M^3Eval: MULTI-MODAL MEMORY EVALUATION THROUGH COGNITIVELY-GROUNDED VIDEO TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

As multi-modal models advance towards long-form video understanding, memory emerges as a critical capability. Despite substantial efforts in developing video datasets and benchmarks, existing works primarily focus on perception and reasoning, without systematically evaluating memory: what models retain, how faithfully information is preserved, and how robust memory remains under interference. To address this gap, we introduce M^3Eval, the first comprehensive evaluation framework and benchmark for probing different memory dimensions in multi-modal models. Grounded in cognitive psychology, our design features carefully constructed tasks that probe key aspects of memory. Leveraging M^3Eval, we conduct extensive experiments across representative multi-modal models, revealing consistent weaknesses and distinctive behaviors. We find that models struggle to maintain disentangled representations when processing parallel video streams, exhibit interference patterns differing substantially from those observed in human memory, ground memory sources more reliably in the spatial domain than the temporal domain, and demonstrate limited symbolic memory. Collectively, our benchmark provides a valuable resource for future research, while our findings highlight memory as a fundamental yet underexplored capability and offer insights for designing more effective memory mechanisms in multi-modal models.

1 INTRODUCTION

Multi-modal models (Bai et al., 2025; Wang et al., 2025b; Qwen Team, 2026; OpenAI, 2026; Google DeepMind, 2026) are rapidly advancing towards long-form video understanding, driven in part by expanding context windows. However, increasing context alone does not guarantee effective memory. A core challenge lies in the memory mechanism itself (Hu et al., 2025a; Jia et al., 2026; Liang et al., 2025; Zhang et al., 2025), the ability to encode, store, retrieve, and synthesize information over long temporal horizons spanning both video and text. Such memory is critical for retaining information across long video streams and multi-turn interactions, and for enabling downstream reasoning that depends on this information (Long et al., 2025; Zuo et al., 2025; Fan et al., 2024; Song et al., 2024; 2025a). Despite growing interest in this capability, there is currently no dedicated evaluation protocol or benchmark for systematically probing memory in multi-modal models. As a result, their memory capabilities remain poorly measured and not well understood.

A large body of multi-modal datasets and benchmarks has been developed for video understanding (Fu et al., 2025; Li et al., 2024). These benchmarks primarily focus on visual perception and reasoning. While some tasks implicitly involve memory, for example, long video understanding (Zhou et al., 2025a; Wang et al., 2025a; Chandrasegaran et al., 2024; Zhou et al., 2025b; Yang et al., 2025a) or video reasoning (Cheng et al., 2025; Yang et al., 2025c), they are not designed to directly evaluate memory. Consequently, they provide only a partial and indirect assessment of memory. In particular, existing benchmarks rarely probe different aspects of memory.

To address this gap, we introduce M^3Eval, a principled evaluation framework and benchmark for probing memory capabilities in multi-modal models. As illustrated in Fig. 1, our design is inspired by the controlled experimental paradigms in cognitive psychology (Kahana & Wagner, 2024; Schwieter & Wen, 2022), where memory is studied through carefully constructed stimuli that probe specific mechanisms. We adapt these principles to video domain by constructing video-based QA tasks that probe memory under controlled yet realistic conditions. Our benchmark characterizes memory

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

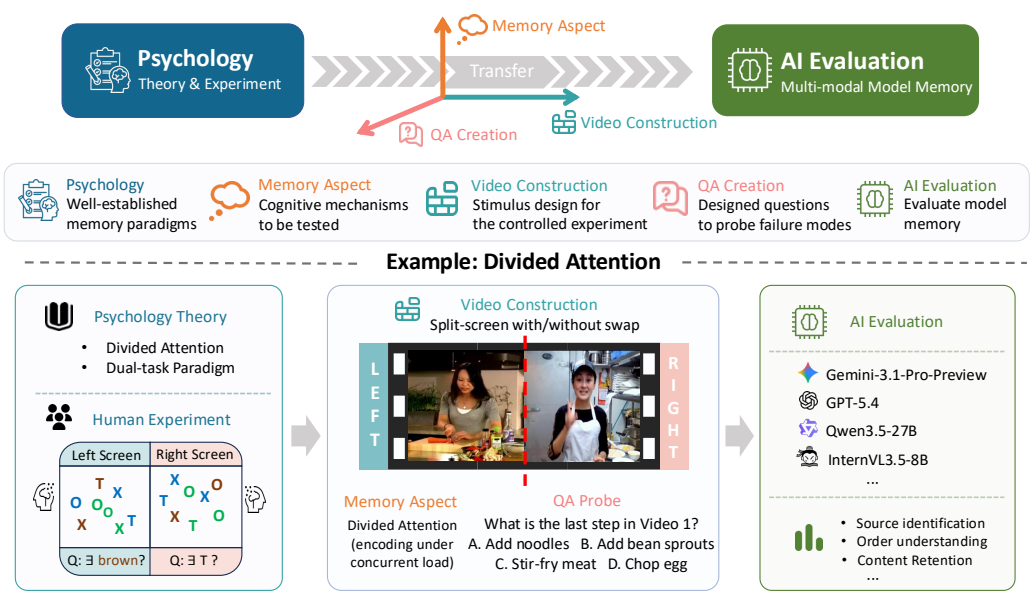

Figure 1: M^3Eval, our principled framework and benchmark for evaluating memory capabilities of multi-modal models. We present an example task of *divided attention*. Grounded in psychological theory, we construct split-screen video scenarios, design memory questions, and analyze multiple models in terms of source identification, order understanding, and content retention.

along four key dimensions: *(1)* the ability to retain information from concurrent inputs (Baddeley, 1998; Craik et al., 1996; Treisman & Gelade, 1980; Treisman & Schmidt, 1982); *(2)* robustness to interference from similar content (McGeoch, 1932; Underwood, 1957; Zaromb et al., 2006; Robinson, 1920); *(3)* the ability to integrate interleaved events into coherent representations (Mandler & Johnson, 1977; Mandler, 1978; Johnson et al., 1993; Schacter et al., 1984); and *(4)* the ability to track abstract attributes across video segments (Kirchner, 1958; Owen et al., 2005). While grounded in cognitive theory, these dimensions also arise in real-world video understanding, such as scene analysis, object tracking, and long context reasoning.

Leveraging M^3Eval, we conduct an extensive evaluation on open-source and proprietary multi-modal models. Our results reveal several notable and, in some cases, unexpected findings. *First*, when processing parallel video streams, the models fail to maintain independent representations for each stream; we hypothesize that such failure stems from attention confusion across concurrent inputs. *Second*, humans exhibit notably stronger retroactive interference than proactive interference, whereas multi-modal models demonstrate comparable interference levels. This contrast indicates a fundamental difference in memory mechanisms between humans and models. Surprisingly, repeating interfering video segments can even improve model understanding of target video segments. *Third*, model memory is less capable than human memory when organizing temporally interleaved information. Further analysis reveals that memory source grounding along temporal dimension is consistently weaker than spatial dimension. *Finally*, the models exhibit far weaker symbolic memory than humans when required to abstract multi-modal information into symbolic attributes and distinguish their relations. We further find that the models struggle to filter out irrelevant information from memory.

Our contributions are summarized as follows.

- We introduce M^3Eval, **the first benchmark** for systematically evaluating different dimensions of memory capabilities of multi-modal models with video tasks.
- **Our key innovation** lies in a cognitively-grounded evaluation design that probes memory through orchestrated video tasks.
- We provide **a systematic evaluation** across diverse models, offering new insights into the limitations of current multi-modal memory and informing the design of future systems.

2 Related Work

Memory Evaluation in LLMs and Agents. Memory evaluation has been recently studied for LLMs and LLM-based agents (Hu et al., 2025a; Jia et al., 2026; Liang et al., 2025). Early benchmarks relied on synthetic needle-in-a-haystack tasks (Song et al., 2025b; Hsieh et al., 2024; Kuratov et al., 2024) or long-range dialogues (Maharana et al., 2024; Jia et al., 2025) to assess retention within a fixed context. (Wei et al., 2025; Zheng et al., 2025) introduced self-evolution settings examining whether models can distill strategies from experience. While the above efforts focus on text, Mem-Gallery (Bei et al., 2026) extends memory evaluation to the multimodal setting with multi-session dialogues grounded in text and images. Inspired by cognitive psychology, recent studies (Gong et al., 2024; Zhang et al., 2024) adopted the N-Back task (Kirchner, 1958) to assess working memory capacity. However, none of these works has explored memory evaluation for video tasks.

Evaluation for Video Understanding. Memory is an essential yet underexplored component for video understanding. Numerous benchmarks evaluate general video understanding (Fu et al., 2025; Li et al., 2024), long-form video tasks (Zhou et al., 2025a; Wang et al., 2025a; Chandrasegaran et al., 2024; Zhou et al., 2025b), streaming evaluation (Niu et al., 2025; Lin et al., 2026; Xiong et al., 2025), and cross-video understanding (Zhu et al., 2025; Li et al., 2026). However, these benchmarks conflate memory with visual perception and reasoning without measuring memory explicitly. Another line adopts synthetic needle-in-a-haystack settings (Zhao et al., 2024; Hu et al., 2025b; Yang et al., 2025b), inserting target segments into distractor footage to test retrieval over extended contexts. Yet these approaches rely on simple probe designs, making it difficult to assess different memory dimensions. A recent effort (Long et al., 2025) probes memory through reasoning tasks, yet does not directly evaluate multiple memory dimensions. Unlike these works, our benchmark leverages existing video datasets and explicitly probes key memory dimensions through cognitively-grounded paradigms.

Memory Investigation in Cognitive Psychology. Cognitive psychology decomposes memory into distinct, measurable processes. Our evaluation framework builds on four such processes: *(1) Divided Attention.* Divided attention during encoding degrades retention and induces illusory conjunctions (Craik et al., 1996; Treisman & Gelade, 1980), due to limited cognitive resources (KAHNEMAN, 1973; Baddeley, 1998). *(2) Memory Interference.* Forgetting arises from competition among similar memory traces, manifesting as proactive or retroactive interference (McGeoch, 1932; Underwood, 1957). *(3) Memory Organization.* Recall relies on implicit story schemata (Mandler & Johnson, 1977). When processing interleaved storylines, individuals default to the underlying event structure (Mandler, 1978). *(4) N-Back and Symbolic Representation.* The N-Back task (Kirchner, 1958; Owen et al., 2005) is widely used to evaluate memory capability and reflects the view that memory operates over abstract representations (Anderson & Bower, 2014; Pylyshyn, 1973).

3 Memory Evaluation

As shown in Fig. 2, our evaluation consists of four paradigms unified under a coherent framework. All evaluations share **a common design principle**: each is grounded in cognitive psychology theory, instantiated as a controlled video task, and equipped with targeted questions and metrics to quantify specific failure modes. Below, we first introduce the design of each evaluation paradigm in detail and then describe the process of evaluation dataset creation.

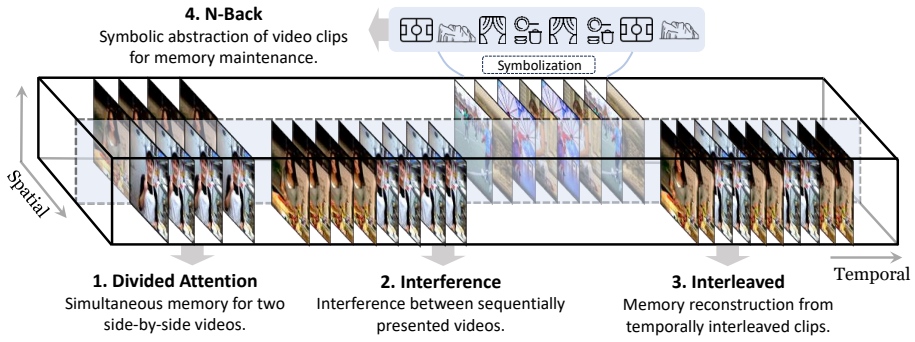

Figure 2: **Overview** of the unified and coherent framework for our four evaluation paradigms.

3.1 Evaluation Design

3.1.1 Divided Attention: Encoding Concurrent Information

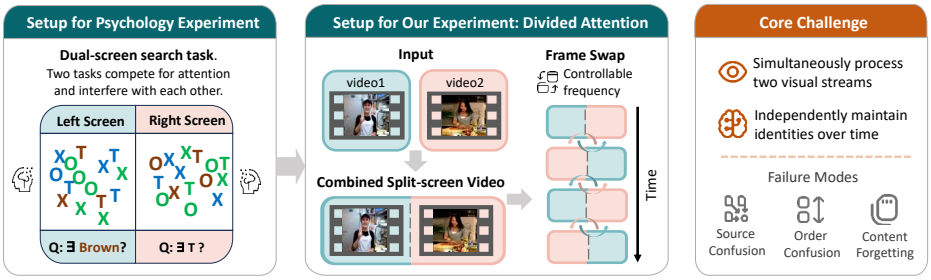

Figure 3: **Divided Attention**. Split-screen presentation with optional frame swaps.

Psychological Theory. The divided attention paradigm originates from research on *limited attentional resources* and *dual-task processing* (KAHNEMAN, 1973; Baddeley, 1998). In classic experiments, participants perform two tasks simultaneously, competing for attentional resources and resulting in reduced encoding quality and impaired memory retention (KAHNEMAN, 1973; Craik et al., 1996).

Instantiation in Video Understanding. Following this paradigm, we adopt a split-screen configuration where two semantically similar videos are displayed synchronously, as shown in Figure 3. We consider two conditions: (1) *No swapping*: V_1 appears on the left and V_2 on the right, evaluating whether the model maintains distinct representations under parallel input. (2) *Swapping*: the positions of V_1 and V_2 are swapped 10 times at uniformly spaced timestamps, examining whether the model can track the correspondence between content identity and spatial location.

Metrics. We construct three types of multiple-choice questions, each targeting a specific failure mode. Each question has one correct option and three distractors of the same error type: (1) *Source Identification*, where content from the distractor video is erroneously attributed to the target, resulting in source confusion; (2) *Order Understanding*, where the temporal or logical sequence of events is inaccurately recalled; and (3) *Content Retention*, where plot points or details from the target video are misremembered or imprecisely recalled.

3.1.2 Memory Interference: Robustness to Distraction

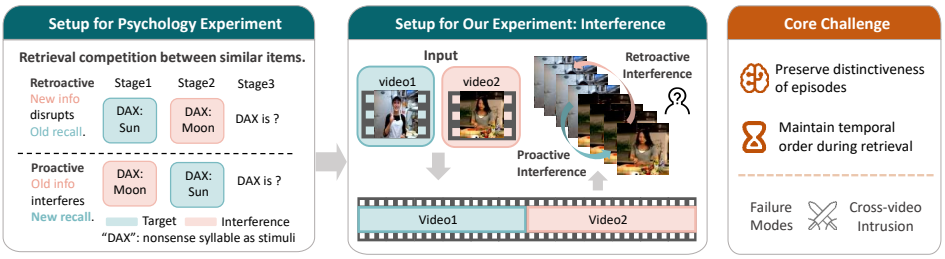

Figure 4: **Memory Interference**. Proactive interference: earlier learning disrupts later memory. Retroactive interference: later learning impairs earlier memory.

Psychological Theory. Memory interference theory explains forgetting as competition among similar traces rather than passive decay (McGeoch, 1932; Underwood, 1957). *Proactive interference* occurs when earlier material disrupts recall of later material, while *retroactive interference* occurs when later material impairs recall of earlier material (Zaromb et al., 2006; Robinson, 1920).

Instantiation in Video Understanding. As shown in Figure 4, we concatenate two semantically similar videos and pose questions about one designated target video. To isolate each interference direction, we evaluate both concatenation orders using identical questions targeting the same video. Specifically, in the order [V1, V2], asking about V1 tests retroactive interference, as the later video V2 may disrupt recall of the earlier target. In the order [V2, V1], asking about the same V1 tests proactive interference, as the earlier video V2 may disrupt encoding of the later target.

Metrics. We design multiple-choice questions with four options: (1) the correct answer for the target video, from which we report Accuracy (Acc); (2) two intrusion options drawn from the competing

video, from which we report Intrusion Rate (IR) following (Zaromb et al., 2006), measuring the proportion of responses that select an option from the competing video; and (3) one unrelated distractor. IR directly quantifies cross-video intrusion.

3.1.3 Interleaved Events: Temporal Organization

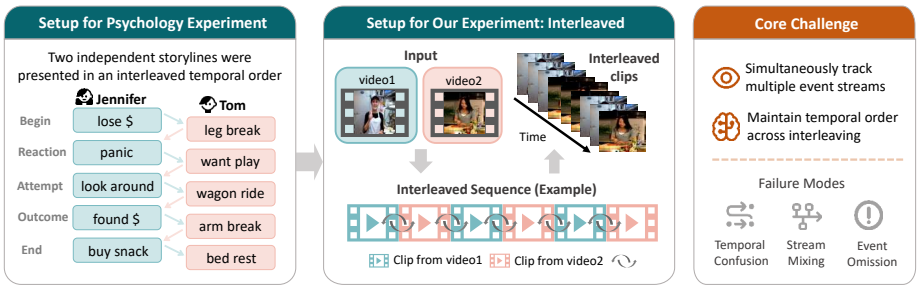

Figure 5: **Interleaved Events**. Interleaved presentation of video clips from two sources.

Psychological Theory. (Mandler, 1978) demonstrated that, when presented with intermixed storylines, individuals spontaneously recover the underlying event structure rather than following surface presentation order. This paradigm has become a classic test for memory organization.

Instantiation in Video Understanding. We divide two source videos with each into 10 temporally ordered segments and interleave them into a single stream in alternating order, e.g., A_1–B_1–A_2–B_2–$\cdots$–A_{10}–B_{10}, as shown in Figure 5. To answer correctly, the model must disentangle segments from the same source and recover the internal temporal order of the target video.

Metrics. We adopt the same three question types as in §3.1.1, and add a fourth: *False Memory Discrimination*, inspired by the DRM paradigm (Roediger & McDermott, 1995; Deese, 1959). Here, a fake question that is relevant to video content is presented, and the model should be aware to choose the option indicating that the query does not belong to either video.

3.1.4 N-Back: Symbol Grounding and Memory Capacity

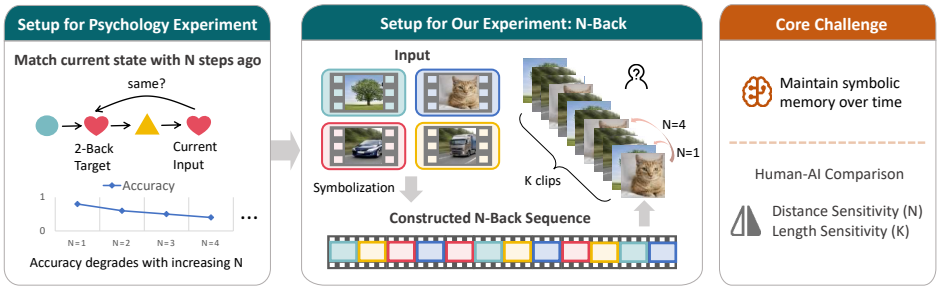

Figure 6: **N-Back**. Abstracting videos into symbols and comparing them.

Psychological Theory. Unlike episodic memory, symbolic memory concerns the ability to abstract events into symbolic representations (Anderson & Bower, 2014; Pylyshyn, 1973). N-Back tasks present sequences of symbolic stimuli (e.g., letters, digits, or simple shapes) and require participants to decide whether the current stimulus matches the one N steps earlier (Kirchner, 1958). This match/mismatch structure naturally requires encoding stimuli as abstract symbols before comparison, making the N-Back format well-suited for probing symbolic grounding and memory capacity.

Instantiation in Video Understanding. We adapt the N-Back paradigm to a multi-video clip sequence setting. As shown in Figure 6, each test sample consists of a sequence of short video clips drawn from different source videos. Two variables control the difficulty: N, the lag distance, where the model determines whether the final clip matches the clip N positions earlier on a designated attribute (e.g., scene or action category); and K, the sequence length, specifying the total number of video clips presented to the model in a single trial.

Metrics. The model is asked to decide whether the final clip matches the clip N positions earlier, producing a Yes/No answer. *Scene* measures whether two clips belong to the same scene or environment category, while *Action* assesses whether they depict the same type of activity.

3.2 Evaluation Dataset Creation

Our video materials are drawn from five publicly available datasets. Questions are automatically generated using Qwen3.5-27B (Qwen Team, 2026) and refined through manual review. In total, our benchmark comprises 2,403 questions over 451 videos spanning approximately 403 hours. We provide benchmark statistics in Appendix A, video and QA construction details in Appendices F and G, and example visualizations in Appendix I.

4 Experiments and Results

We evaluate **two proprietary models** (Gemini-3.1-Pro-Preview (Google DeepMind, 2026) and GPT-5.4 (OpenAI, 2026)), **five open-weight models** (Qwen3-VL-8B-Instruct (Bai et al., 2025), Qwen3.5-{4B, 9B, 27B} (Qwen Team, 2026), and InternVL3.5-8B (Wang et al., 2025b)), and **two agentic methods**: VideoLucy (Zuo et al., 2025), which adopts Qwen3.5-4B as the VLM and DeepSeek-V4-Pro (DeepSeek-AI, 2026) as the LLM, and M3-Agent (Long et al., 2025) with its default configuration. We additionally report human performance as a reference. Further details are provided in Appendix C, D, and H.

4.1 Divided Attention: Encoding Concurrent Information

Table 1: **Divided Attention**. Accuracy (%) on three divided attention metrics under the split-screen setting without swaps and with frequent left/right swaps.

Acc(%)	No swapping			Swapping		
	Source Identification	Order Understanding	Content Retention	Source Identification	Order Understanding	Content Retention
Human	89.58	90.00	92.16	81.25 (-8.33)	85.00 (-5.00)	86.27 (-5.89)
Random	25.00	25.00	25.00	25.00 (0.00)	25.00 (0.00)	25.00 (0.00)
Closed-Source Models						
Gemini-3.1-Pro-Preview	62.50	52.50	49.02	37.50 (-25.00)	52.50 (0.00)	56.86 (+7.84)
GPT-5.4	27.08	35.00	47.06	35.42 (+8.34)	30.00 (-5.00)	49.02 (+1.96)
Open-Source Agents						
VideoLucy	16.67	42.50	37.25	14.58 (-2.09)	25.00 (-17.50)	39.22 (+1.97)
M3-Agent	27.08	30.00	23.53	31.25 (+4.17)	35.00 (+5.00)	23.53 (0.00)
Open-Source Models						
Qwen3.5-4B	18.75	25.00	31.37	14.58 (-4.17)	22.50 (-2.50)	33.33 (+1.96)
Qwen3-VL-8B-Instruct	16.67	25.00	37.25	12.50 (-4.17)	30.00 (+5.00)	35.29 (-1.96)
InternVL3.5-8B	29.17	37.50	33.33	25.00 (-4.17)	40.00 (+2.50)	27.45 (-5.88)
Qwen3.5-9B	35.42	25.00	25.49	18.75 (-16.67)	30.00 (+5.00)	13.73 (-11.76)
Qwen3.5-27B	41.67	25.00	35.29	27.08 (-14.59)	32.50 (+7.50)	35.29 (0.00)

Main results. Table 1 shows that divided attention is challenging for existing models. They all exhibit a substantial gap from human performance, with most near chance across metrics except Gemini-3.1-Pro-Preview, indicating that effective dual-stream understanding remains beyond current memory mechanisms. With frequent swapping, the most prominent drop occurs on *source identification*, while other categories are largely unaffected. This suggests that swapping mainly disrupts source identification rather than order understanding or content retention.

Further experiment. To better understand this failure mode, we examine attention visualizations from representative examples. As shown in Figure 7, in the single screen format, model attention concentrates on the queried region. However, in a split-screen format, the attention maps become notably more diffused and disorganized. Based on this observation, we hypothesize that the poor performance may stem from attention confusion across concurrent visual streams, preventing the model from selectively attending to the relevant stream.

> ***Finding 1****: Existing multi-modal models lack robust memory for parallel tasks, probably due to attention confusion across concurrent visual streams.*

Discussion. In real-world settings, events often unfold simultaneously, requiring systems to process and reason over multi-view or multi-stream inputs, as in autonomous driving (Jiang et al., 2023; Tian et al., 2024; Liu et al., 2025) and household robotics (Intelligence et al., 2025; Song et al., 2026).

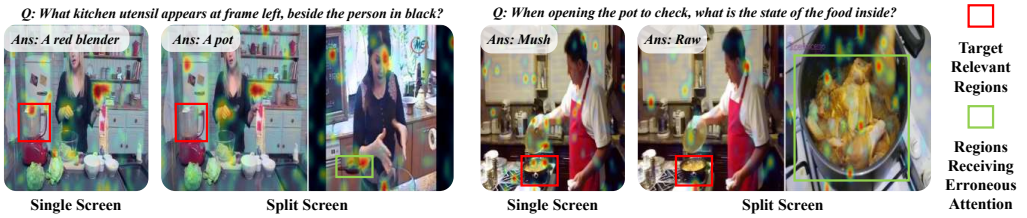

Figure 7: **Attention shifts induced by split-screen interference**. For each case, the left panel shows the single-video condition, whereas the right panel shows the split-screen condition. In the split-screen setting, the question asks specifically about the left video. However, the model's attention is disrupted by the concurrent right video, resulting in erroneous responses.

Although existing models perform well on single-video, our experiments suggest they still struggle with parallel streams, multiple objects, and concurrent scenes.

4.2 Memory Interference: Robustness to Distraction

Table 2: **Memory Interference**. Proactive: the first video disrupts recall of the second video. Retroactive: the second disrupts recall of the first. Δ denotes proactive minus retroactive.

	Accuracy (%, ↑)			Intrusion Rate (%, ↓)		
	Proactive	**Retroactive**	Δ	**Proactive**	**Retroactive**	Δ
Human	94.55	74.55	20.00	3.64	20.00	-16.36
Random	25.00	25.00	0.00	50.00	50.00	0.00
Closed-Source Models						
Gemini-3.1-Pro-Preview	63.64	54.55	9.09	23.64	30.91	-7.27
GPT-5.4	43.64	40.00	3.64	43.64	34.55	9.09
Open-Source Agents						
VideoLucy	29.09	43.64	-14.55	43.64	34.55	9.09
M3-Agent	43.64	36.36	7.28	40.00	34.55	5.45
Open-Source Models						
Qwen3.5-4B	29.09	38.18	-9.09	45.45	38.18	7.27
Qwen3-VL-8B-Instruct	25.45	29.09	-3.64	54.55	52.73	1.82
InternVL3.5-8B	52.73	49.09	3.64	32.73	41.82	-9.09
Qwen3.5-9B	29.09	38.18	-9.09	50.91	41.82	9.09
Qwen3.5-27B	45.45	40.00	5.45	40.00	43.64	-3.64

Main results. As shown in Table 2, most models achieve low accuracy, indicating that memory interference poses a significant challenge. Further, humans demonstrate a clear asymmetry between proactive and retroactive interference ($\Delta = 20.00\%$), yet models exhibit a small delta between two conditions. This suggests that the models differ from humans in memory mechanism where later information tends to overwrite earlier memories for humans. Notably, intrusion rates are high across most models, and thus most errors come from the interference of competing video. This indicates that models struggle to resist interference from semantically similar content.

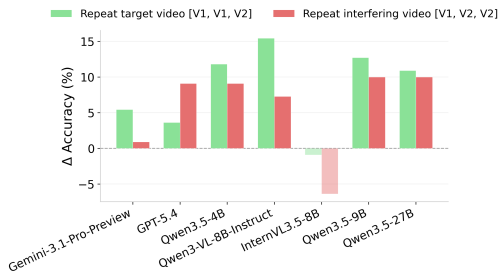

Figure 8: **Video repetition improves accuracy under interference.** Repeating either the target or interfering video yields performance gains, suggesting repetition as a promising strategy for enhancing model memory.

Further experiment. We test whether a repetition strategy can improve robustness to interference. This is done by repeating the target or the interfering video, forming [V1, V1, V2] and [V1, V2, V2] with questions about V1. As shown in Figure 8, both settings surprisingly improve accuracy over the baseline. We hypothesize that repetition helps models better distinguish the target video from the interfering video. Without repetition, causal attention allows later frames to attend to earlier frames, but not the reverse; with repetition, the later copy can attend to the earlier occurrence of the same video. This gives the model a clearer view of the repeated video, consistent with recent findings (Leviathan et al., 2025).

***Finding 2**: Retroactive interference exceeds proactive interference in humans, whereas both occur comparably in multi-modal models. Further experiments surprisingly find that repeating target or interfering videos can both enhance the understanding of target video.*

Discussion. Humans exhibit pronounced retroactive interference, whereas most models do not. We empirically observe that positional and recency biases in causal Transformers are not strong enough in the tested models to produce the pronounced directional asymmetry characteristic of human memory. Repetition strategy benefits both humans and models yet through different mechanisms. Humans leverage repetition to reinforce memory anchors (Hintzman, 1976), whereas models benefit from the strengthened representations of both target and interfering videos via causal attention.

4.3 Interleaved Events: Temporal Organization

Table 3: **Interleaved Events**. Accuracy (%) on four interleaved reconstruction metrics.

Acc(%)	Source Identification	Order Understanding	Content Retention	False Memory Discrimination
Human	75.95	80.00	83.64	82.11
Random	25.00	25.00	25.00	25.00
Closed-Source Models				
Gemini-3.1-Pro-Preview	43.04	50.00	49.09	26.32
GPT-5.4	43.04	40.00	47.27	7.37
Open-Source Agents				
VideoLucy	30.38	23.33	43.64	40.00
M3-Agent	27.85	40.00	21.82	15.79
Open-Source Models				
Qwen3.5-4B	30.38	20.00	41.82	23.16
Qwen3-VL-8B-Instruct	21.52	23.33	30.91	3.16
InternVL3.5-8B	25.32	26.67	41.82	1.05
Qwen3.5-9B	26.58	40.00	25.45	7.37
Qwen3.5-27B	39.24	33.33	34.55	3.16

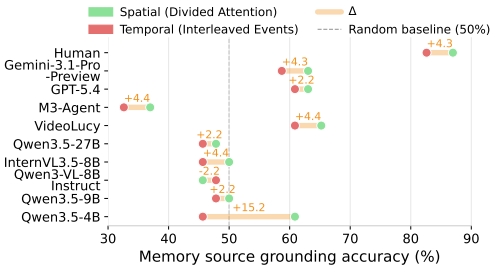

Figure 9: **Spatial source grounding outperforms temporal source grounding.** Spatial source uses the split-screen format with frequent left/right swaps (§4.1); temporal source uses the interleaved format (§4.3).

Main results. As shown in Table 3, humans substantially outperform all models across all four question types. These results demonstrate that reorganizing temporally interleaved events remains a significant challenge. Agentic methods show no clear advantage, suggesting that rule-based memory strategies are insufficient for handling temporal interleaving. Notably, most models achieve below the 25% random baseline on false memory discrimination, revealing severe hallucination.

Further experiment. To further examine the ability of memory source grounding (Johnson et al., 1993; Schacter et al., 1984), we compare grounding accuracy under spatial (split-screen with frequent left/right swaps, §4.1) versus temporal (interleaved, §4.3) conditions. As shown in Figure 9, spatial source grounding generally yields higher accuracy than temporal source grounding, where many models even fall below the random baseline. These results suggest that for both humans and models, accurately grounding the temporal source is more difficult than grounding the spatial source.

***Finding 3**: Compared to human memory, multi-modal models are less capable of organizing temporally interleaved information. Further analysis reveals that memory source grounding along the spatial dimension is consistently stronger than along the temporal dimension.*

Discussion. Models exhibit stronger spatial source grounding than temporal source grounding, mirroring an asymmetry observed in human cognition (Torres-Morales & Cansino, 2025; Pathman et al., 2018) and AI research (Upadhyay et al., 2025). This suggests that temporal memory organization is inherently more challenging. One potential direction is building models to better capture sequential relationships across events.

4.4 N-BACK: SYMBOL GROUNDING AND MEMORY CAPACITY

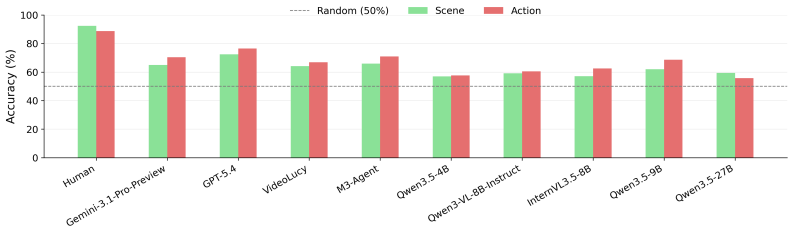

Figure 10: **Overall accuracy on the N-Back task.** Performance of each model and human under two symbolic attributes (scene and action), averaged over all K and N configurations.

Main results. Existing multi-modal models substantially lag behind humans, with many only slightly exceeding the random baseline. Among them, GPT-5.4 achieves the best performance. Interestingly, humans recall scene attributes more accurately than action attributes, whereas most models show the opposite pattern, with action accuracy being higher than scene accuracy.

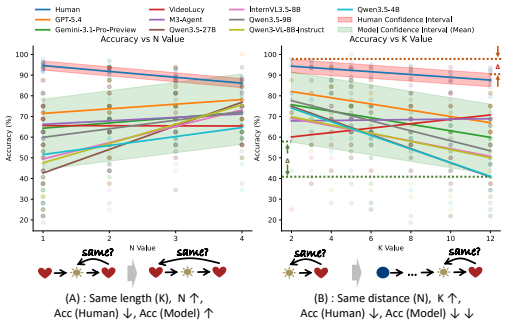

Figure 11: **Effects of N and K on accuracy.** Points show per-model accuracy under different (N, K) settings, with linear fits for each model. The colored filled regions indicate ± 1 standard deviation around the fit lines.

Further experiment. As shown in Figure 11, behavior discrepancy emerges between humans and models. For humans, monotonic decline of accuracy is observed with increasing N, reflecting capacity limit. When increasing K, accuracy decreases modestly, demonstrating an ability to discard irrelevant information. For example, when $N = 2$ and $K = 9$, the first six video clips are no longer relevant to the final decision and thus can be discarded. In contrast, model accuracy remains flat or even improves as N increases, likely due to the Transformer architecture that retrieves temporally distant clips through global attention. However, accuracy drops sharply with increasing K, suggesting that models struggle to filter out irrelevant information from memory.

> ***Finding 4**: Multi-modal models lag far behind humans in symbolic memory. Unlike humans, models do not decay when increasing temporal gap (N), yet degrade largely with increasing number of total symbols (K), revealing a fundamental inability to filter irrelevant memory.*

Discussion. In the N-Back task, humans typically maintain only recent items in working memory while gradually forgetting earlier ones. By contrast, current models retain all past inputs at a similar level of accessibility due to the attention mechanism. We hypothesize that introducing an appropriate forgetting mechanism could help multi-modal models overcome the limitations of symbolic memory, complementing recent explorations in AI research (Hu et al., 2023; Wang et al., 2024).

5 CONCLUSION

In this work, we introduce M^3Eval, the first benchmark for systematically measuring multi-modal memory across different dimensions. M^3Eval is grounded in cognitive psychology and instantiated through orchestrated video tasks, moving beyond conventional video understanding benchmarks to probe memory mechanisms critical for multi-modal models. Our experiments reveal consistent weaknesses and meaningful characteristics across models, pointing to several future directions: *(1)* refining attention mechanisms to better handle parallel streams; *(2)* leveraging repetition strategy to mitigate interference between similar memory traces; *(3)* strengthening temporal source grounding, which substantially lags behind spatial grounding; and *(4)* improving symbolic memory to support abstraction and filtering of task-irrelevant memory. We hope that M^3Eval will serve as a diagnostic tool for future research and motivate the development of multi-modal systems equipped with robust, structured, and human-aligned memory capabilities.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, or interpreting results, and formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, editing a research paper to improve readability, and creating or editing software code. We have reviewed all AI-assisted work. Specifically:

1. During data construction, the QA pairs in our benchmark were initially generated by AI based on the video content and our specifications. All AI-generated content was manually reviewed and verified by the authors.
2. AI-powered coding assistants were used during development. All code was verified and tested for correctness by 2 authors.
3. AI tools were used to polish the paper text and assist in creating figures. All polished content and figures were manually reviewed by the authors.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The benchmark construction process is described in Section 3, Section 4, Appendix F and Appendix G. The experimental settings are detailed in Section 3, Section 4, Appendix C and Appendix H. Based on these descriptions, reported results can be reproduced. Our code and data will be publicly available.

REFERENCES

John R Anderson and Gordon H Bower. *Human associative memory*. Psychology press, 2014.

Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A benchmark for large multi-modal models in long-form movies and tv shows. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19496–19523, 2025.

Alan Baddeley. Working memory. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, 321(2-3):167–173, 1998.

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

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

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Fergus IM Craik, Richard Govoni, Moshe Naveh-Benjamin, and Nicole D Anderson. The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, 125(2):159, 1996.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. Hugging Face model card, April 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro. Accessed: 2026-05-02.

James Deese. On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of experimental psychology*, 58(1):17, 1959.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Dongyu Gong, Xingchen Wan, and Dingmin Wang. Working memory capacity of chatgpt: An empirical study. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10048–10056, 2024.

Google DeepMind. Gemini 3.1 pro model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-05-02.

Douglas L Hintzman. Repetition and memory. *Psychology of learning and motivation*, 10:47–91, 1976.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025a.

Zi-Yuan Hu, Shuo Liang, Duo Zheng, Yanyang Li, Yeyao Tao, Shijia Huang, Wei Feng, Jia Qin, Jianguang Yu, Jing Huang, et al. Nemo: Needle in a montage for video-language understanding. *arXiv preprint arXiv:2509.24563*, 2025b.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Zixi Jia, Qinghua Liu, Hexiao Li, Yuyan Chen, and Jiqiang Liu. Evaluating the long-term memory of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19759–19777, 2025.

Zixia Jia, Jiaqi Li, Yipeng Kang, Yuxuan Wang, Tong Wu, Quansen Wang, Xiaobo Wang, Shuyi Zhang, Junzhe Shen, Qing Li, et al. The ai hippocampus: How far are we from human memory? *arXiv preprint arXiv:2601.09113*, 2026.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

Marcia K Johnson, Shahin Hashtroudi, and D Stephen Lindsay. Source monitoring. *Psychological bulletin*, 114(1):3, 1993.

Michael J Kahana and Anthony D Wagner. *The Oxford handbook of human memory, two volume pack: foundations and applications*. Oxford University Press, 2024.

D KAHNEMAN. Attention and effort. *Experimental psychology*, 1973.

Wayne K Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*, 2025.

Jingyao Li, Jingyun Wang, Molin Tan, Haochen Wang, Cilin Yan, Likun Shi, Jiayin Cai, Xiaolong Jiang, and Yao Hu. Crossvid: A comprehensive benchmark for evaluating cross-video reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6244–6252, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Jiafeng Liang, Hao Li, Chang Li, Jiaqi Zhou, Shixin Jiang, Zekun Wang, Changkai Ji, Zhihao Zhu, Runxuan Liu, Tao Ren, et al. Ai meets brain: Memory systems from cognitive neuroscience to autonomous agents. *arXiv preprint arXiv:2512.23343*, 2025.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Ruixun Liu, Lingyu Kong, Derun Li, and Hang Zhao. Occvla: Vision-language-action model with implicit 3d occupancy supervision. *arXiv preprint arXiv:2509.05578*, 2025.

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. *arXiv preprint arXiv:2508.09736*, 2025.

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

Jean M Mandler. A code in the node: The use of a story schema in retrieval. *Discourse processes*, 1 (1):14–35, 1978.

Jean M Mandler and Nancy S Johnson. Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, 9(1):111–151, 1977.

John A McGeoch. Forgetting and the law of disuse. *Psychological review*, 39(4):352, 1932.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18902–18913, 2025.

OpenAI. Gpt-5.4 thinking system card, March 2026. URL `https://openai.com/zh-Hans-CN/index/introducing-gpt-5-4/`. Accessed: 2026-05-02.

Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59, 2005.

Thanujeni Pathman, Christine Coughlin, and Simona Ghetti. Space and time in episodic memory: Effects of linearity and directionality on memory for spatial location and temporal order in children and adults. *PLoS One*, 13(11):e0206999, 2018.

Zenon W Pylyshyn. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological bulletin*, 80(1):1, 1973.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`. Accessed: 2026-05-02.

Edward Stevens Robinson. Some factors determining the degree of retroactive inhibition. *Psychological Monographs*, 28(6):i, 1920.

Henry L Roediger and Kathleen B McDermott. Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 21(4): 803, 1995.

Daniel L Schacter, Joanne L Harbluk, and Donald R McLachlan. Retrieval without recollection: An experimental analysis of source amnesia. *Journal of verbal learning and verbal behavior*, 23(5): 593–611, 1984.

John W Schwieter and Zhisheng Edward Wen. *The Cambridge handbook of working memory and language*. Cambridge University Press, 2022.

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, 2024.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Mingyang Song, Mao Zheng, and Xuan Luo. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3753–3763, 2025b.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18549–18557, 2026.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

César Torres-Morales and Selene Cansino. Neurophysiological distinctions between spatial and temporal context in episodic memory. *International Journal of Psychophysiology*, pp. 113302, 2025.

Anne Treisman and Hilary Schmidt. Illusory conjunctions in the perception of objects. *Cognitive psychology*, 14(1):107–141, 1982.

Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

Benton J Underwood. Interference and forgetting. *Psychological review*, 64(1):49, 1957.

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time blindness: Why video-language models can't see what humans can? *arXiv preprint arXiv:2505.24867*, 2025.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. Symbolic working memory enhances language models for complex rule application. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17583–17604, 2024.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

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

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28885–28900, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L. Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*, 2025c.

Franklin M Zaromb, Marc W Howard, Emily D Dolan, Yevgeniy B Sirotin, Michele Tully, Arthur Wingfield, and Michael J Kahana. Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792, 2006.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. Working memory identifies reasoning limits in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16896–16922, 2024.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms. *arXiv preprint arXiv:2406.09367*, 2024.

Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, ZhongZhi Li, Yingying Zhang, Le Song, and Qianli Ma. Lifelongagentbench: Evaluating llm agents as lifelong learners. *arXiv preprint arXiv:2505.11942*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025a.

Wenqi Zhou, Kai Cao, Hao Zheng, Yunze Liu, Xinyi Zheng, Miao Liu, Per Ola Kristensson, Walterio Mayol-Cuevas, Fan Zhang, Weizhe Lin, et al. X-lebench: A benchmark for extremely long egocentric video understanding. *arXiv preprint arXiv:2501.06835*, 2025b.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. *arXiv preprint arXiv:2508.19542*, 2025.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. *arXiv preprint arXiv:2510.12422*, 2025.

A BENCHMARK STATISTICS

A.1 QUESTION COUNT

The full M^3Eval benchmark comprises 2,403 questions, organized into two parts.

Non-N-Back Questions. This part contains 739 questions covering divided attention, memory interference, and interleaved events (§4.1, 4.2, 4.3). We construct the question-answer pairs from 451 videos sourced from six public long-video understanding datasets. Table 4 details the distribution of questions across these datasets.

Table 4: Composition of the non-N-Back portion of M^3Eval by source dataset.

Dataset	Questions	Videos
CrossVid-CC	138	85
CrossVid-NC	96	70
HourVideo	100	54
InfiniBench-TVQA	184	95
LVBench	102	71
Video-MME-L	119	76
Total (Non-N-Back)	739	451

N-Back Questions. This part includes 1,664 questions. We generate them from 64 carefully selected 12-clip sequence instances, evenly split into 32 for the *action* attribute and 32 for the *scene* attribute. Each instance yields 26 valid $K \times N$ combinations, ensuring comprehensive coverage across different memory loads and temporal gaps.

A.2 FINE-GRAINED PER-METRIC BREAKDOWN

Table 5 provides a detailed breakdown of question and video counts for each paradigm and metric.

Table 5: Fine-grained question and video counts per paradigm and metric.

Paradigm	**Metric**	**Questions**	**Videos**
Divided Attention	Source Identification	96	96
	Order Understanding	80	80
	Content Retention	102	102
	Subtotal	*278*	*190*
Memory Interference	Proactive	55	55
	Retroactive	55	55
	Subtotal	*110*	*110*
Interleaved Events	Source Identification	79	79
	Order Understanding	30	30
	Content Retention	55	55
	False Memory Discrimination	95	95
	Subtotal	*259*	*132*

Note: Different metrics may share constructed videos, so per-metric video counts do not sum to subtotals.

For the N-Back paradigm, 1,664 questions span a full $K \times N$ matrix with $K \in \{2, 3, 4, 5, 6, 8, 10, 12\}$ and $N \in \{1, 2, 3, 4\}$. Each cell contains 64 questions (32 action + 32 scene).

A.3 ANSWER DISTRIBUTION

For the non-N-Back 4-choice questions, the construction pipeline performs a random shuffle on options after generation. This results in a uniform distribution of correct answers across A/B/C/D

with no position bias. For the N-Back Yes/No questions, match and mismatch instances are balanced by construction within each $K \times N$ combination.

A.4 Video Duration

Figure 12 shows the duration distribution of the 451 source videos used for the non-N-Back tasks. Each clip used in the N-Back tasks is trimmed from these source videos.

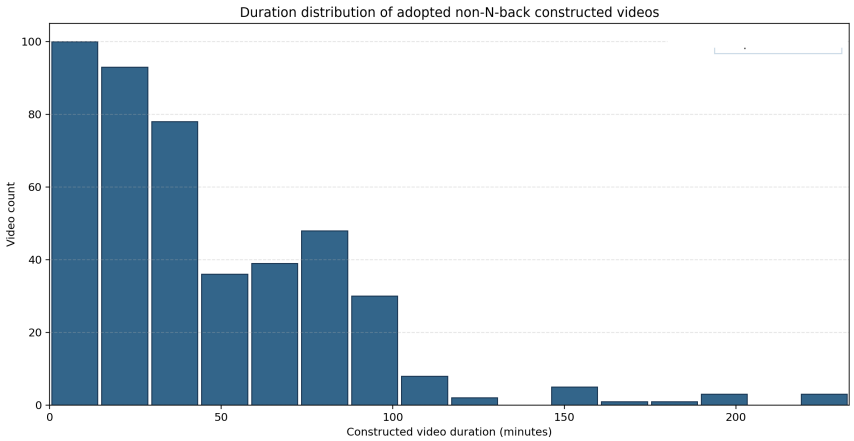

Figure 12: Duration histogram for the source videos in the non-N-Back portion of M^3Eval.

B Single-Video Baseline Experiments

To quantify the performance drop caused by multi-video memory demands, we conduct single-video baseline experiments for all three non-N-Back paradigms. In each baseline, only the target video is presented to the model, removing the interfering or concurrent video. The same questions are used as in the multi-video evaluation. This within-model comparison directly measures how much performance degrades when a second video is introduced.

B.1 Divided Attention

Table 6 reports accuracy under the dual-video (split-screen) and single-video conditions. Because the single-video input is identical regardless of the swap setting, the single-video baseline is the same across the No-Swap and Swap columns.

Table 6: **Single-video baselines for Divided Attention.** Format: dual-video → single-video accuracy (%). The Avg. column shows the average improvement per model.

Model	No Swapping			Swapping			Avg.
	Source Ident.	Order Underst.	Content Ret.	Source Ident.	Order Underst.	Content Ret.	
GPT-5.4	27.08→72.92	35.00→52.50	47.06→64.71	35.42→72.92	30.00→52.50	49.02→64.71	↑26.11
Qwen3.5-4B	18.75→56.25	25.00→35.00	31.37→50.98	14.58→56.25	22.50→35.00	33.33→50.98	↑23.15
Qwen3-VL-8B	16.67→66.67	25.00→47.50	37.25→50.98	12.50→66.67	30.00→47.50	35.29→50.98	↑28.93
InternVL3.5-8B	29.17→68.75	37.50→55.00	33.33→39.22	25.00→68.75	40.00→55.00	27.45→39.22	↑22.25
Qwen3.5-9B	35.42→56.25	25.00→27.50	25.49→54.90	18.75→56.25	30.00→27.50	13.73→54.90	↑21.49
Qwen3.5-27B	41.67→64.58	25.00→50.00	35.29→52.94	27.08→64.58	32.50→50.00	35.29→52.94	↑23.04
VideoLucy	16.67→58.33	42.50→37.50	37.25→49.02	14.58→58.33	25.00→37.50	39.22→49.02	↑19.08

All models show consistent accuracy improvements (+19–29 pp on average) from dual-video to single-video, confirming that the performance drops in the main experiments arise from multi-video memory demands rather than perceptual limitations on individual videos.

B.2 Memory Interference

Table 7 reports accuracy under the two-video concatenation and single-video conditions.

Table 7: **Single-video baselines for Memory Interference.** Format: dual-video → single-video accuracy (%).

Model	Proactive Acc	Retroactive Acc	Avg.
GPT-5.4	43.64→52.73	40.00→52.73	↑10.91
Qwen3.5-4B	29.09→54.55	38.18→54.55	↑20.91
Qwen3-VL-8B	25.45→54.55	29.09→54.55	↑27.28
InternVL3.5-8B	52.73→61.82	49.09→61.82	↑10.91
Qwen3.5-9B	29.09→52.73	38.18→52.73	↑19.09
Qwen3.5-27B	45.45→58.18	40.00→58.18	↑15.45
VideoLucy	29.09→60.00	43.64→60.00	↑23.63

Single-video accuracy is consistently higher than dual-video accuracy across all models, with improvements ranging from +10.91 to +27.28 pp. These results validate that performance drops under the memory interference paradigm are driven by the presence of semantically similar competing content.

B.3 Interleaved Events

Table 8 reports accuracy under the interleaved and single-video conditions.

Table 8: **Single-video baselines for Interleaved Events.** Format: interleaved → single-video accuracy (%).

Model	Source Identification	Order Understanding	Content Retention	Avg.
GPT-5.4	43.04→74.68	40.00→46.67	47.27→58.18	↑16.40
Qwen3.5-4B	30.38→56.96	20.00→33.33	41.82→50.91	↑16.34
Qwen3-VL-8B	21.52→69.62	23.33→36.67	30.91→52.73	↑27.76
InternVL3.5-8B	25.32→67.09	26.67→43.33	41.82→43.64	↑20.08
Qwen3.5-9B	26.58→62.03	40.00→30.00	25.45→61.82	↑20.60
Qwen3.5-27B	39.24→70.89	33.33→43.33	34.55→50.91	↑19.33
VideoLucy	30.38→59.49	23.33→43.33	43.64→43.64	↑16.37

The same pattern holds: all models improve substantially (+16–28 pp on average) when the interleaved structure is removed, confirming that temporal interleaving poses a genuine memory challenge beyond basic perception.

C Human Evaluation Protocol

This section describes the manual review process for benchmark construction and the human baseline evaluation protocol.

C.1 Manual Review Process

All candidate questions generated by Qwen3.5-27B undergo a manual review before inclusion in the benchmark. Three reviewers (co-authors) independently assessed each question on two criteria: (1) factual accuracy with respect to the video content, and (2) conformity to the design requirements of the corresponding task paradigm.

The initial pairwise agreement rate was 91.4%, with Fleiss' $\kappa = 0.84$, indicating high inter-annotator reliability. Disagreements were resolved through discussion. The entire review process required approximately 37 person-hours. As a result, 137 questions were discarded and 125 were revised.

C.2 Human Baseline Protocol

Five expert volunteers participated in the human baseline evaluation. None of the five overlapped with the three reviewers above. The evaluation followed these procedures:

Question assignment. The 2,403 questions were evenly divided into five groups (∼480 each), balanced by task type. Each question was answered by exactly one participant.

Viewing format. Participants watched the same constructed videos as the models (i.e., the complete split-screen, concatenated, or interleaved videos), not individual sampled frames. This means human participants had access to the full visual information.

Pause and rewind. Participants were allowed to pause and rewind. Our benchmark includes agent-based methods (e.g., VideoLucy (Zuo et al., 2025)) that iteratively re-access video content during inference, so we grant human participants the same capability.

Cross-participant variance. The standard deviation of overall accuracy across the five participants was ∼2.8%, indicating low inter-participant variability.

D Human Agreement Analysis

To further assess the reliability of the human baseline, we conducted a human agreement test on a representative subset.

D.1 Setup

We randomly sampled 200 questions, uniformly stratified across all task types and their sub-categories. Four participants independently answered the same 200 questions. This sample size is sufficient for reliably computing agreement metrics.

D.2 Overall Results

Table 9: Human agreement analysis on a 200-question representative subset.

Metric	Value
Average accuracy (4 participants)	88.62%
Per-participant accuracy	89.50% / 90.00% / 86.00% / 89.00%
Inter-participant SD	1.797 pp
Pairwise Agreement	84.58%
Fleiss' κ	0.8043

The four participants achieve similar accuracy (SD = 1.797 pp), and their answers on the same questions show high agreement (Pairwise Agreement 84.58%, Fleiss' κ = 0.8043). These results indicate that individual differences have minimal impact on human performance.

D.3 Per-Task Breakdown

Table 10: Human agreement breakdown by task type.

Task	Inter-participant SD	Fleiss' κ	Pairwise Agreement
Divided Attention	2.062 pp	0.7897	84.52%
Memory Interference	4.083 pp	0.8400	88.33%
Interleaved Events	2.722 pp	0.7531	83.33%
N-Back	2.049 pp	0.7189	84.29%

All four task types show consistently high agreement and low variance. These results suggest that our questions do not suffer from notable ambiguity, and that the human baseline is stable and reliable.

E Downstream Task Correlation

To examine whether memory performance correlates with broader video understanding, we evaluate models on two established long-video benchmarks: LVBench (Wang et al., 2025a) and VideoMME-long (Fu et al., 2025).

E.1 Ranking Comparison

Table 11 shows model rankings on M^3Eval alongside rankings on LVBench and VideoMME-long. Models scoring higher on M^3Eval consistently score higher on both downstream benchmarks.

Table 11: Model performance and rankings on M^3Eval versus downstream benchmarks.

Model	M^3Eval	**Rank**	**LVBench**	**Rank**	**VideoMME-long**	**Rank**
Gemini-3.1-Pro-Preview	55.17%	1	78.20%	1	80.30%	1
InternVL3.5-8B	41.62%	2	42.22%	2	52.22%	6
Qwen3.5-27B	40.17%	3	42.09%	4	64.44%	2
Qwen3.5-9B	37.14%	4	21.30%	6	57.11%	3
Qwen3.5-4B	36.00%	5	42.16%	3	57.00%	4
Qwen3-VL-8B	33.23%	6	40.28%	5	56.33%	5

Note: M^3Eval scores are computed as the average accuracy across all evaluated tasks.

E.2 Correlation Analysis

We compute Pearson correlations using an exact permutation test (720 permutations).

Table 12: Pearson correlation between M^3Eval and downstream benchmarks.

	Pearson r	R^2	**Permutation** p
$M^3Eval \rightarrow$ LVBench	0.853	0.728	0.069
$M^3Eval \rightarrow$ VideoMME-long	0.857	0.735	0.081
$M^3Eval \rightarrow$ Combined (z-scored)	0.893	0.797	0.071

All Pearson $r > 0.85$ with $R^2 > 0.72$. Pairwise rank consistency is 80.0% with LVBench (Spearman $\rho = 0.714$) and 73.3% with VideoMME-long. These results indicate that memory performance, as measured by M^3Eval, is a meaningful predictor of broader long-video understanding.

F Video Construction Details

We source video materials from five public datasets and benchmarks: HourVideo (Chandrasegaran et al., 2024), Video-MME (long-video subset) (Fu et al., 2025), LVBench (Wang et al., 2025a), InfiniBench (TVQA subset) (Ataallah et al., 2025), and CrossVid (Li et al., 2026). These datasets cover egocentric daily activities, diverse web videos, TV episodes, cooking tutorials, and movies, ensuring broad coverage of real-world video scenarios. We construct video pairs following a semantic similarity-first principle. Within each dataset, videos with similar topics, scenes, or narrative structures are paired. This design is motivated by findings in cognitive psychology that memory interference is strongest between semantically similar materials (Underwood, 1957).

We use a metadata-guided intra-dataset pairing strategy rather than per-frame visual embeddings or aggregated visual distances. Each video is paired only with videos from the same source dataset, using existing structured annotations such as category labels, scene tags, episode/season information, subtitle text similarity, question-type labels, and video duration.

For example, in InfiniBench-TVQA, we pair episodes from the same show and season with close episode numbers. We additionally consider subtitle text similarity (TF-IDF-based, not visual features), skill-label distribution similarity, and video duration. This ensures high comparability in narrative content and task structure. All video pairs underwent manual verification to confirm that the pairing effectively induces the memory interference effects our benchmark requires.

G QA Construction Details

Questions are automatically generated using Qwen3.5-27B (Qwen Team, 2026) and refined through manual review. We design two separate pipelines for the different memory dimensions targeted by our benchmark. The first, non-N-Back QA construction, covers Divided Attention, Memory Interference, Interleaved Events, and source-memory judgment sub-tasks. The second, N-Back QA construction, targets tracking abstract symbolic attributes (scene or action categories) over a video stream.

G.1 Non-N-Back QA Construction

This pipeline generates questions through a multi-stage process: (1) video segmentation; (2) hierarchical description extraction; (3) model-based question generation using predefined prompts; and (4) manual filtering and verification. We manually filter out controversial and potentially composite scenarios, and ensure that the labels within each group are free from interference and ambiguity.

G.1.1 Video Segmentation

Instead of processing videos end-to-end, we first segment them into short, localized units by sampling frames and grouping them into local segments. These segments serve as the basic units for all subsequent description extraction and question generation steps.

G.1.2 Hierarchical Description Extraction

For each video, we extract structured evidence at both the local and global levels to prevent the language model from hallucinating or relying on unstructured information. Each segment is described using a predefined six-key schema, as summarized in Table 13.

Table 13: Structured caption schema for video segments.

Key	Description
`main_storyline`	Plot progression, operational step, or event stage.
`spatial_relation_binding`	Layout of people, objects, text, or markers relative to local screen space (left/right, above/below, on-screen region).
`short_term_action_state`	Transient action states or instantaneous changes (open/close, pick up/put down, brief flashes).
`tool_prop`	Tools, containers, handheld objects, or props relevant to the current operation.
`fine_visual_attribute`	Colors, textures, materials, shapes, accessories, or other fine-grained appearance details.
`text_symbol`	Text, numbers, icons, labels, logos, or other symbolic information.

G.1.3 Question Generation Prompts

Below, we show the prompts fed to the model during the question generation stage for each task and failure mode.

VIDEO CAPTION GENERATION

```
You are a professional video evidence describer preparing context for
difficult memory questions over two highly similar videos.
You will receive a small batch of nearby frames from one video.
Describe only grounded, directly visible evidence from these frames.
Return STRICT JSON only with exactly these 6 keys: main_storyline,
spatial_relation_binding, short_term_action_state, tool_prop,
fine_visual_attribute, text_symbol.
Write every value as one grounded English string. If a category is
not supported by visible evidence in this frame batch, return an
empty string for that key. Never replace these fields with a single
caption key. Do not add wrapper keys such as result, response, or
data.
Output format:
{
  "main_storyline": "",
  "spatial_relation_binding": "",
  "short_term_action_state": "",
  "tool_prop": "",
  "fine_visual_attribute": "",
  "text_symbol": ""
}
Key definitions:
1. main_storyline: State the local storyline progression or process
step shown by this batch. Also include the directly visible people,
entities, objects, actions, and scene context that support that
storyline reading. If the frames imply a short local progression,
describe that progression clearly.
2. spatial_relation_binding: Describe detail-level spatial
relations or position binding. Focus on who or what is to the
left/right/front/behind, what is on top of or inside something, where
a marking appears, or where text/icons appear within the frame.
3. short_term_action_state: Describe short-lived action states
or momentary changes visible in this batch. Focus on brief
states such as on/off, open/half-open/closed, held/released,
appearing/disappearing, flashing/persistent, steaming/not steaming,
and other transient states.
4. tool_prop: Describe visible tools, props, utensils, containers,
handheld objects, manipulated small items, or operation-relevant
objects.
5. fine_visual_attribute: Describe subtle visual attributes such as
colors, patterns, materials, textures, shapes, markings, accessories,
or other precise appearance details.
6. text_symbol: Record readable text, numbers, labels, logos, icons,
signs, subtitles, packaging text, or symbolic graphics. If only part
is readable, say what is readable and what is uncertain.
Requirements: Use objective, concrete English. Do not speculate
beyond what is visible or strongly implied by the shown frames.
Keep each field concise but specific. Preserve short-lived and
discriminative evidence instead of collapsing everything into broad
summary. Output exactly one JSON object and nothing else.
```

SOURCE IDENTIFICATION

Role You are a professional video memory test question designer.
Task Two semantically similar videos are shown side by side throughout the full clip. You will receive sparse frame-batch caption contexts for Video 1 and Video 2. Design exactly 1 multiple-choice storyline-reconstruction question asking which narrative paragraph most accurately describes what happened in one target video.
Focus Build the question around the main storyline, process stages, or high-level event progression of the target video rather than tiny fleeting details.
2×2 Combinatorial Design
- This prompt is for source_confusion.
- Option A must be the only correct paragraph before downstream option shuffle.
- Select exactly two non-adjacent manipulable positions, slot_1 and slot_2, inside one shared storyline scaffold.
- For each slot, prepare: the real target-video main-stage detail as the correct version; a corresponding main-stage detail from the other video as the incorrect version.
- The four options must realize the full 2×2 grid: A: slot_1=correct, slot_2=correct; B: slot_1=incorrect, slot_2=correct; C: slot_1=correct, slot_2=incorrect; D: slot_1=incorrect, slot_2=incorrect.
- All three wrong options must stay within the single error family source_confusion.

Writing Constraints
1. The question must ask: ''Which of the following options most accurately describes what happened in Video X?''
2. Each option must be one coherent narrative paragraph of about 90--160 words.
3. All four options must keep nearly the same paragraph structure, writing style, and overall narrative scaffold.
4. Differences should center on meaningful storyline stages, not trivial wording noise.
5. Do not turn the paragraph into bullet-like stage labels; keep fluent prose.
6. The four option strings must all be distinct; never output two identical paragraphs.

Self-Check Before Finalizing
1. All four options are individually coherent and share the same high-level storyline scaffold.
2. A/B/C/D exactly realize the correct-correct / incorrect-correct / correct-incorrect / incorrect-incorrect grid.
3. option_role_by_letter stays coarse-grained: A=correct, B/C/D=source_confusion.
4. storyline_2x2_design.slot_1 and slot_2 each use exact verbatim anchor spans copied from the option text so a checker can verify the expected presence/absence; do not paraphrase or inflect those anchor spans across options.
5. The two slots are separated in the paragraph and do not collapse into the same local phrase.
6. B differs from A only at slot_1, C differs from A only at slot_2, and D differs from A at both slots.

ORDER UNDERSTANDING

Role You are a professional video memory test question designer.
Task Two semantically similar videos are shown side by side throughout the full clip. You will receive sparse frame-batch caption contexts for Video 1 and Video 2. Design exactly 1 multiple-choice storyline-reconstruction question asking which narrative paragraph most accurately describes what happened in one target video.
Focus Build the question around the main storyline, process stages, or high-level event progression of the target video rather than tiny fleeting details.
2×2 Combinatorial Design
- This prompt is for order_disruption.
- Option A must be the only correct paragraph before downstream option shuffle.
- Select exactly two non-overlapping reversible event pairs, slot_1 and slot_2, inside one shared storyline scaffold.
- For each slot, prepare: the true temporal order as the correct version; a locally plausible reversed order as the incorrect version.
- Represent each slot with the same two short event clauses reused word-for-word across all four options. Form the incorrect version only by swapping the order of those exact clauses.
- The four anchor clauses across slot_1 and slot_2 must all be distinct. Do not reuse one event clause in both slots.
- In every option, each anchor clause must appear exactly once. Do not duplicate or omit an anchor clause when building a distractor.
- Keep each slot as one local two-clause micro-sequence in place. Reverse the two clauses inside the same local sentence window; do not move one anchor clause to another sentence or another slot region.
- The four options must realize the full 2×2 grid: A: slot_1=correct, slot_2=correct; B: slot_1=incorrect, slot_2=correct; C: slot_1=correct, slot_2=incorrect; D: slot_1=incorrect, slot_2=incorrect.
- All three wrong options must stay within the single error family order_disruption.
- Both slots must genuinely change somewhere in the final options. Do not leave one slot text unchanged across all four options.
- Construct A first. Then create B, C, and D by copying A and swapping only the required slot order in place. Do not rewrite the paragraph from scratch for each letter.

Writing Constraints
1. The question must ask: ''Which of the following options most accurately describes what happened in Video X?''
2. Each option must be one coherent narrative paragraph of about 90--160 words.
3. All four options must keep nearly the same paragraph structure, writing style, and overall narrative scaffold.
4. Differences should center on meaningful storyline stages rather than trivial wording noise.
5. Do not turn the paragraph into bullet-like stage labels; keep fluent prose.
6. The four option strings must all be distinct; never output two identical paragraphs.
7. Build slot_1 and slot_2 as two separate local windows, preferably two separate sentences. Do not let the two slots overlap or share a sentence fragment.

Self-Check Before Finalizing
1. All four options are individually coherent and share the same high-level storyline scaffold.

2. A/B/C/D exactly realize the correct-correct / incorrect-correct / correct-incorrect / incorrect-incorrect grid.
3. option_role_by_letter stays coarse-grained: A=correct, B/C/D=order_disruption.
4. For each slot, correct_anchors and incorrect_anchors must be ordered anchor lists copied verbatim from the option text so a checker can verify the local order; the same anchor wording must be reused word-for-word in every option.
5. The two slots are separated in the paragraph and the four anchor clauses are all distinct; no anchor phrase is reused across slots.
6. Every option contains each anchor clause exactly once; no anchor clause is duplicated or omitted.
7. Each slot is reversed in place inside its own local sentence window; no anchor clause is moved to another sentence or another slot region.
8. B differs from A only at slot_1, C differs from A only at slot_2, and D differs from A at both slots.
9. Before finalizing, literally check that all four anchor clauses from A still appear once each in B, C, and D.

CONTENT RETENTION

Role You are a professional video memory test question designer.
Task Two semantically similar videos are shown side by side throughout the full clip. You will receive sparse frame-batch caption contexts for Video 1 and Video 2. Design exactly 1 multiple-choice storyline-reconstruction question asking which narrative paragraph most accurately describes what happened in one target video.
Focus Build the question around the main storyline, process stages, or high-level event progression of the target video rather than tiny fleeting details.
2×2 Combinatorial Design
- This prompt is for content_forgetting.
- Option A must be the only correct paragraph before downstream option shuffle.
- Select exactly two non-adjacent manipulable positions, slot_1 and slot_2, inside one shared storyline scaffold.
- For each slot, prepare: the real target-video main-stage detail as the correct version; a plausible but nonexistent main-stage detail as the incorrect version.
- The four options must realize the full 2×2 grid: A: slot_1=correct, slot_2=correct; B: slot_1=incorrect, slot_2=correct; C: slot_1=correct, slot_2=incorrect; D: slot_1=incorrect, slot_2=incorrect.
- All three wrong options must stay within the single error family content_forgetting.

Writing Constraints
1. The question must ask: ``Which of the following options most accurately describes what happened in Video X?''
2. Each option must be one coherent narrative paragraph of about 90--160 words.
3. All four options must keep nearly the same paragraph structure, writing style, and overall narrative scaffold.
4. Differences should center on meaningful storyline stages rather than trivial wording noise.
5. Do not turn the paragraph into bullet-like stage labels; keep fluent prose.
6. The four option strings must all be distinct; never output two identical paragraphs.

Self-Check Before Finalizing

```
1. All four options are individually coherent and share the same
   high-level storyline scaffold.
2. A/B/C/D exactly realize the correct-correct / incorrect-correct /
   correct-incorrect / incorrect-incorrect grid.
3. option_role_by_letter stays coarse-grained:  A=correct,
   B/C/D=content_forgetting.
4. storyline_2x2_design.slot_1 and slot_2 each use exact verbatim
   anchor spans copied from the option text so a checker can verify
   the expected presence/absence; do not paraphrase or inflect those
   anchor spans across options.
5. The two fabricated details never appear in either source video and
   do not collapse into the same local phrase.
6. B differs from A only at slot_1, C differs from A only at slot_2,
   and D differs from A at both slots.
```

FALSE MEMORY DISCRIMINATION

You are generating exactly ONE 4-option false-source question for two similar videos named video1 and video2.
Canonical type: question_type_id: false_source_attribution; question_type: False Source Attribution.
Goal Ask about a semantically related event or detail that appears in neither video. The false event should feel relevant to the two videos, not random.
Required Design
- Use only the provided segment summaries, global summaries, and pair summary as grounding for what is related and what is absent.
- The stem detail must be absent from both videos.
- Use exactly these 4 option meanings: Video 1, Video 2, Both videos, Neither video.
- The options themselves must be plain source labels. Do NOT paraphrase them as correct video / wrong video.
- Use the source labels directly as answer texts, for example A. Video 1, B. Video 2, C. Both videos, D. Neither video.
- The correct answer must therefore be the Neither video option.
- Set target_video to neither.

MEMORY INTERFERENCE

Role You are a professional video memory test question designer.
Task Two semantically similar videos are concatenated into a single continuous video for the model to watch. You will receive sparse frame-caption contexts for each video. Design exactly 1 multiple-choice question that tests whether recall of one video's fine visual attribute is invaded by a similar attribute from the other video.
Focus Prioritize subtle appearance evidence such as: colors, patterns, materials, shapes, markings, textures, accessories, or small appearance cues; details visible only briefly and easy to confuse across similar scenes; grounded visual attributes rather than broad plot summaries.
Runtime Compatibility
- Treat the provided Video 1 context and Video 2 context as the only source evidence.
- Generate exactly ONE 4-option multiple-choice question.
- Canonical type: question_type_id: memory_interference; question_type: Memory Interference.
- Set target_video to video1 or video2.
- Set target_memory_ability to memory_interference_fine_visual_attribute.
- Include option_role_by_letter with exactly one correct, two intrusion, and one irrelevant_distractor.
- Set intrusion_option_letter to one of the intrusion letters.
- Return STRICT JSON only. No markdown. No explanation.

SOURCE MEMORY

Role You are a professional video memory test question designer.
Task Two semantically similar videos are presented to the model. You will receive sparse frame-caption contexts for Video 1 and Video 2. Design exactly 1 source-memory multiple-choice question about a short-lived fine visual attribute that truly appeared in only one video.
Focus Prioritize subtle appearance evidence such as: colors, patterns, materials, shapes, markings, textures, accessories, or small appearance cues; details that are visible briefly and are easy to confuse across similar scenes; grounded visual attributes rather than broad plot summaries.
Runtime Compatibility
- Treat the provided Video 1 context and Video 2 context as the only source evidence.
- Generate exactly ONE 2-option multiple-choice question.
- Canonical type: question_type_id: source_memory; question_type: Source Memory.
- Use exactly these 2 answer texts: A. Video 1, B. Video 2.
- Set target_video to video1 or video2.
- Set target_memory_ability to source_memory_fine_visual_attribute.
- Include option_role_by_letter as a JSON object with exactly: A: video1_only; B: video2_only.
- Return STRICT JSON only. No markdown. No explanation.

G.1.4 MANUAL REVIEW AND VERIFICATION

All generated candidate questions undergo a rigorous manual review. We verify the logical consistency of the options, ensure that distractors are plausible but demonstrably incorrect, and revise wording when necessary to eliminate ambiguity. Only questions that pass this quality check are included in the final benchmark.

G.2 N-BACK QA CONSTRUCTION

We first uniformly segment each video into clips. Then, we employ Qwen3.5-27B to annotate each clip with its scene and action attributes via sequential prompting. Specifically, we process clips in temporal order and maintain a running list of all previously assigned attribute phrases. At each step, the existing action and scene labels are injected into the prompt, encouraging the model to reuse consistent phrasing for recurring attributes. A new phrase is introduced only when a genuinely novel action or scene appears. The annotation prompt is shown below.

N-BACK ATTRIBUTE ANNOTATION

```
You are analyzing a sequence of video segments from the same context.
Your task is to provide two short phrases for the current segment:
1. The main Action/Activity.
2. The Scene/Environment.
Current Global Memory for this group:
Existing Actions: [list of action phrases assigned to previous
clips]
Existing Scenes: [list of scene phrases assigned to previous clips]
Instructions:
- If the current action/scene matches one in the memory, use the
EXACT phrase from the list.
- If it is new, create a concise new phrase (2-4 words).
- OUTPUT FORMAT: Action: [phrase] | Scene: [phrase]
```

Based on the similarity of these clip-level attributes, we select four groups of videos. From each group, three clips are sampled and randomly combined to construct N-Back testing video sets. All

generated N-Back probes undergo the same manual review process described above, ensuring that the annotated attributes are accurate and that each probe has a single unambiguous correct answer.

H Experimental Details

For frame sampling, we use 0.5 FPS for Gemini-3.1-Pro-Preview and 96 uniform frames for all other models by default, with two exceptions: 144 frames for the repeated-trial experiments in Fig. 8 and 8 frames per clip for the N-Back experiments in Sec. 4.4. *For all other settings, we use the defaults*. All experiments with locally deployed models were conducted on a server equipped with 4 NVIDIA A800 GPUs. Proprietary models were evaluated through their official APIs.

Different models in our evaluation use different frame sampling configurations. Specifically, Gemini-3.1-Pro-Preview natively supports FPS-based frame input and adapts naturally to videos of varying durations in long-video scenarios. Other models are better suited to fixed-count uniform frame sampling. We follow each model's native recommended input setting to ensure evaluation under its optimal operating condition.

This practice is standard in the field. Established video understanding benchmarks, including Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025a), and InfiniBench (Ataallah et al., 2025), all adopt model-specific frame sampling configurations rather than enforcing a uniform frame count. The choice of frame rate is part of the model configuration itself.

I EXAMPLE VISUALIZATION

Here we present examples for each type of task. Each example includes: 1) the format in which the video is presented, 2) the task-specific prompt given to the model, and 3) the specific question asked.



Figure 13: Example of Divided Attention targeting *Source Identification*. The three distractors replace certain content in the target video's narrative with content from the distractor video, while the correct option (highlighted in yellow) faithfully describes only the target video.

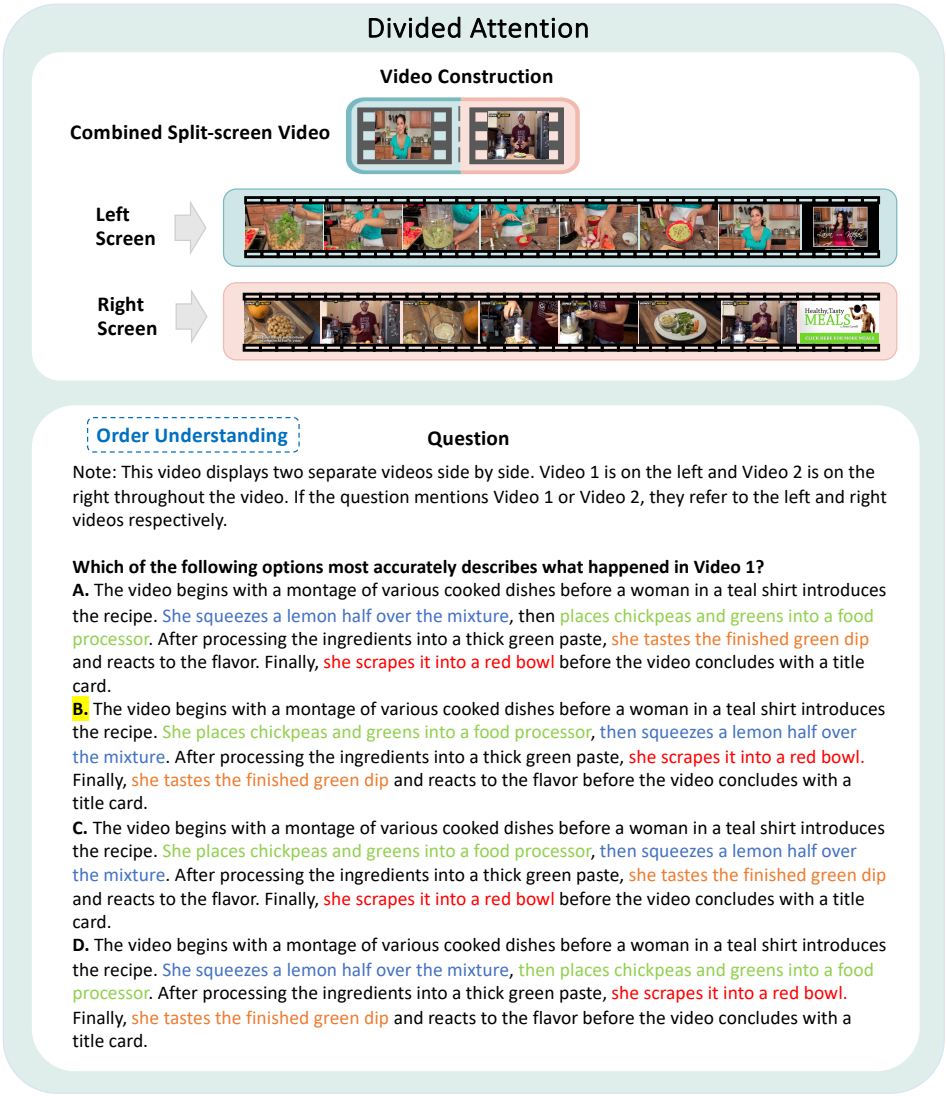

Figure 14: Example of Divided Attention targeting *Order Understanding*. The three distractors swap the temporal or logical sequence of events in the target video's narrative, while the correct option (highlighted in yellow) preserves the original order.

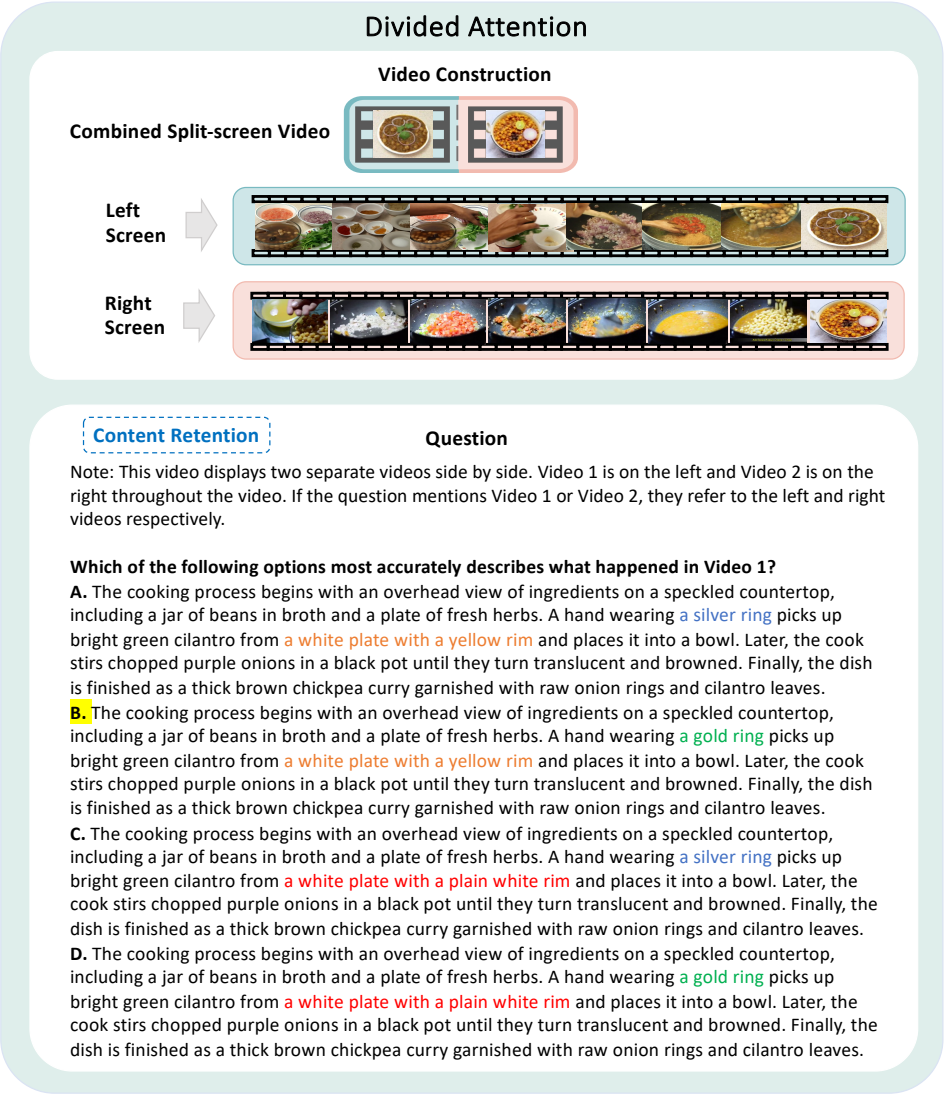

Figure 15: Example of Divided Attention targeting *Content Retention*. The three distractors replace certain content in the target video's narrative with plausible but fabricated content, while the correct option (highlighted in yellow) faithfully describes only the target video.

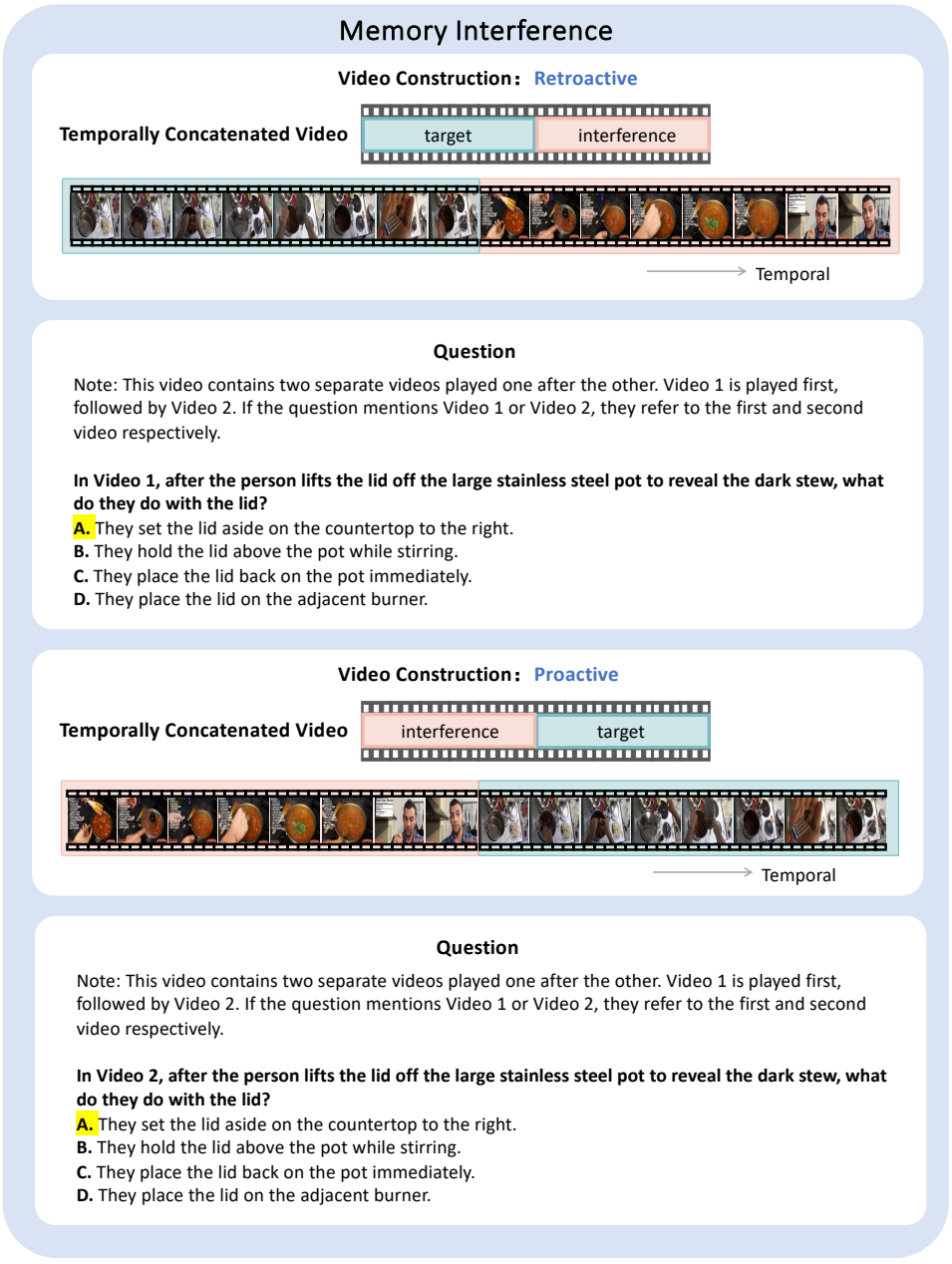

Figure 16: Example of Memory Interference. Each question comprises the correct answer (highlighted in yellow) for the target video, two intrusion options drawn from the interfering video, and one unrelated distractor.

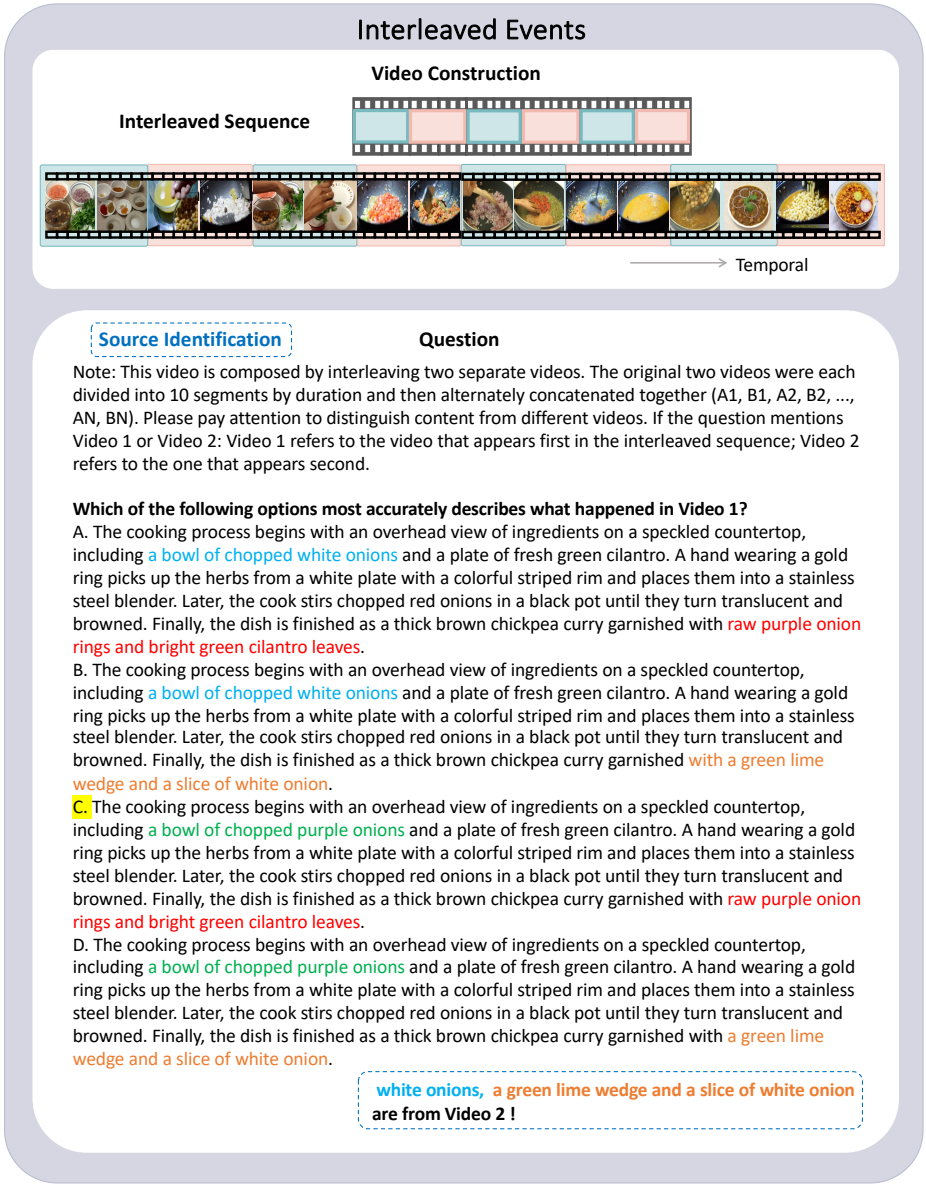

Figure 17: Example of Interleaved Events targeting *Source Identification*. The three distractors replace certain content in the target video's narrative with content from the distractor video, while the correct option (highlighted in yellow) faithfully describes only the target video.

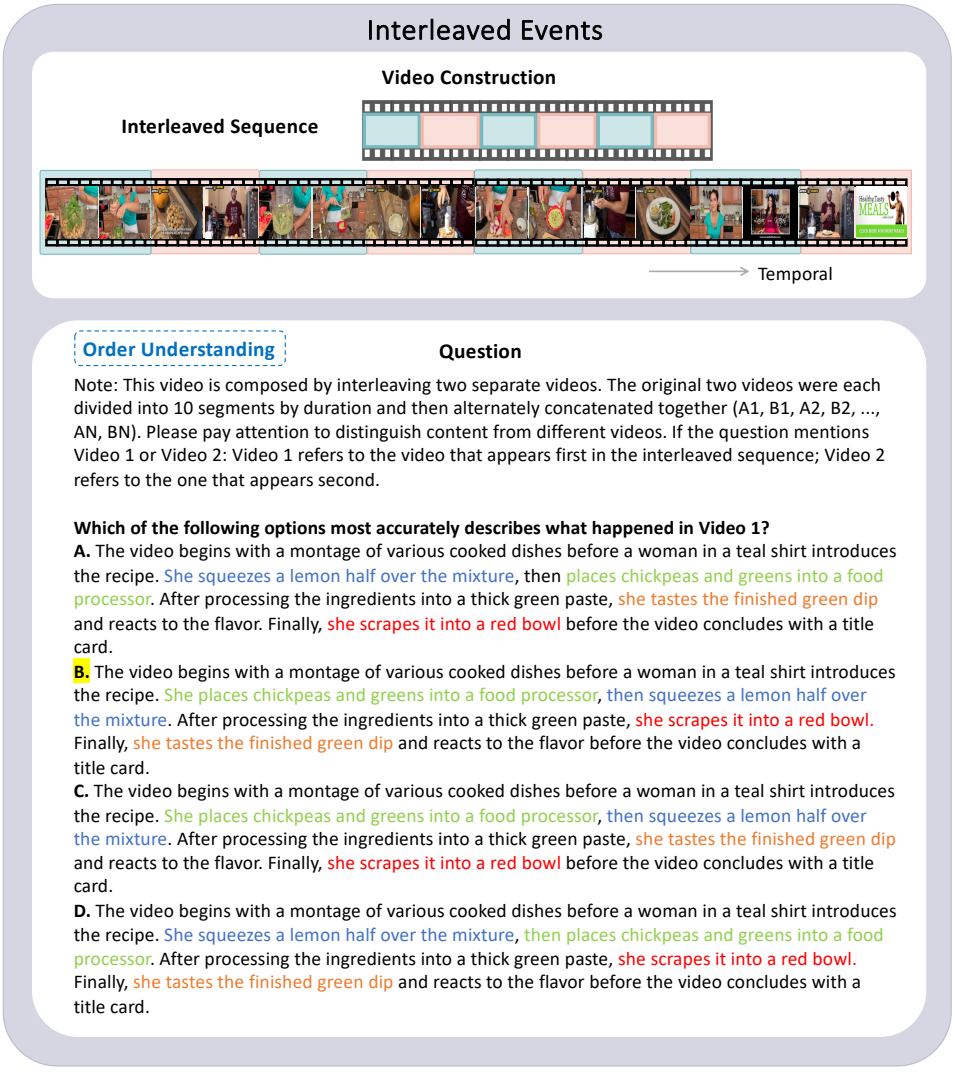

Figure 18: Example of Interleaved Events targeting *Order Understanding*. The three distractors swap the temporal or logical sequence of events in the target video's narrative, while the correct option (highlighted in yellow) preserves the original order.

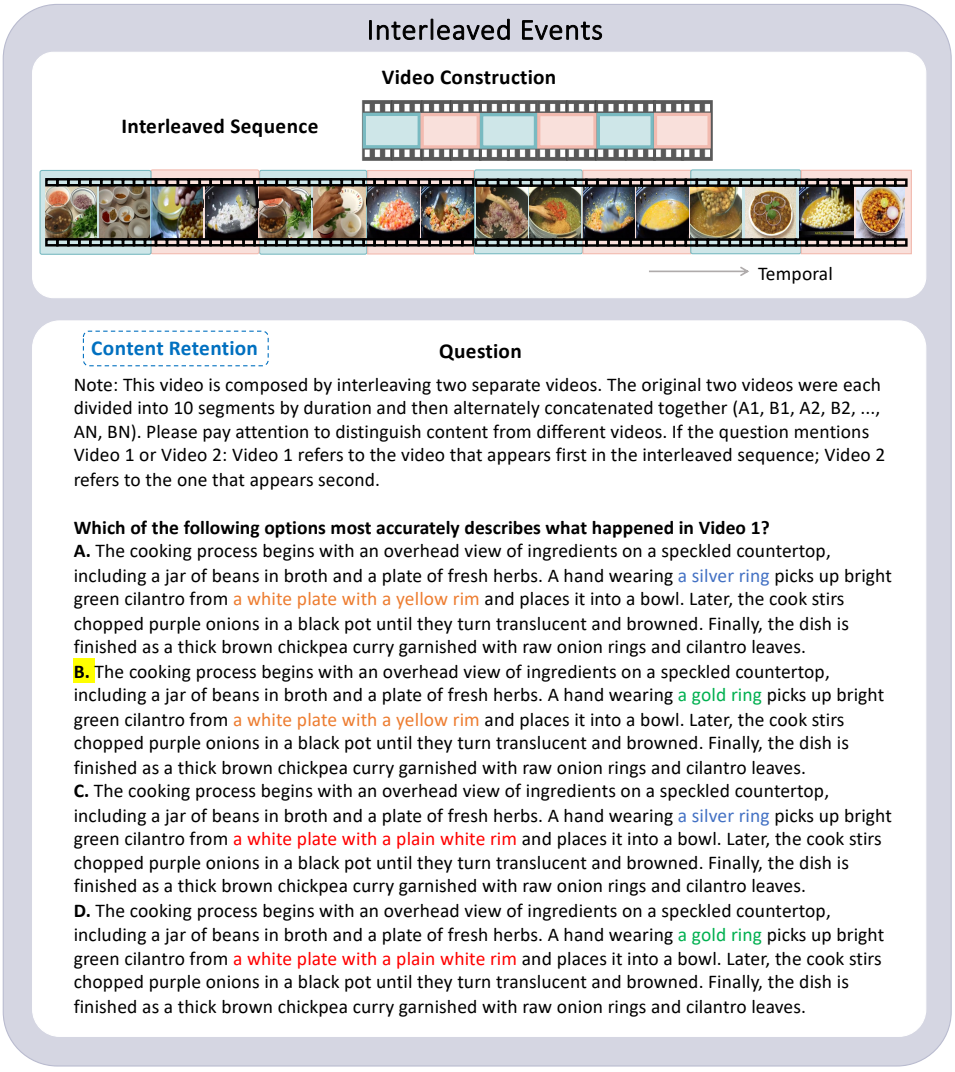

Figure 19: Example of Interleaved Events targeting *Content Retention*. The three distractors replace certain content in the target video's narrative with plausible but fabricated content, while the correct option (highlighted in yellow) faithfully describes only the target video.

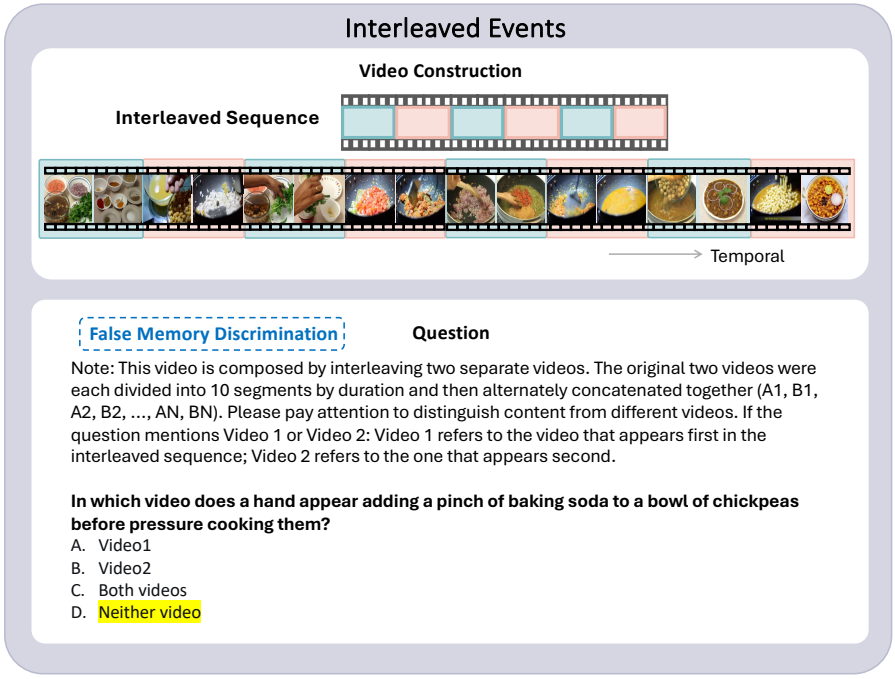

Figure 20: Example of Interleaved Events targeting *False Memory Discrimination*. A fake question that is relevant to video content is presented, and the model should be aware to choose the option indicating that the query does not belong to either video. The correct answer is highlighted in yellow.

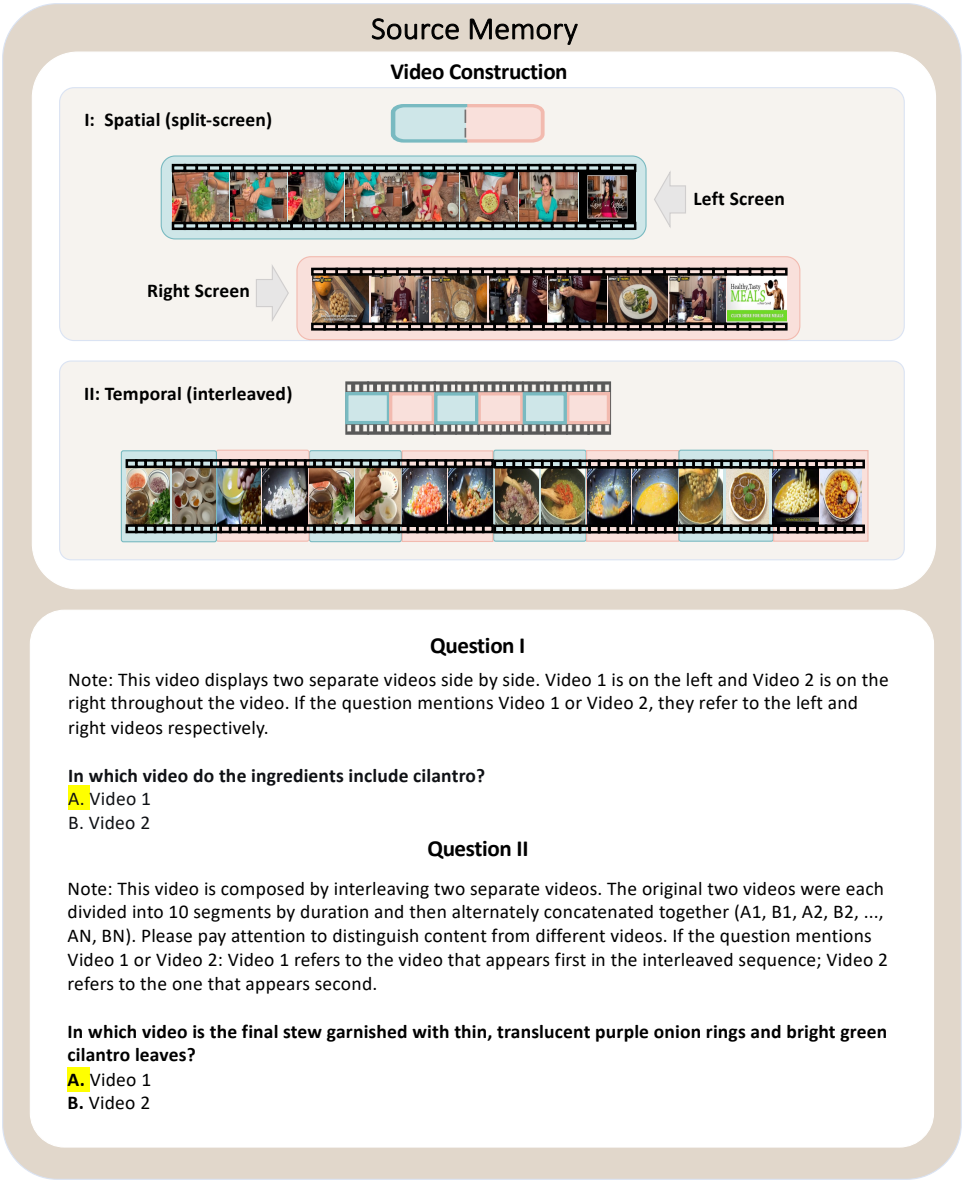

Figure 21: Example of Source Memory. Spatial refers to a split-screen format with frequent left/right swaps. The correct answer is highlighted in yellow.

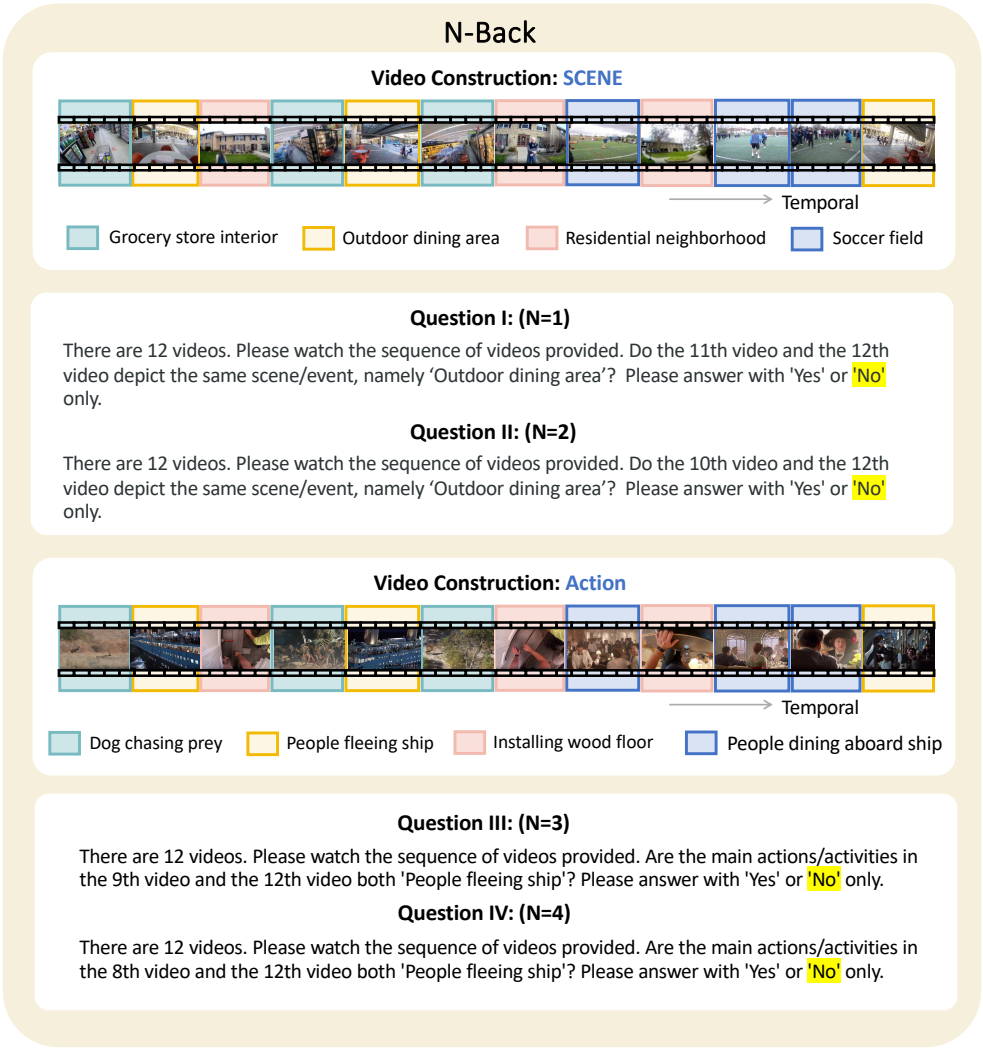

Figure 22: Example of N-Back. The model is asked to decide whether the final clip matches the clip N positions earlier with a Yes/No answer, on two attributes: *Scene* (same environment category) and *Action* (same type of activity). The correct answer is highlighted in yellow.