",
     "evidence": "short extractive snippet proving the info is absent",
     "explanation": "one-sentence rationale why we must abstain"
   }
5) Do NOT add extra keys or commentary outside the JSON.

---------- Example ----------
INPUT
CONTEXT:
User: I upgraded my old 10-gallon tank, which has my betta fish, Bubbles.
User: I added decorations to the 20-gallon tank for more hiding places.

OUTPUT
{
  "question": "How many fish are there in my 30-gallon tank?",
  "answer": "NOT_MENTIONED",
  "evidence": "The user never mentions owning a 30-gallon tank.",
  "explanation": "No sentence discusses fish in a 30-gallon tank, so the
                  answer cannot be inferred."
}
-----------------------------
```

F.4 Evidence Session Construction

The evidence pipeline wraps each generated needle fact into a multi-turn session that is structurally indistinguishable from a haystack session. The box below contains the prompt that drives the assistant side of these sessions. Length is targeted at 250 to 350 words per turn, with a knowledge-oriented follow-up rather than a personal-social one, so that the evidence session looks informative without advertising the role as the resolution channel for any specific needle question. The user side is generated by a paired template that injects the needle fact into the message body using a directive that is itself constrained to first-person conversational style, after which a six-stage validator chain (rule-based length, photo-directive, n-gram leakage, semantic leakage, ambiguity preservation, and end-marker checks) decides whether the turn is accepted or regenerated.

Evidence Session: Assistant Turn Template

```
You are {assistant_name}, a helpful AI assistant having a casual
conversation about {topic}.

Current conversation:
{chat_history}

{user_name}'s last message: "{last_msg}"

{image_context}

Generate a helpful, conversational response (TARGET: 250-350 words - be
concise, not lecture-like):

STRUCTURE YOUR RESPONSE:
1. Acknowledge what they shared (1-2 sentences)
   - If they shared a photo, briefly acknowledge it (1 sentence max),
     then move on
2. Share your perspective or relevant information (main body, 150-250
```

```
   words)
   - Provide helpful context, explanations, or suggestions
   - Keep it focused - cover 1-2 key points, not exhaustive lists
   - If applicable, offer practical tips or recommendations
3. End with a follow-up thought or question (1 sentence). Ask
   knowledge-oriented questions ("Would you like tips on X?" or "What
   part of this are you most stuck on?"), NOT personal/social questions
   ("Do you prefer X or Y?" or "Have you ever tried X?").

IMPORTANT: Write like a helpful chat assistant, NOT a textbook. Keep
responses focused and avoid padding with extra context the user didn't
ask for.

STYLE GUIDELINES:
- Be warm, conversational, and genuinely helpful
- Use natural paragraph breaks for readability
- Keep photo acknowledgment BRIEF (don't describe or analyze the image
  in detail)
- Provide substantive value - don't just agree, add to the conversation
- Do NOT repeat back personal details verbatim
- Do NOT use emojis

Output only your response, nothing else.
```

F.5 Anti-Shortcut Quality Filter

After question generation, every candidate is passed through a text-only judge that simulates a model with no access to images. The box below shows the prompt used for this judge. A candidate is rejected when the judge returns `answerable_without_image = true`, which screens for question-only solvability under this judge rather than guaranteeing failure for every text-only model. A second visual judge (omitted from this appendix, but structurally identical apart from the role swap) checks the converse, that an image together with the question is sufficient to determine the answer; the two filters together enforce the cross-modal grounding contract of the construction pipeline.

Filter: Text-Only Leakage Judge

```
[SYSTEM]
You are a judge evaluating whether a question can be answered from text
alone, WITHOUT seeing any images.

Your task is to determine if the answer to a question can be inferred
from the conversation text.

Rules:
1. Carefully read the conversation text
2. Try to answer the question using ONLY the text (no images)
3. If you can confidently answer the question from text alone ->
   answerable_without_image = true
4. If the answer requires visual inspection of an image ->
   answerable_without_image = false

Be strict: even if you can make an educated guess, if the text doesn't
explicitly contain the answer, mark it as NOT answerable from text.

[USER]
<conversation_text>{conversation_text}</conversation_text>
<question>{question}</question>
<reference_answer>{answer}</reference_answer>

<task>
Can this question be answered from the conversation text alone, without
seeing any images?

Return JSON:
{
  "answerable_without_image": true|false,
  "text_based_answer": "your answer attempt from text only, or
                        'CANNOT_DETERMINE'",
  "confidence": "high|medium|low",
  "explanation": "why the answer is/isn't determinable from text",
  "leakage_evidence": ["list of text snippets that reveal the answer,
                       if any"]
}
</task>
```

G Supplementary Experiments and Analysis

G.1 Extended Results Tables

Table 13: Comprehensive per-type accuracy (%) for all 27 LVLMs at 32K / 64K / 128K contexts (LLM-as-Judge, $n = 789$). IE: Information Extraction, MSR: Multi-Session Reasoning, TR: Temporal Reasoning, KU: Knowledge Update, AR: Answer Refusal. Cells marked — indicate the model's context window does not support the input length. (T) denotes Thinking mode and (I) denotes Instruct mode. This table is the full roster corresponding to the representative subset in Figure 2; agents are reported separately in Table 14.

	32K						64K						128K					
Model	IE	MSR	TR	KU	AR	Ov.	IE	MSR	TR	KU	AR	Ov.	IE	MSR	TR	KU	AR	Ov.
Claude Sonnet 4.5	28.05	13.29	40.21	29.31	97.78	36.50	26.42	9.09	35.05	21.55	94.44	32.45	19.51	2.80	32.99	16.38	93.33	27.76
Gemini-3.1-Pro	63.60	32.17	40.93	49.14	97.75	54.10	58.94	31.47	42.27	48.62	95.56	52.52	55.79	29.37	41.24	46.03	96.67	50.65
GPT-5.4	63.68	28.18	39.18	47.41	97.78	52.72	63.01	27.27	42.27	43.10	96.67	52.34	60.16	21.68	39.69	43.10	94.44	49.56
Kimi-K2.5	51.20	39.90	53.09	50.86	97.78	54.88	52.44	35.66	52.06	49.14	97.78	53.99	51.63	28.67	48.25	48.59	93.33	50.95
Qwen3.5-122B-A10B	74.39	30.07	51.55	49.14	88.89	58.68	67.07	27.27	51.55	48.59	84.44	55.30	43.09	22.38	41.75	46.03	83.33	44.03
Qwen3.5-27B	70.33	29.37	39.18	42.24	86.67	52.98	62.20	20.98	35.57	43.97	67.78	46.13	54.07	13.99	34.02	38.79	63.33	40.68
Qwen3.5-9B	67.48	16.78	38.66	43.10	91.11	50.32	60.57	13.99	33.51	37.93	80.00	44.36	45.12	5.59	26.80	25.86	62.22	32.57
Qwen3.5-4B	56.91	11.89	38.66	34.48	86.67	44.36	49.19	9.79	30.41	28.45	73.33	37.14	29.67	3.50	24.74	19.83	55.56	25.22
Qwen3.5-2B	64.23	17.48	32.47	33.62	72.22	44.36	2.03	0.00	2.06	0.86	15.56	3.04	1.22	0.00	1.55	0.00	1.11	0.89
Nemotron-Nano-12B	63.82	16.78	44.85	39.66	45.56	44.99	56.10	20.98	42.78	34.48	35.56	40.94	—	—	—	—	—	—
Cosmos-Reason2-8B	55.69	23.78	48.45	31.90	68.89	46.13	46.34	20.28	45.88	24.14	54.44	39.16	36.99	20.28	43.81	12.93	35.56	31.94
Phi4-Multimodal	31.71	20.28	43.81	19.83	43.33	32.19	28.05	20.98	39.18	24.14	21.11	28.14	22.76	22.38	30.93	14.66	15.56	22.69
GLM-4.6V	59.35	20.28	53.09	43.97	93.33	52.34	61.79	23.78	53.09	34.48	64.44	49.05	52.85	23.08	43.81	28.45	30.00	39.04
GLM-4.5V	61.79	15.38	39.69	24.14	51.11	41.19	45.93	15.38	33.51	14.66	32.22	31.18	—	—	—	—	—	—
Gemma3-27B	34.96	23.78	41.24	30.17	68.89	37.64	26.42	22.38	44.33	27.59	61.11	34.22	19.11	18.18	37.11	13.79	48.89	25.98
Gemma3-12B	36.18	22.38	43.81	29.31	53.33	36.50	30.89	22.38	44.33	20.69	38.89	32.07	23.58	18.18	28.35	16.38	28.89	23.32
Gemma3-4B	25.20	21.68	41.75	23.28	14.44	27.12	17.89	18.88	28.87	14.66	8.89	19.26	14.63	12.59	24.23	10.34	5.56	14.96
Qwen3-VL-235B (T)	51.63	25.17	49.48	36.21	88.89	48.29	38.62	18.18	42.78	35.34	75.56	39.67	29.27	15.38	43.30	30.17	67.78	34.73
Qwen3-VL-235B (I)	60.98	18.88	55.67	40.52	97.78	53.23	56.10	20.98	54.12	36.21	95.56	50.82	51.22	18.18	54.64	30.17	92.22	47.66
Qwen3-VL-30B (T)	41.46	11.89	35.57	30.17	71.11	36.38	39.02	11.89	37.11	29.31	65.56	35.23	33.74	10.49	36.60	20.69	60.00	31.31
Qwen3-VL-30B (I)	65.45	18.88	60.82	37.93	93.33	55.01	56.91	20.28	55.15	39.66	78.89	49.81	56.50	16.08	52.58	24.14	63.33	44.23
Qwen3-VL-8B (T)	51.22	17.48	52.58	32.76	80.00	46.01	42.28	18.88	50.52	30.17	54.44	39.67	33.33	8.39	39.18	24.14	45.56	30.29
Qwen3-VL-8B (I)	52.85	17.48	58.25	33.62	90.00	49.18	44.72	19.58	47.42	29.31	82.22	42.84	32.52	20.28	44.85	15.52	66.67	34.73
Qwen3-VL-4B (T)	46.75	20.98	43.30	33.62	82.22	43.35	28.05	11.19	26.80	18.10	70.00	28.01	17.07	9.09	13.92	10.34	52.22	17.87
Qwen3-VL-4B (I)	52.03	17.48	42.78	25.86	92.22	44.23	42.28	19.58	35.57	19.83	75.56	37.01	28.46	16.08	35.05	13.79	50.00	28.14
Qwen3-VL-2B (T)	26.83	9.09	19.07	11.21	87.78	26.36	19.11	6.99	11.86	16.38	54.44	18.76	16.67	12.59	5.67	8.62	14.44	11.79
Qwen3-VL-2B (I)	32.52	16.78	50.52	20.69	72.22	36.88	32.52	21.68	40.72	16.38	65.56	33.97	25.20	18.88	32.99	8.62	47.78	26.11

G.2 Evaluation Coverage and Matched Comparisons

Table 14 reports Mem0, MemOS, MemAgent-7B, M3-Agent, and M3C on all 789 benchmark questions at each context length. The question IDs are identical across these five systems and all four lengths; the type counts are 246 IE, 143 MSR, 194 TR, 116 KU, and 90 AR. Memory-T1 is reported only on the canonical 195-question subset, with no full-benchmark results included in this paper; M2A also retains its subset results. Because their sample sizes and judge differ from the full-benchmark panel, direct numerical differences between the two groups are not paired comparisons.

Canonical subset. The canonical subset contains 195 questions, approximately one quarter of the full benchmark, selected by stratified sampling (seed=42). Memory-T1 and M2A use the same subset at all four lengths. Table 15 shows its per-type composition. The proportions differ from the full benchmark by less than 0.2 percentage points for every type; this describes type balance and does not by itself establish that system rankings are unchanged.

Matched comparison for Memory-T1 and M2A. Table 16 compares Memory-T1 and M2A with six representative direct LVLMs on the same 195 question IDs under the same Qwen3-VL-235B judge. The LVLMs cover 32K/64K/128K, and the two agents cover 32K/64K/128K/256K. This panel retains the reported per-type and overall accuracies. For the six direct LVLMs at 32K, the reported correlation between subset and full-benchmark rankings is Spearman $\rho = 1.00$ ($p < 0.01$, $n = 6$); it is specific to this model panel and length.

Table 14: Per-type accuracy (%) for seven memory agents at 32K / 64K / 128K / 256K. Mem0, MemOS, MemAgent-7B, M3-Agent, and M3C are evaluated on all 789 questions at each length (GPT-5.4-mini judge). M2A and Memory-T1 are reported on the 195-question canonical subset (Qwen3-VL-235B judge). Experimental settings are specified in Appendix B. IE: Information Extraction; MSR: Multi-Session Reasoning; TR: Temporal Reasoning; KU: Knowledge Update; AR: Answer Refusal; Ov.: Overall.

	32K						64K					
Model	IE	MSR	TR	KU	AR	Ov.	IE	MSR	TR	KU	AR	Ov.
Full benchmark ($n = 789$)												
Mem0	8.54	17.48	29.38	22.41	94.44	27.12	10.16	16.08	27.84	15.52	92.22	25.73
MemOS	19.92	19.58	43.30	25.86	81.11	33.46	19.11	20.28	41.24	23.28	78.89	32.19
MemAgent-7B	18.70	20.98	46.91	28.45	4.44	25.86	16.26	20.28	47.42	21.55	7.78	24.46
M3-Agent	23.58	16.78	34.02	11.21	27.78	23.57	21.14	18.18	28.87	5.17	16.67	19.65
M3C	17.07	14.69	32.99	12.07	26.67	20.91	17.07	16.08	29.38	6.90	25.56	19.39
Canonical subset ($n = 195$)												
Memory-T1	18.03	25.71	62.50	13.79	9.09	28.72	19.67	22.86	56.25	17.24	18.18	28.72
M2A	34.41	8.57	2.08	0.00	22.73	15.38	18.03	14.29	25.00	0.00	13.64	15.90

	128K						256K					
Model	IE	MSR	TR	KU	AR	Ov.	IE	MSR	TR	KU	AR	Ov.
Full benchmark ($n = 789$)												
Mem0	8.94	17.48	27.84	15.52	90.00	25.35	8.54	16.78	27.84	15.52	82.22	24.21
MemOS	18.70	19.58	41.24	25.86	84.44	32.95	18.29	18.88	40.72	23.28	84.44	32.19
MemAgent-7B	13.82	17.48	44.85	17.24	6.67	21.80	15.45	17.48	46.91	19.83	1.11	22.56
M3-Agent	22.36	16.08	31.96	4.31	25.56	21.29	20.73	12.59	28.87	3.45	21.11	18.76
M3C	14.23	17.48	26.29	5.17	15.56	16.60	16.67	17.48	25.26	6.03	16.67	17.36
Canonical subset ($n = 195$)												
Memory-T1	21.31	22.86	56.25	27.59	13.64	30.26	18.03	25.71	58.33	24.14	9.09	29.23
M2A	21.31	11.43	27.08	0.00	13.64	16.92	21.31	11.43	25.00	0.00	13.64	16.41

Table 15: Per-type composition of the 195-question canonical subset compared against the full 789-question benchmark. Proportions are preserved to within 0.2 percentage points per type.

Type	Full 789 (%)	195-subset	195 share (%)
Information Extraction (IE)	31.18	61	31.28
Multi-Session Reasoning (MSR)	18.12	35	17.95
Temporal Reasoning (TR)	24.59	48	24.62
Knowledge Update (KU)	14.70	29	14.87
Answer Refusal (AR)	11.41	22	11.28
Total	100.00	195	100.00

Bootstrap confidence intervals for Memory-T1 and M2A. The reported confidence intervals use question-level bootstrap resampling on the canonical 195 questions (1000 iterations, percentile method). At 32K, the 95% intervals (mean $\pm$ half-width, $n = 195$) are Memory-T1 28.72 ± 6.15 and M2A 15.38 ± 5.38; at 128K, they are Memory-T1 30.26 ± 6.41 and M2A 16.92 ± 5.13. These intervals describe the two subset evaluations.

Scope of agent claims. The full-benchmark results in Table 14 describe the five evaluated pipelines with the backbones, input interfaces, and retrieval settings specified in Appendix B. They are not upper bounds on memory-augmented architectures. The matched comparison in Table 16 concerns Memory-T1 and M2A on the canonical subset. Neither panel alone isolates the causal effect of memory architecture from backbone quality or access to visual evidence.

Table 16: Memory-T1 and M2A versus six representative direct LVLMs on the canonical 195-question subset, using the Qwen3-VL-235B judge. Overall and per-type accuracy (%) are reported at 32K/64K/128K for LVLMs and 32K/64K/128K/256K for the two agents. The reported values are retained. Per-type denominators are 61 IE, 35 MSR, 48 TR, 29 KU, and 22 AR.

System / Ctx	Overall (n=195)	IE (61)	MSR (35)	TR (48)	KU (29)	AR (22)
Direct LVLMs on the 195-subset						
Gemini-3.1-Pro / 32K	55.38	60.66	45.71	39.58	48.28	100.00
Gemini-3.1-Pro / 64K	58.46	59.02	48.57	41.67	65.52	100.00
Gemini-3.1-Pro / 128K	58.46	68.85	42.86	43.75	48.28	100.00
GPT-5.4 / 32K	51.28	63.93	17.14	35.42	55.17	100.00
GPT-5.4 / 64K	56.92	65.57	37.14	47.92	44.83	100.00
GPT-5.4 / 128K	51.79	67.21	28.57	37.50	41.38	90.91
Qwen3.5-122B / 32K	63.59	75.41	45.71	58.33	44.83	95.45
Qwen3.5-122B / 64K	59.49	67.21	37.14	52.08	58.62	90.91
Qwen3.5-122B / 128K	49.23	45.90	31.43	43.75	58.62	86.36
Qwen3-VL-30B (I) / 32K	58.46	65.57	25.71	62.50	48.28	95.45
Qwen3-VL-30B (I) / 64K	52.31	57.38	25.71	56.25	44.83	81.82
Qwen3-VL-30B (I) / 128K	50.77	57.38	25.71	66.67	27.59	68.18
Qwen3-VL-8B (I) / 32K	50.77	52.46	22.86	58.33	44.83	81.82
Qwen3-VL-8B (I) / 64K	47.69	52.46	25.71	50.00	41.38	72.73
Qwen3-VL-8B (I) / 128K	34.36	27.87	22.86	54.17	13.79	54.55
Qwen3-VL-2B (I) / 32K	38.46	27.87	20.00	62.50	24.14	63.64
Qwen3-VL-2B (I) / 64K	35.38	29.51	37.14	47.92	10.34	54.55
Qwen3-VL-2B (I) / 128K	27.69	18.03	28.57	41.67	10.34	45.45
Memory-augmented agents on the 195-subset						
M2A / 32K	15.38	34.41	8.57	2.08	0.00	22.73
M2A / 64K	15.90	18.03	14.29	25.00	0.00	13.64
M2A / 128K	16.92	21.31	11.43	27.08	0.00	13.64
M2A / 256K	16.41	21.31	11.43	25.00	0.00	13.64
Memory-T1 / 32K	28.72	18.03	25.71	62.50	13.79	9.09
Memory-T1 / 64K	28.72	19.67	22.86	56.25	17.24	18.18
Memory-T1 / 128K	30.26	21.31	22.86	56.25	27.59	13.64
Memory-T1 / 256K	29.23	18.03	25.71	58.33	24.14	9.09

Table 17: Coverage and Per-Answer Accuracy on the retained 32K diagnostic runs ($n = 789$, 699 answerable; Gemini: $n = 787$, 698 answerable). Coverage = % of answerable questions attempted; PA = accuracy on attempted answers. Judge is the diagnostic-run accuracy, not the selected main-table score. Keyword-detected refusals include one judge-correct answerable response per row.

Model	**M**	**Judge**	**Coverage**	**PA**	**Refused**
Claude Sonnet 4.5	I	36.50	49.50	57.51	353
Gemini-3.1-Pro	I	52.10	66.76	69.10	232
GPT-5.4	I	52.72	63.95	73.15	252
Kimi-K2.5	I	54.88	72.39	67.98	193
Qwen3.5-122B	I	58.68	87.70	62.32	86
Cosmos-Reason2-8B	I	46.13	99.43	43.31	4
GLM-4.6V	I	52.34	95.85	48.96	29
Qwen3-VL-235B	I	53.23	79.26	59.75	145
Qwen3-VL-30B	I	55.01	97.00	51.47	21
Qwen3-VL-8B	I	49.18	92.13	47.52	55

G.3 Coverage and Per-Answer Accuracy

Coverage–accuracy trade-off. For these diagnostic runs, overall judge accuracy equals $(C_{\text{attempted}}+C_{\text{refused}}+C_{\text{AR}})/n$, where C counts judge-correct responses in each group; $C_{\text{refused}} = 1$ in every row. Refusal detection and judge correctness are separate labels. Decomposing performance into coverage (fraction of answerable questions attempted) and per-answer accuracy (PA) reveals contrasting response patterns. GPT-5.4 reaches the highest PA (73.15%) but refuses 36.05% of answerable questions; Qwen3.5-122B attempts 87.70% but achieves only 62.32% PA. Claude Sonnet 4.5 scores 36.50% Overall and 57.51% PA, with a 50.50% refusal rate on answerable questions. The

full decomposition appears in Table 17. At 128K, most models answer *more* questions than at 32K (coverage rises) despite lower accuracy, indicating more attempted answers without higher accuracy; coverage alone does not measure confidence or identify the cause of that change.

Claude Sonnet 4.5: frequent refusal on answerable questions. Per-type decomposition shows how refusal and attempted-answer accuracy contribute to this result. Claude's coverage ranges from 41.26% (MSR) to 64.95% (TR); the corresponding reported ranges are 45–83% for GPT-5.4 and 52–80% for Gemini-3.1-Pro. On IE and KU, Claude's overall accuracies are 28.05% and 29.31%, while its per-answer accuracies reach 63.55% and 62.96%; refusing approximately 55% of answerable questions accounts for much of this difference. TR has its highest coverage (64.95%), but this does not measure the model's confidence or establish that timestamps cause greater willingness to answer. MSR combines low coverage with low per-answer accuracy (41.26% / 32.20%), so frequent refusal alone does not explain its low score. These patterns characterize the evaluated system under the retained protocol. They do not isolate calibration, retrieval, or reasoning as the cause, and they do not resolve possible benchmark or generator-family effects.

G.4 Wrong-Answer Error Analysis

This appendix supplements the wrong-answer error analysis in §4.3 with (a) the motivation for the seven-label taxonomy, (b) the per-label detection procedure used to produce Figure 16 and Figure 17 (Appendix G.4), and (c) the modality re-grouping (Table 19) that turns the seven labels into the five categories shown in Figure 4b of the main text.

Folding into two meta-categories. The four near-miss labels (grounding failure, computation slip, closed-set selection, stale retrieval) identify outputs that retain some relationship to the reference answer or task structure. The three total-miss labels (unsupported answer, answerability failure, non-answer pathology) describe unsupported or non-extractable responses. Labels are assigned from questions, references, predictions, and image-dependency annotations, rather than internal retrieval traces. Their role in §4.3 is diagnostic: the shift toward total misses characterizes the loss of answer grounding as context grows. We interpret this pattern together with the modality breakdown and the oracle-retrieval diagnostic in §G.7 to assess evidence-access difficulties. Thus, the retrieval interpretation draws on these complementary analyses; it is not inferred from a near-miss or total-miss label alone.

Two labels are not enough. The 32K$\rightarrow$128K aggregate shift is $\pm 15.17\%$ by partition identity, so reporting only the meta-category totals hides which labels absorb the shift. Unsupported answer alone carries $+10.23\%$ of the shift, while answerability failure and non-answer pathology each move by less than 3%. The single dominant label inside total-miss cannot be identified without the seven-way split.

Grouping by question type is not enough either. Partitioning wrong answers by question type (five categories) obscures the per-type failure signature: MSR Arithmetic errors are computation slips, TR order-ranking errors are closed-set swaps, and KU errors include stale retrievals. These signatures surface only when the error label is separated from the question type; collapsing either dimension merges failure modes that have different responses to context length.

Why exactly seven. Each of the seven labels either binds to a structural constraint of a question-type subset (numeric answer $\Rightarrow$ computation slip; closed answer set $\Rightarrow$ closed-set selection; match to a KU item's `old_answer` $\Rightarrow$ stale retrieval; AR item $\Rightarrow$ answerability failure; non-terminating generation $\Rightarrow$ non-answer pathology), or is a catch-all for free-form answers (grounding failure when the prediction is evidence-anchored but wrong; unsupported answer when it is not). Every attempted-but-incorrect answer receives exactly one label, and no label admits a strictly finer partition without introducing sub-labels that collapse to zero support on subsets of the benchmark. Table 18 summarises the seven labels with their detection method and category assignment.

Detection, Phase 1 (rule-based). Five labels are assigned by deterministic rules that depend only on the question type and the prediction string. *Stale retrieval* (KU only): the prediction fuzzy-matches the item's `old_answer` field with `SequenceMatcher` ratio ≥ 0.7 on normalized (lower-cased,

alphanumeric) strings. Correctly answering a requested historical state is not a stale-retrieval error. *Computation slip* (MSR Counting and Arithmetic): the prediction and the reference both parse as numbers, and the two differ. *Closed-set selection* (TR Duration A/B, MSR Yes/No, TR order-ranking): the prediction lies in the closed answer set for the item but is not the reference element. *Answerability failure* (AR items): the prediction is a substantive answer rather than a refusal, detected by the absence of any refusal keyword from a fixed list of 14 phrases such as "insufficient information" and "cannot be determined". *Non-answer pathology* (any item): the prediction is empty, or is verbose reasoning with no "answer:", "therefore", or "final answer" anchor. Phase 1 labels take priority over Phase 2 when both would apply; Phase 1 covers roughly 40% of attempted-but-incorrect items at each context.

Detection, Phase 2 (LLM-judged). The remaining items have free-form answers with no Phase 1 rule matching. A GPT-4o-mini call reads the question, the reference answer, and the prediction, and assigns one of the two free-form labels. *Grounding failure*: the prediction is semantically related to the reference but incomplete, over-specific, mislocalized, or at the wrong granularity. *Unsupported answer*: the prediction is a specific entity or value with no meaningful semantic overlap with the reference. LLM classifications are cached in `phase2_cache_{ctx}.jsonl` and reused across runs, so the results are reproducible without new API calls.

Coverage and caveats. Phase 1 and the Phase 2 cache jointly cover roughly 90% of wrong answers per context; the residual are invalid answers that failed to produce a confirmed final answer. This inflates the unsupported-answer share by at most $\sim 10\%$ of its own value; rerunning with the full LLM pass is left as an extension because the direction and magnitude of the near-miss $\rightarrow$ total-miss shift are stable under that default assignment.

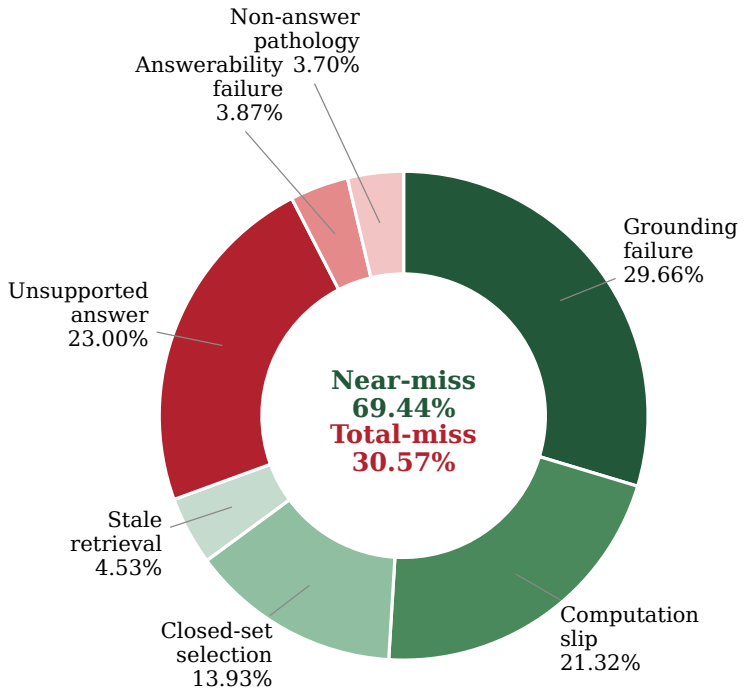


Figure 16: Distribution of wrong-answer types at 32K context ($n = 5{,}592$ attempted-but-incorrect LVLM answers, out of $27 \times 789 = 21{,}303$ evaluations). Near-miss labels account for 69.44% of wrong answers and total-miss labels for 30.57%, using the output-based definitions in Table 18.

Five-category modality view of wrong answers (Figure 4b). The wrong-answer pie (Figure 16) reports the seven-label distribution at the attempted-but-incorrect level; the per-type bars in Figure 4b (main text) re-group every wrong answer along a complementary modality axis by joining its seven-label tag with the per-question image-dependency annotation (image-essential / image-supportive

Table 18: Seven-label wrong-answer taxonomy. Labels partition attempted-but-incorrect outputs into near-miss and total-miss categories using the observable answer patterns below.

Label	Category	Detection	Definition
Grounding failure	Near-miss	LLM	Related answer, wrong detail.
Computation slip	Near-miss	Rule	Numeric output differs from reference.
Closed-set selection	Near-miss	Rule	Right set, wrong element.
Stale retrieval	Near-miss	Rule	Incorrect answer matching the annotated `old_answer`.
Unsupported answer	Total-miss	LLM	No meaningful reference overlap.
Answerability failure	Total-miss	Rule	Answered an unanswerable item.
Non-answer pathology	Total-miss	Rule	Never produced a final answer.

/ text-sufficient, defined in §3.4). The result is five disjoint modality categories that exhaust every attempted-but-incorrect answer; Table 19 gives the exact mapping.

- **Visual.** Grounding failure or unsupported answer on an image-essential question—the visual evidence was the source of the missing information.
- **Textual.** Either (i) grounding failure or unsupported answer on a text-sufficient question, or (ii) any stale-retrieval error (intrinsically a textual-update miss on KU)—text was the source of the missing information.
- **Mixed.** Grounding failure or unsupported answer on an image-supportive question, where image and text both contribute to the answer and either could be at fault.
- **Reasoning.** Computation slip, closed-set selection, or answerability failure—a label-specific reasoning-shaped error (off-by-one arithmetic, wrong A/B choice, or substantive answer to an unanswerable item).
- **Output.** Non-answer pathology (empty or non-extractable response).

Table 19: Mapping from the seven-label wrong-answer taxonomy ($\times$ per-question image dependency) to the five modality categories of Figure 4b. A single column entry indicates that the dependency does not affect the assignment.

Seven-label	image-essential	image-supportive	text-sufficient
Grounding failure	Visual	Mixed	Textual
Unsupported answer	Visual	Mixed	Textual
Stale retrieval		Textual	
Computation slip		Reasoning	
Closed-set selection		Reasoning	
Answerability failure		Reasoning	
Non-answer pathology		Output	

The seven-label and five-modality views connect answer patterns to the evidence demands of each task. The seven-label view records the *form* of an incorrect answer; the modality view combines this label with the question's *evidence channel*. This explains the per-type asymmetries in Figure 4b: IE and KU concentrate errors in the Visual category, whereas the numeric and closed-set structure of MSR produces a large Reasoning share under the Phase-1 rules. A wrong total can reflect missing or misread operands as well as incorrect arithmetic. The oracle-retrieval diagnostic in §G.7 complements this taxonomy by testing performance when the required facts and images are supplied. Read together, the analyses support the main text's emphasis on evidence-access fidelity: a reasoning-shaped output error can originate upstream of aggregation.

Case studies. To illustrate each error category concretely, we present one representative failure case per label below. Each card shows the question, the reference answer, and the model's incorrect prediction at the 32K context length.

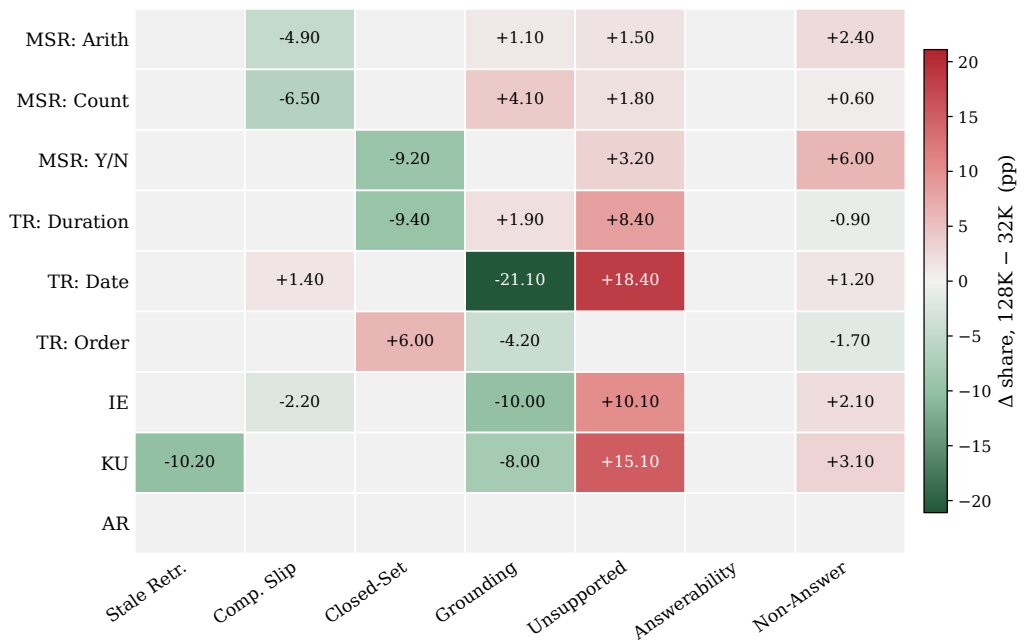


Figure 17: Wrong-answer error-type shift from 32K to 128K by question type ($n = 789$). Unsupported answer (+10.23%) replaces grounding failure (−5.38%) and computation slip (−4.82%).

Case Study: Grounding Failure

The prediction refers to the same object or attribute as the reference but gives an incorrect detail, location, or level of specificity.

Model: claude-sonnet-4.5 **Task:** Information Extraction **Context:** 32K

Question. Where on the sign is the photo of the mountains located?

Reference Answer. In the bottom right corner
Model Prediction. In the upper right corner of the kiosk sign.

Case Study: Computation Slip

The prediction and reference are numeric but differ. Such a mismatch may reflect incomplete evidence, misread operands, or an arithmetic error.

Model: claude-sonnet-4.5 **Task:** MSR — Arithmetic **Context:** 32K

Question. How much total have I spent on coffee makers?

Reference Answer. $260.00
Model Prediction. $150.00

Case Study: Closed-Set Selection Error

The model recognised the closed answer set (A/B, Yes/No, or a fixed ordering) but selected the wrong element—typically swapping two options.

Model: glm4.6v **Task:** TR — Duration Comparison **Context:** 32K

Question. Which of the following two durations is longer? Duration 1: the time spent working at JTB Corporation; Duration 2: the time spent working at Rakuten Group.

Reference Answer. B
Model Prediction. a

Case Study: Stale Retrieval

The model retrieved a factually correct but outdated *answer—returning a previous state instead of the most recent update in the knowledge-update chain.*

Model: cosmos-reason2-8b **Task:** Knowledge Update **Context:** 32K

Question. What's my favorite dip now?

Reference Answer. spinach artichoke dip
Model Prediction. Hummus
Previous (Outdated) Answer. hummus

Case Study: Unsupported Answer

The model's prediction has no meaningful semantic overlap with the reference; the response appears fabricated with no correct evidence anchor.

Model: cosmos-reason2-8b **Task:** MSR — Arithmetic **Context:** 32K

Question. How much total have I spent on blenders?

Reference Answer. $433.99
Model Prediction. You have spent a total of $40 on blenders.

Case Study: Answerability Failure

The question is deliberately unanswerable (evidence image removed), yet the model produced a confident substantive answer instead of refusing.

Model: glm4.5v **Task:** Answer Refusal **Context:** 32K

Question. What color are the words "rice is life" written in?

Reference Answer. `NOT_MENTIONED`
Model Prediction. The words "rice is life" are written in black.

Case Study: Non-Answer Pathology

The model either produced an empty response or generated verbose reasoning without an extractable final answer.

Model: glm4.5v **Task:** MSR — Entity Resolution **Context:** 32K

Question. Based on the tweets and my observations, is my vehicle, Bluebird, subject to the recall mentioned by @Woolworths_SA?

Reference Answer. No
Model Prediction. To determine if your vehicle, Bluebird, is subject to the recall mentioned by @Woolworths_SA, follow these steps: 1. Identify the Recall Details: Check the Woolworths tweet for the specific vehicle mo...

G.5 Context-Length Analysis

This appendix collects the context-length-centric analyses that complement the memory-ability and multimodal perspectives in the main text.

Size helps long-context retention within a family. Within the Qwen3-VL Instruct family, the ratio of 128K to 32K Overall accuracy rises from 0.71 at 8B to 0.90 at 235B; within the Qwen3.5 dense series, 27B retains 0.77 versus 0.65 at 9B and 0.57 at 4B. These examples indicate better retention at larger scales within the evaluated families. They do not isolate parameter count from model architecture or training recipe.

Architectural family shifts both the ceiling and the degradation profile. Qwen3.5-122B-A10B loses 14.65 points in Overall accuracy from 32K to 128K, compared with 5.57 points for Qwen3-VL-235B Instruct. The type-wise profile also differs: at 128K, Qwen3-VL-235B Instruct reaches 54.64% TR and 30.17% KU, while Qwen3.5-122B-A10B reaches 41.75% TR and 46.03% KU. The systems differ in size, architecture, and training, so these observations characterize model-specific trade-offs rather than isolate an architectural cause.

No model family wins on every type. The per-type leaders differ even among high-scoring systems. Kimi-K2.5 leads MSR at 32K (39.90%) but reaches 28.67% at 128K. At 128K, GPT-5.4 leads IE (60.16%), Kimi-K2.5 leads KU (48.59%), and Gemini-3.1-Pro leads MSR (29.37%). These profiles support evaluating different abilities separately, without attributing specialization to training recipe alone.

Hard questions stay hard across context lengths. The 280 questions that are hard at 32K (solved by $<$20% of models) average 9.07% accuracy at 32K and 8.52% at 128K, so expanding context does not unlock them. These items remain difficult in this paired diagnostic, but the means do not separate task demands from context effects.

Nominal improvements and diagnostic scope. The aggregate decline with context length does not imply that every model and type decreases monotonically. Across the five types and two transitions (32K$\rightarrow$64K and 64K$\rightarrow$128K), the selected result tables contain 26 positive LVLM transitions, 13 among the five full-benchmark agents, and 9 among Memory-T1 and M2A. These are descriptive counts, not significance tests; no new paired tests are applied to the full-agent results.

A separate paired LVLM diagnostic used per-question outcomes to examine nominal improvements in its own evaluation panel. It reported 38 positive transitions and one uncorrected $p < 0.05$ result: Qwen3-VL-2B Thinking MSR at 64K$\rightarrow$128K ($+5.59$ points, $p = 0.022$, 9 flip-to-correct versus 1 flip-to-wrong). Eight of the nine recovered questions hit the 2,048-token generation cap at 64K but produced normal outputs at 128K. This diagnostic illustrates how output instability can produce local reversals; its transition counts and tests are not recomputed from, or applied to, the selected aggregate result tables.

Bidirectional churn underlies apparent reversals. Per-question tracking across 21 LVLMs with judge data at all three context lengths reveals that context transitions induce *bidirectional* churn: at each step, models simultaneously lose 60–150 questions they previously answered correctly and gain 30–80 new correct answers. The observed accuracy decay is the net of these two opposing flows, not a uniform loss of signal. Of questions that flip at 32K$\rightarrow$64K, 38.7% flip back at 64K$\rightarrow$128K, showing that many single-step reversals do not persist; identifying stochastic variation would require repeated runs under the same condition. Within this separate paired diagnostic, Gemini-3.1-Pro flips approximately 71 questions in each direction at 64K$\rightarrow$128K, and Kimi-K2.5 approximately 83. Such balanced turnover can yield a small net change despite substantial question-level instability. These rounded diagnostic counts are not used to reconstruct the frozen Overall values in Table 13. The diagnostic therefore cautions against interpreting a small aggregate change as stability on individual questions.

We also tested whether MSR questions are systematically easier at 64K, since 10 individual LVLMs show MSR improvement at that length. The separate question-level analysis did not detect a systematic advantage for MSR at 64K: more questions became harder than easier across its model panel. This observation does not exclude all possible positioning effects.

Overall, the aggregate results show declining accuracy with length, together with local reversals. The separate paired diagnostics illustrate bidirectional churn and output variability; they do not establish a cause for each reversal in the full-benchmark agent results.

AR accuracy degradation. AR accuracy declines with context length across the reported Qwen3-VL sizes, with different drops by scale and decoding mode (Table 20). The 235B Instruct model loses 5.6 points from 32K to 128K, whereas 2B Thinking falls from 87.8% to 14.4% (-73.3 points). The separate output diagnostic associates some failures with truncated reasoning traces; the aggregate table alone does not establish truncation as the cause of every decline.

G.6 Agent Underperformance: Where in the Pipeline Is Information Lost?

This subsection examines *where in the pipeline* information is lost using existing diagnostic outputs on retained subsets, including the canonical 195-question subset (Appendix G.2). We retain the subset diagnostics, including the backbone ablation and retrieval-attribution analysis, under their original evaluation conditions. These supplemental analyses characterize failure patterns within their

Table 20: Answer Refusal (AR) accuracy across input lengths for Qwen3-VL Instruct and Thinking modes. Both modes lose accuracy; the relative decline varies with model size. Differences are in percentage points.

Size	Mode	32K AR	64K AR	128K AR	$\Delta_{32\to128}$
235B	Instruct	97.8	95.6	92.2	−5.6
235B	Thinking	88.9	75.6	67.8	−21.1
30B	Instruct	93.3	78.9	63.3	−30.0
30B	Thinking	71.1	65.6	60.0	−11.1
8B	Instruct	90.0	82.2	66.7	−23.3
8B	Thinking	80.0	54.4	45.6	−34.4
4B	Instruct	92.2	75.6	50.0	−42.2
4B	Thinking	82.2	70.0	52.2	−30.0
2B	Instruct	72.2	65.6	47.8	−24.4
2B	Thinking	87.8	54.4	14.4	−73.3

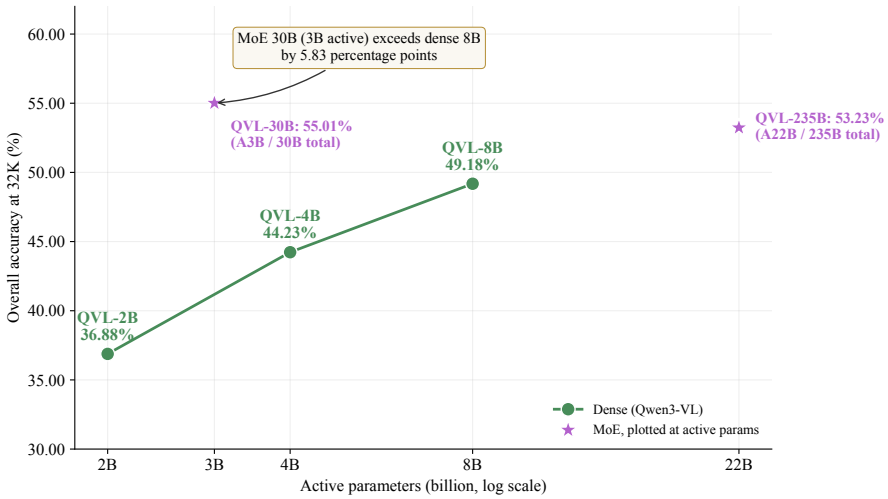


Figure 18: Model-size scaling within the Qwen3-VL Instruct family at 32K. Dense models improve from 2B to 8B; the 30B MoE (3B active) exceeds the 8B dense model by 5.83 percentage points. MoE models are plotted at their active parameter counts. This comparison does not isolate parameter count from architecture or training.

diagnostic panels; they are separate from the full-benchmark results in Table 14, and their absolute accuracies should be interpreted within the corresponding panel.

Session-recall thresholds separate different failure patterns. For the three agents with retrieval logs (Mem0, Memory-T1, and M3C), we partition wrong answers by session-level evidence recall: the fraction of ground-truth evidence sessions surfaced by the retriever. The reported "retrieval failure" category uses recall below 0.5. For M3C, 78.1% of errors fall below this threshold, with mean recall 0.26. Mem0 and Memory-T1 have mean recall of 0.82–0.89, and 87–95% of their errors occur at or above the threshold (Figure 19). The latter category was labeled "comprehension failure" in the diagnostic, but retrieving at least half the sessions does not establish delivery of all required facts or visual details. Errors above the threshold may still involve incomplete retrieval, lossy evidence representation, extraction, or reasoning. The pattern motivates testing session coverage for M3C and evidence sufficiency after retrieval for Mem0 and Memory-T1, rather than assigning each pipeline a single causal failure mode.

Backbone capacity does not determine complete-system accuracy. The matched canonical-subset comparison (Table 16) illustrates the difference between backbone capability and complete-system accuracy. At 32K, M2A scores 15.38%, while direct Qwen3-VL-8B-Instruct scores 50.77%

on the same 195 items, a 35.39-point gap with the same reported answer backbone. Memory-T1 reaches 28.72% using its smaller Qwen2.5-3B backbone and a different memory pipeline. These comparisons show that a capable answer backbone does not guarantee high accuracy after memory construction and retrieval. They do not isolate architecture from training, visual access, evidence selection, or prompting. In the separate backbone-ablation panel below, Mem0 and MemOS use Qwen3-VL-8B as the diagnostic default; their absolute scores and configurations remain specific to that panel. The appropriate conclusion is to evaluate backbone choice within a fixed pipeline and the complete pipeline on its own terms, without comparing unlike score gaps as causal effect sizes.

Controlled backbone ablation. The cross-agent comparison above is confounded by architecture, training recipe, and backbone quality varying simultaneously. To isolate the backbone factor, we re-evaluate Mem0 and MemOS with alternative backbones while keeping each pipeline unchanged (Table 21). Within the Mem0 FAISS architecture, swapping the diagnostic default Qwen3-VL-8B for gpt-4.1-mini lifts Overall accuracy by 10.65 percentage points to 43.15%, while substituting Qwen2.5-7B lowers it by 4.00 points to 28.50%, for a 14.65-point spread. Within MemOS, replacing the default Qwen3-VL-8B with Qwen2.5-7B counter-intuitively raises Overall accuracy by 2.50 points, showing that the tested newer-generation backbone does not improve this pipeline. These within-pipeline effects establish sensitivity to backbone choice; they cannot be compared with the M2A system gap to determine which component dominates.

Per-type profiles also shift across backbones. In Mem0, the diagnostic default Qwen3-VL-8B reaches 77.27% AR while the Qwen2.5-7B variant reaches 100%; Qwen3-VL-8B instead leads on TR (50.00% versus 32.69%). These results show that backbone choice changes the ability profile within the fixed FAISS pipeline, without isolating intrinsic calibration or retrieval quality. The retained length diagnostic reports a 32K-to-256K Overall range below five percentage points for each tested variant. This is evidence of aggregate stability under those conditions, not a guarantee of architectural invariance or unchanged accuracy for every question type. It is consistent with the main result that modest Overall changes can coexist with pronounced type-specific failures.

Table 21: Backbone ablation within fixed agent architectures at 32K. Each row uses the same pipeline with a different backbone model, evaluated on the 200-question stratified subset (197 for Mem0 with gpt-4.1-mini). Δ: spread between best and worst backbone within each framework.

Framework	Backbone	IE	MSR	TR	KU	AR	Ov.
Mem0	Qwen3-VL-8B (default)	13.11	25.00	50.00	17.24	77.27	32.50
Mem0	gpt-4.1-mini	31.15	33.33	42.86	44.83	90.91	43.15
Mem0	Qwen2.5-7B	3.28	19.44	32.69	31.03	100.00	28.50
	Δ						14.65
MemOS	Qwen3-VL-8B (default)	18.03	22.22	40.38	34.48	81.82	34.00
MemOS	Qwen2.5-7B	29.51	22.22	40.38	20.69	90.91	36.50
	Δ						2.50

Captioner control. We used the caption-input configuration of MemAgent-7B to test a stronger captioner on the canonical 195 questions at 32K. Only the question-blind captioner changes, from BLIP-2 to GLM-4.6V; the memory writer, retriever, answer backbone, prompts, and Qwen3-VL-235B judge are held fixed. Each arm has two runs. Overall accuracy increases from 26.92% to approximately 29.17% ($+2.25$ percentage points), with the largest improvement on IE ($+9.67$ points). We report point estimates without a claim of statistical significance in either direction. The improvement indicates that captioner quality matters for visually grounded questions. The upgraded pipeline nevertheless remains below the reported 50.77% of direct Qwen3-VL-8B-Instruct on the same 195 items (Table 16). This latter comparison uses the reported direct-model score and the agent re-judge; it does not isolate the effects of the caption interface, memory pipeline, or answer backbone. The diagnostic supports improving visual-evidence preservation, while its scope remains one pipeline, one context length, and a subset of the benchmark.

Aggregate stability is limited to the tested diagnostic. Within this retained diagnostic subset, six of seven agents stay within the reported $\pm 7\%$ band across 32K–256K. Correct-answer sets also overlap:

adjacent-length Jaccard overlap exceeds an independent-draw baseline at the observed accuracies by factors of 3.3–6.9 across agents. This comparison indicates persistence beyond that particular random baseline, but neither identical correct-answer sets nor context invariance. The diagnostic reports a smaller gap to the Qwen3-VL-8B tier at 128K than at 32K, while substantial absolute deficits remain. These observations concern the retained panel and do not identify information loss as their sole cause. The full-benchmark results must be read separately: their smaller Overall declines coexist with low IE/KU accuracy and type-specific changes, including the 12.22-point AR decline of Mem0.

A common question subset defeats multiple agents regardless of architecture. On the shared answerable subset at 32K, 52% of the questions are answered incorrectly by all four agents with complete per-question logs (M3-Agent, M3C, M2A, and Memory-T1, with overall accuracy in the 14 to 30% range), despite architecturally disjoint retrieval pipelines that include ColPali (M3-Agent), LoRA session retrieval (M3C), dual-layer SQLite (M2A), and BM25 (Memory-T1). This overlap diagnostic uses the four retained logs with aligned question identifiers; Mem0, MemOS, and MemAgent-7B are outside this diagnostic panel. The four included agents span distinct retrieval paradigms. The type-specific all-wrong rates are 69% for KU and MSR and 21% for TR; these are rates within types, not shares of the all-wrong set. Whether this ceiling persists for stronger agents or instead reflects a shared bottleneck of current sub-10B retrieval-based memory systems is an open question that would require matched-scale implementations we leave to future work.

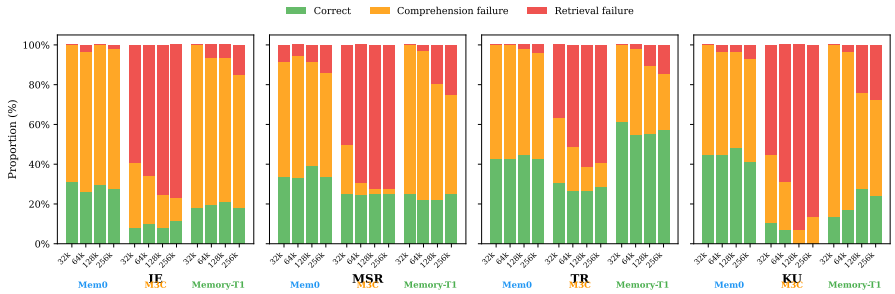


Figure 19: Session-recall diagnostic on the canonical 195-question subset at 32K. Correct answers are green; wrong answers with session recall at least 0.5 are yellow (originally labeled "comprehension failure), and those below 0.5 are red ("retrieval failure). These operational categories do not establish complete evidence delivery or isolate comprehension errors.

G.7 MSR Oracle-Retrieval Diagnostic

MSR has the lowest type-level ceiling among LVLMs at 32K (§4.2), and its cross-session aggregation over three to eight sessions remains difficult. We use an oracle-retrieval diagnostic to examine whether supplying the required evidence directly restores performance. Here the oracle provides question-relevant facts and images: evidence access includes both locating relevant sessions and selecting and extracting the information needed for aggregation.

The cross-modality ablation (Table 3) supplies the question, required textual evidence facts, and associated evidence images, without the haystack. Under this condition, reported MSR accuracy reaches 100.00% for GPT-5.4 and 89.74% for Gemini-3.1-Pro. The input contains extracted facts rather than complete evidence-session transcripts, bypassing long-history search and much of the within-session selection and fact-extraction problem while retaining the need to combine the supplied information. These high scores show that the models can answer most MSR questions once the required evidence is directly available, supporting the main text's emphasis on failures in acquiring and preserving that evidence during long-context answering.

This diagnostic also clarifies the large Reasoning share in the output taxonomy (§G.4). An incorrect count or sum in a long history can result from combining incomplete or misread evidence, even when the model can perform the required aggregation with sufficient inputs. We therefore interpret the improvement as evidence that long-context evidence access is a substantial bottleneck, motivating the main text's focus on cross-modal evidence-retrieval fidelity. The oracle supplies facts and images jointly, so it measures the combined benefit of simplifying evidence access rather than the separate

contribution of each retrieval, reading, or extraction stage. Gemini-3.1-Pro's remaining approximately 10% error also leaves room for reasoning and interpretation improvements.

G.8 Session Indistinguishability Validation

Evidence sessions and haystack sessions are produced by the same generation pipeline. Both use GPT-5.1 as the user and Gemini-3-Pro as the assistant, follow the same prompt templates with identical style constraints (250–350 words per turn, knowledge-oriented follow-ups rather than personal or social ones), and share the same image integration pathway (§F.4). The sole structural difference is that an evidence session embeds one or more needle facts into the user turns; a six-stage validator chain (rule-based length and format checks, n-gram overlap detection, semantic leakage filtering, chain-ambiguity preservation, and end-marker verification) then screens each generated turn, rejecting and regenerating any output that departs from the haystack register or reveals the embedded fact through surface cues. These shared generation constraints aim to reduce stylistic differences. They do not guarantee indistinguishability: inserting evidence facts may leave residual cues even when the same generator and validators are used.

The conversation history assembly stage imposes type-specific ordering constraints beyond the random placement and non-contamination invariants described in the main text. For MSR and KU questions, whose answers depend on the temporal sequence of facts across sessions (a four-step knowledge update chain or a cross-session counting task, for instance), the evidence sessions are inserted into the haystack sequence in their original fact order, preserving the chronological progression that the question probes. For TR questions, each evidence session is assigned a timestamp that strictly precedes the question's reference date, so that temporal reasoning over session dates remains internally consistent. These ordering constraints, together with the non-contamination invariant and the visual-density normalization by text-only filler sessions, reduce obvious placement cues while preserving task validity; they do not prove that all structural signals are absent.

To validate the indistinguishability claim quantitatively, we trained two binary classifiers to distinguish evidence from haystack sessions. From the evidence and haystack sessions in the benchmark, we drew a balanced 698/698 sample and applied five-fold cross-validation with question-grouped splits, ensuring that no question appears in both the training and test sets of any fold. Speaker labels were normalized to a uniform User:/Assistant: format and the `<image>` placeholder token was removed, eliminating trivial classification shortcuts. A TF-IDF logistic regression model (unigrams and bigrams, 10,000 features, sublinear term frequency) achieved $56.71\% \pm 1.20\%$ accuracy with an ROC-AUC of 0.59, and a fine-tuned DeBERTa-v3-base classifier (512-token truncation, three epochs) reached $57.92\% \pm 0.85\%$ with an ROC-AUC of 0.55. Both results sit only marginally above the 50% chance floor for balanced binary classification (Cohen's $h < 0.16$, a small effect). A pooled binomial test against H_0: $p = 0.5$ on the TF-IDF predictions (805 of 1,396 correct) yields a two-sided $p < 10^{-7}$; the classifiers detect a small text-distribution difference. Its utility for query-conditioned retrieval within a long history is not measured by balanced binary-classification accuracy.

A follow-up vocabulary ablation progressively removed the K most discriminative n-grams from the TF-IDF feature set ($K = 20, 50, 100, 200, 500, 1{,}000$; balanced removal from each class direction, ranked on training-fold coefficients only to prevent test-fold leakage). Across the entire sweep, accuracy remained in a narrow band between 54.77% and 56.71%, never exceeding 7 percentage points above chance. Removing up to 1,000 top-ranked n-grams leaves accuracy essentially unchanged, indicating that the weak signal is diffuse rather than concentrated in identifiable stylistic markers. Together, these controls characterize a weak but detectable surface signal under the tested classifiers. They do not establish the absence of exploitable shortcuts, nor directly measure whether such cues help a question-conditioned retriever locate all required evidence.

G.9 Controlled Generator Swap

We examine generator-family overlap by replacing Gemini with GLM-4.6V as the question generator while holding the underlying (entity, image, abstracted paragraph) triples, prompt templates, validator chain, answer models, and judge fixed. The reported experiment generated 169 questions spanning IE, MSR, KU, and TR; Answer Refusal is excluded because these items have no source triple to rewrite. In the reported comparison, GLM-4.6V scores 53.33% on Gemini-generated questions and 54.00% on its own questions, an increase of 0.67 percentage points whose reported confidence

interval includes zero. Gemini-3.1-Pro changes by -2.08 points. GLM-4.6V's arm-wise change was reported as not significantly different from that of any of the four non-generator controls, yielding no detectable generator-specific advantage in this controlled comparison.

This conclusion is limited to the tested generator swap, question set, and answering condition. It does not establish that generator-family effects are absent from the benchmark, and it does not replace the scores or rankings in the main result tables. In particular, failure to detect a GLM-specific advantage on GLM-generated questions does not by itself rule out a Gemini-family advantage on the released benchmark.

G.10 Memory-Oriented Post-Training and Evidence-Sensitive Abstention

Correct abstention requires recognizing when the available memory does not support an answer. MemLens constructs Answer Refusal (AR) questions by removing their supporting evidence, so the appropriate response is to acknowledge insufficient information. Retrieving related content or generating a plausible answer does not satisfy this requirement. A reliable memory system must therefore both use available evidence and recognize its limits. Success on answerable questions alone does not establish that this second behavior will survive optimization for memory-based answering.

Our results show that strong abstention is achievable, but varies substantially across memory systems. At 32K, the released Qwen3-VL-8B-Instruct baseline, evaluated without additional memory-specific post-training, reaches 90.00% AR. Mem0 and MemOS use frozen Qwen3-VL-8B backbones with text-only memory harnesses and obtain 94.44% and 81.11% on the full benchmark. These results demonstrate that external memory can coexist with effective abstention. The memory-oriented post-trained MemAgent-7B and M3-Agent obtain 4.44% and 27.78%, respectively. Memory-T1 obtains 9.09% on the separately reported canonical195 subset. This pattern motivates examining how memory-oriented optimization treats evidence insufficiency, beyond its ability to produce correct answers when the required evidence is present.

The reported training objectives suggest a mechanism through which abstention could be weakened. MemAgent and M3-Agent emphasize final-answer correctness, while Memory-T1 additionally rewards evidence grounding and temporal consistency; their reported objectives contain no dedicated abstention component (Yu et al., 2025; Du et al., 2025b; Long et al., 2025). A correctness reward can still reinforce refusal if unanswerable examples and suitable reference responses are included. However, when such cases are insufficiently represented, optimization may reward successful answering without adequately preserving the decision to withhold an unsupported answer. This yields a specific hypothesis: memory-oriented post-training can weaken evidence-sensitive abstention when its training signals do not sufficiently distinguish answerable from unanswerable inputs.

Prior work supports this mechanism with comparisons across training stages. AbstentionBench evaluates successive Tülu 3 checkpoints and reports degraded abstention after reinforcement learning with verifiable rewards, although earlier supervised and preference-training stages can improve abstention (Kirichenko et al., 2025, Section 4.2). The Hallucination Tax of Reinforcement Finetuning likewise compares models before and after mathematical reinforcement finetuning and finds reduced refusal on unanswerable problems (Song et al., 2025, Section 5.1, Figure 2). These studies establish that training for successful problem solving can impair abstention in controlled settings. They provide external support for the proposed mechanism beyond the cross-system association observed in MemLens.

Interventions on training supervision further support treating abstention as an explicit optimization target. Song et al. (2025, Section 5.2, Table 2) show that mixing unanswerable problems into reinforcement finetuning can recover refusal, with a trade-off in performance on solvable problems. Abstain-R1 trains refusal and subsequent clarification through verifiable rewards; its ablation removing unanswerable supervision substantially reduces correct refusal, even as answerable-question accuracy increases (Zhai et al., 2026, Section 6.4, Table 2). Together, these interventions indicate that the composition of training data and rewards can change the balance between answering and abstaining. Post-training can therefore improve this balance when evidence insufficiency receives

appropriate supervision; degradation is a risk of particular optimization choices, rather than an inevitable consequence of reinforcement learning.

Taken together, our observations and these external interventions support memory-oriented post-training as a plausible contributor to abstention failures. Our comparison does not isolate the training effect: it does not evaluate each model before and after memory-specific training with its inference pipeline and evaluation protocol held fixed. Other mechanisms also matter: M3C reaches only 26.67% AR despite using a stock answering checkpoint in our evaluated configuration. The low scores therefore cannot be attributed exclusively to training. Nevertheless, the observed pattern, the reported objectives, and the controlled evidence from prior work jointly support the risk identified in the main text. Memory systems should evaluate and optimize evidence access, answer correctness, and appropriate abstention together, including checking that gains on answerable questions preserve refusal when supporting evidence is absent.

H Detailed Limitations and Future Work

Synthetic conversations and human-in-the-loop naturalness. The evidence and haystack sessions in MemLens are LLM-generated (GPT-5.1 user, Gemini-3-Pro assistant; §3.2), with ShareGPT and UltraChat (Ding et al., 2023) dialogues used as filler in the surrounding context. Naturalness is enforced through an intensive human-in-the-loop review pipeline rather than automatic metrics alone (§E.1): Round 1 audits each candidate question; Round 2 audits each session against a conversational-naturalness criterion covering colloquial phrasing, turn-taking coherence, and indirect embedding of factual content, returning stilted or overly formal sessions to the generation pipeline until annotators accept them as plausible user–assistant exchanges; Round 3 audits the assembled haystack. This pipeline reduces the candidate pool from $\sim$20k to 789 final questions, following the human-review practice established by LongMemEval (Wu et al., 2025). As a complementary surface-shortcut control, the session-indistinguishability validation (§G.8) trains binary classifiers to separate evidence from haystack on text features and reports only marginally above-chance accuracy (DeBERTa accuracy 57.92%, TF-IDF accuracy 56.71%). These small but detectable signals do not establish the absence of retrieval shortcuts. MemLens therefore stands as a controlled diagnostic stress test for cross-modal evidence retrieval, multi-session aggregation, and length-controlled comparability under uniform construction; characterizing the gap to the distribution of real long-term human–assistant interactions remains an open research question.

Question generator and test-taker overlap. The final question-generation model in our pipeline is Gemini-3-Pro (§3.2, §C.4), while Gemini-3.1-Pro is one of the evaluated models. The 32K Overall leader is Qwen3.5-122B-A10B at 58.68%; at 128K, Gemini-3.1-Pro and Kimi-K2.5 reach 50.65% and 50.95% (Table 13). These rankings do not resolve possible generator-family familiarity. Conditioning generation on fixed entity–image–paragraph triples and applying leakage checks and human review constrain the task, but do not eliminate this concern. The evidence-supplied results in Table 3 also do not isolate generator effects, because they change evidence availability rather than the generator. The controlled same-triple swap in Appendix G.9 reports GLM-4.6V changing from 53.33% to 54.00% ($+0.67$ points) and Gemini-3.1-Pro changing by -2.08 points, with no detectable generator-specific advantage in that comparison. This is evidence about the tested swap, not proof that the released benchmark is free of generator-family effects.

Judge limitations. LLM-as-judge validation in this paper covers 800 of 73,784 judge calls ($\approx$1.08%) for cross-family agreement against GPT-5.4-mini and 484 items for human verification by three annotators with consensus adjudication. Three limitations follow. First, $\kappa = 0.86$ measures judge-vs-consensus-label agreement; we do not separately report inter-annotator κ among the three raters, since disagreements were resolved to a single label before release, so inter-annotator reliability on this benchmark remains an open question. Second, the 29 false positives versus 2 false negatives indicate the judge is systematically lenient on borderline partial-match and verbose-correct outputs; the format-dependent bias correction (Appendix E.2) addresses the short-output subset but not the full leniency channel. Third, per-subtype judge agreement (Appendix E.2) inherits the reporting taxonomy of §C.3 (9 subtypes, all $n \geq 46$); finer per-generator splits inside MSR Entity and TR Temporal Grounding remain statistically noisy. A separate inter-annotator reliability study on the 120 double-annotation-tagged items and a per-rare-type audit are left to future work. The

retained VLM and subset-agent scores use the Qwen judge, whereas the five full-benchmark agents use GPT-5.4-mini. The reported validation does not establish that this judge difference is negligible for the current cross-system comparisons; rankings and small score gaps should be interpreted within their evaluation panels.

Static-length vs. streaming evaluation. MEMLENS queries each frozen multi-session history offline at four context lengths (32K–256K tokens); a complementary online streaming protocol that preserves temporal causality between memory writes and queries (Zheng et al., 2026) would further probe irreversible-update and forgetting dynamics in the multimodal setting, and is left to future work.

Broader impacts. MEMLENS provides a length-controlled diagnostic for multimodal conversational memory, intended to make memory-faithfulness regressions in long, multi-session multimodal assistants visible to model developers before deployment. Two negative impacts are foreseeable. First, leaderboard targeting can encourage over-fitting to the construction pipeline (§3.2); the released anti-shortcut filter (§F.5) and the canonical 195-question subset (§G.2) are designed so that such over-fitting becomes detectable rather than rewarded. Second, the synthetic dialogue construction does not capture the full distribution of real user–assistant interactions, so accuracy on MEMLENS is a necessary but not sufficient indicator of deployed memory quality. As mitigation, MEMLENS is released with versioned frozen tags so that any specific evaluation run remains reproducible; images are sourced from the public web and each retains its original source-site license, as documented in §C.7.

Ethics statement. MEMLENS is released for the evaluation of multimodal long-term conversational memory, and we do not endorse using its dialogues for training or fine-tuning, since exposure of evaluation items would compromise diagnostic value; the versioned frozen tags (§C.7) and the per-image perceptual hashes support detection of leakage in derived models. Two dual-use considerations follow. First, memory-augmented assistants whose deployed memory faithfulness regresses below the levels reported here may amplify mis-recall in user-facing settings; the per-ability decomposition (§4.3) is designed so that such regressions are visible before deployment. Second, the synthetic-dialogue construction (§3.2) does not capture the distribution of real long-term human–assistant interactions, so MEMLENS scores are a necessary but not sufficient indicator of deployed memory quality. Image sourcing, licensing, privacy, and takedown protocol are documented in Appendix C.7.

I DESIGN RATIONALE FOR THE MEMORY ABILITY TAXONOMY

This appendix motivates the five memory abilities evaluated in MEMLENS and argues that they jointly cover the space of capabilities required for long-term multimodal conversational agents.

Recent memory-agent benchmarks converge on three functional dimensions of conversational memory (Bei et al., 2026; Wu et al., 2025): retrieval (can the agent recall stored information?), reasoning (can it synthesize inferences across distributed evidence?), and knowledge update (can it correctly update its internal state as the conversation evolves?). Our taxonomy is designed so that the five abilities collectively span all three dimensions. IE addresses retrieval at single-session granularity; MSR and TR exercise reasoning from complementary angles; and KU and AR probe opposing facets of knowledge update. We justify each ability below.

Information Extraction (IE): single-session retrieval. IE operationalizes the most fundamental function of memory: accurate recall of previously encountered information. Each IE question requires the agent to locate a specific evidence session within a long interaction history and extract the relevant fact. The multimodal extension adds a layer of image understanding: Entity-subtype questions present an abstracted entity visible only in an image, requiring visual recognition before textual retrieval, while PrevInfo-subtype questions ask the agent to recall a visual detail from an image shared in an earlier session. This two-hop structure ensures that IE measures genuine cross-modal memory rather than unimodal text matching.

Multi-Session Reasoning (MSR): cross-session aggregation. MSR extends single-session retrieval to aggregative reasoning over information distributed across three to eight sessions. The tested

operations (counting, arithmetic, and entity resolution) all require the agent to identify multiple relevant sessions, extract the pertinent facts from each, and combine them into a coherent answer. We retain a subset of text-only MSR questions alongside multimodal ones. These text-only items serve as a controlled modality ablation: they isolate cross-session reasoning from the added difficulty of visual retrieval, yielding a cleaner signal of the reasoning capability itself. Including them also preserves ecological validity, since not every piece of cross-session information in a realistic conversation history involves images. Mem-Gallery (Bei et al., 2026) adopts a compatible design in its Multi-entity Reasoning subtask, where the entities involved can be textual or visual.

Temporal Reasoning (TR): temporal awareness. Temporal reasoning has emerged as a distinct research focus in LLM evaluation (Wang et al., 2023; 2022; Wei et al., 2025; Wu et al., 2025). In long-running interaction histories, temporal information takes heterogeneous forms: natural-language date expressions in utterances, session-level timestamps in metadata, and implicit ordering from the session sequence itself. Our TR questions test whether the agent can jointly process these signals together with visual content (e.g., clock faces and calendar images that replace textual temporal expressions) to answer questions about duration comparison, event ordering, and temporal grounding. This ability is especially relevant for personal assistants, where users frequently ask questions that require temporal contextualization of past interactions.

Knowledge Update (KU): current and historical state tracking. A defining requirement for personalized conversational agents is the ability to track how user attributes evolve over time. Users routinely update preferences, correct earlier statements, and revise plans across sessions (Xu et al., 2024). Each KU question presents a chain of four successive updates to a single attribute and asks for either the current state or a specified historical state. This tests whether the agent can distinguish states across updates and retrieve the value appropriate to the question's temporal reference. Mem-Gallery (Bei et al., 2026) identifies the current-state aspect under Knowledge Resolution, framing it as maintaining consistency when "new, contradictory information appears in the dialogue."

Answer Refusal (AR): epistemic calibration. AR tests whether the agent can recognize the limits of its available evidence. Each AR question is constructed by removing all evidence sessions from an otherwise answerable instance, so that no supporting information remains in the conversation history. A correct agent must decline to answer rather than hallucinate a plausible response. This ability addresses a well-documented failure mode of large language models (Zhang et al., 2024a) and is essential for trustworthy deployment: a personal assistant that fabricates answers from absent evidence is more harmful than one that acknowledges uncertainty.

Coverage and empirical distinctiveness. The five abilities cover different demands within the retrieval–reasoning–update space. IE and MSR require access to evidence at single-session and cross-session granularity, respectively. TR probes temporal inference, KU probes revision of stored facts, and AR probes abstention when evidence is absent. The correlations in §4.3 show substantial overlap between IE and KU, alongside weaker or context-dependent associations for other pairs. This pattern supports separate reporting of the abilities without claiming statistical independence or minimal redundancy.

Code is provided in the supplementary material, including pointers to the dataset and per-image metadata.

J Additional Related Work

Personalization and modality-specific shortcut controls. LCMP evaluates whether an MLLM tracks changes in personalized concepts and produces responses that reflect those changes; TAME provides a training-free baseline with separate dynamic and persistent memories (Hong et al., 2026). This personalization setting is adjacent to our evaluation of five conversational-memory abilities under controlled context lengths. Outside conversation, AVUT filters audio-centric video questions that can be answered from the question and options alone, using answer permutations to test textual shortcuts (Yang et al., 2025). Video-MMMU evaluates knowledge acquisition from professional videos through perception, comprehension, and adaptation questions, and measures the knowledge gained after viewing (Hu et al., 2025a). These controls differ from our context-level leakage test, which retains the conversational text while removing the evidence images (§3.4).