VIDEOMEM-BENCH: BENCHMARKING QUERY-AGNOSTIC MEMORY IN VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

While recent Video Large Language Models (Video-LLMs) show strong performance in processing long context windows, it remains unclear whether they genuinely memorize information or merely retrieve it from a large context cache. Existing benchmarks often overestimate actual memory capabilities because they rely on query-guided joint encoding, allowing models to retrieve video frames conditioned on the question. Furthermore, these benchmarks lack a structured cognitive hierarchy. To address these limitations, we introduce VideoMem-Bench, a benchmark designed to systematically evaluate genuine memory in the field of long video understanding. We establish a four-level hierarchical framework inspired by cognitive neuroscience, ranging from basic semantic perception to predictive causal simulation. Crucially, we propose a query-agnostic, decoupled evaluation paradigm. By enforcing an information bottleneck, this paradigm requires models to compress continuous video streams into a capacity-constrained memory state without prior knowledge of the queries. Models must then answer subsequent questions relying solely on this compressed memory representation. Comprising 3,433 videos and 5,360 annotated QA pairs, VideoMem-Bench prevents retrieval shortcuts and provides a rigorous standard for assessing autonomous long-term memory in dynamic environments.

1 INTRODUCTION

The transition from short-clip pattern recognition to long-video understanding represents a critical milestone for computer vision and artificial general intelligence. Since real-world visual data unfolds as a continuous and dynamic stream, intelligent agents must simultaneously perceive transient actions and integrate spatiotemporal contexts over extended periods. Recently, Video Large Language Models (Video LLMs) (Zhang et al., 2023; Maaz et al., 2024; Cheng et al., 2024) have demonstrated the capacity to process hour-long video sequences, facilitating complex reasoning over temporal dynamics and causal logic. To assess this progress, benchmarks such as EgoSchema (Mangalam et al., 2023) and Video-MME (Fu et al., 2025) formulate long-video understanding as complex question-answering tasks, promoting the expansion of context windows and advancements in multimodal alignment. Consequently, several benchmarks claim to evaluate the memory capabilities required for comprehensive video understanding. However, their evaluation protocols typically allow models to access the raw video alongside the question. This joint processing mechanism introduces a fundamental ambiguity: are these models genuinely encoding and memorizing information, or are they merely performing query-guided retrieval from a fully accessible sensory cache?

To the best of our knowledge, there lacks a benchmark specifically designed to systematically evaluate the memory capability of long-video models. Existing datasets involving memory evaluation exhibit two main limitations. First, they lack a structured hierarchy of memory capabilities. Most are designed for general video understanding rather than specific memory traits. Their question formats are often straightforward, making it difficult to assess higher-order cognitive and memory functions. Second, they heavily rely on joint encoding mechanisms (see Figure 1). Current benchmarks typically allow models to use the query as a condition to retrieve relevant video content. While convenient for offline evaluation, this paradigm diverges from practical applications like streaming video processing (Ren et al., 2024; Yang et al., 2025b; Wang et al., 2025d) and embodied AI (Ding et al., 2026a;b; Li et al., 2026a). In real-world scenarios, agents cannot cache massive raw sensory data indefinitely. They must operate in a query-agnostic manner, continuously filtering

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.

and compressing key information into a capacity-constrained memory state $\mathbf{M}$. The model must then process user queries relying solely on $\mathbf{M}$. Because existing evaluation protocols do not decouple the perception and question-answering stages, they tend to overestimate a model's true memory capability in continuous dynamic environments.

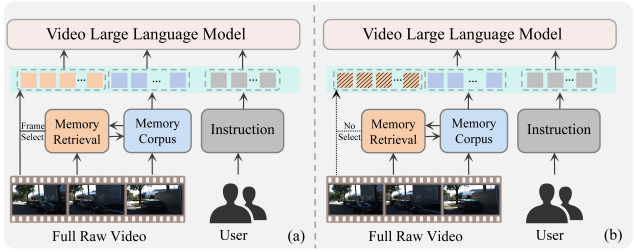


Figure 1: Comparison of evaluation paradigms. (a) Joint encoding enables query-guided retrieval from raw video. (b) Our decoupled paradigm enforces a query-agnostic bottleneck over compressed memory.

To accurately assess the genuine memory capabilities of models in video understanding, we introduce VideoMem-Bench. This benchmark comprises 3,433 long videos and 5,360 manually annotated question-answering pairs across diverse long-form video scenarios. The data covers diverse scenarios, including embodied operations, news broadcasts, instructional demonstrations, cinematic narratives, and various indoor and outdoor activities. Theoretically, inspired by Tulving's taxonomy of consciousness (Tulving, 2002) in cognitive neuroscience, we construct a four-level hierarchical evaluation framework: semantic perception, episodic memory, spatio-temporal scene reconstruction, and predictive causal simulation. This framework quantifies a model's memory performance across dimensions such as object attribute recognition, action state identification, spatial relation reasoning, and causal inference. Methodologically, we decouple the memory evaluation into independent encoding and decoding stages. During the memory encoding stage, the model processes a continuous video stream $\mathcal{V}$ without access to the query set $\mathcal{Q}$, filtering and compressing the visual information into a capacity-constrained memory state $\mathbf{M}$. During the memory retrieval stage, the model relies exclusively on $\mathbf{M}$ to answer any given query $q_i \in \mathcal{Q}$, which is formulated as estimating $P(a \mid \mathbf{M}, q_i)$. By introducing this information bottleneck constraint, our evaluation protocol prevents the model from using the query to retrospectively retrieve raw video frames. This provides an objective measurement of long-term memory capabilities under query-agnostic conditions.

The main contributions of this work are summarized as follows:

- We introduce VideoMem-Bench, the first benchmark specifically designed to evaluate genuine memory capabilities in video understanding systems.
- We establish a hierarchical video memory framework inspired by neuroscience.
- We design a decoupled, query-agnostic paradigm leveraging information bottlenecks to assess compression and reconstruction.

2 Related Work

Benchmarks for Long Video Understanding. Current benchmarks for long video understanding, such as EgoSchema (Mangalam et al., 2023), Video-MME (Fu et al., 2025), LVBench (Wang et al., 2025b), Neptune (Nagrani et al., 2024), CG-Bench (Chen et al., 2025), and MomentSeeker (Yuan et al., 2026), primarily evaluate spatial-temporal integration and long-context retention via complex video QA tasks. Meanwhile, other studies target specific memory mechanisms. For instance, the Ego4D (Grauman et al., 2022) episodic memory task requires locating specific objects in first-person videos. OVBench (Huang et al., 2025) tests continuous perception of dynamic contexts in online streaming, while EgoMem (Zuo et al., 2026) focuses on fine-grained feature extraction and event reconstruction in extremely long videos. However, existing benchmarks predominantly formulate video memory evaluation as flat QA or retrieval tasks, lacking a systematic and hierarchical structure. Consequently, they fail to comprehensively assess long-video memory capabilities, necessitating a multi-dimensional, cognitively-grounded evaluation framework.

Memory Mechanisms in Video Understanding. To address computational and storage bottlenecks in long videos (Jin et al., 2025; Xu et al., 2025), current memory mechanisms generally fall into two categories: internal token compression and external agent-based architectures. Token com-

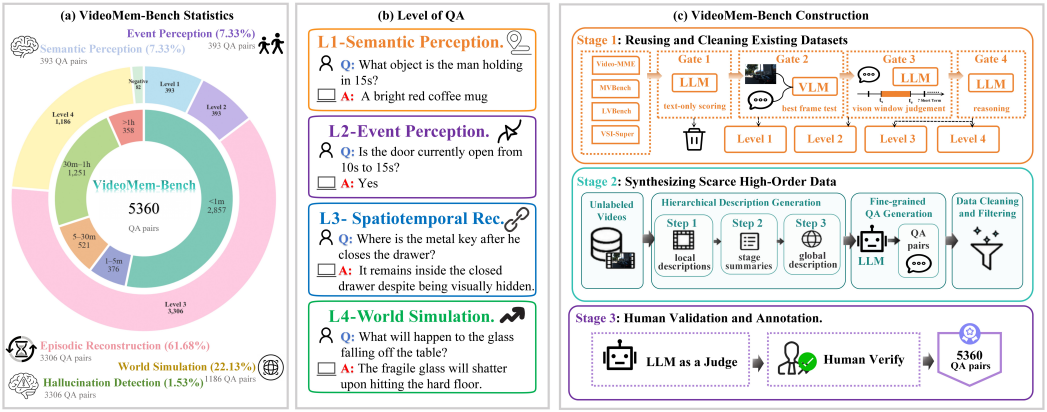


Figure 2: Overview of **VideoMem-Bench**. **(a)** Statistical distribution across cognitive levels and video lengths. **(b)** Examples of the progressive four-level QA taxonomy. **(c)** The three-stage pipeline for benchmark construction, including existing data filtration, scarce data synthesis, and human validation.

pression methods (Liu et al., 2025) focus on optimizing internal feature sequences to build dense memory representations. For example, MovieChat (Song et al., 2024) and MovieChat+ (Song et al., 2025) design long- and short-term memory modules to merge redundant visual tokens, enabling efficient sequence compression. Similarly, BIMBA (Islam et al., 2025) uses feature dimensionality reduction to maintain compact context representations. Conversely, agent-based architectures build explicit external memories and structured retrieval pipelines to assist long-range reasoning. VideoAgent (Fan et al., 2024), VideoLucy (Zuo et al., 2026), and WorldMM (Yeo et al., 2025) construct dynamic multimodal memory frameworks that allow agents to effectively backtrack video events. Works like Seeing, Listening, Remembering, and Reasoning (Long et al., 2025) integrate multimodal perception modules with long-term memory mechanisms. For highly complex tasks or embodied scenarios, Ego-R1 (Tian et al., 2025), and LongVideo-R1 (Feng et al., 2025; Wang et al., 2025a; Li et al., 2025; Jin et al., 2026) introduce external tool invocation and multi-step reasoning chains to manage memory. VideoTree (Wang et al., 2025e) utilizes a hierarchical tree structure to aggregate and retrieve massive video features. These models are still evaluated via query-guided protocols, allowing them to revisit video frames conditioned on the queries. This fails to verify their autonomous memory capabilities when queries are unknown. Consequently, we propose a query-agnostic, decoupled evaluation paradigm to quantify compression and retrieval capabilities under an information bottleneck.

3 The Definition of Video Memory

3.1 Formal Formulation

To contrast existing benchmarks and our proposed genuine memory setting, we formulate the VideoQA task as a probabilistic inference process. Let $\mathcal{V} = \{v_1, \ldots, v_T\}$ denote a continuous video stream, and let q be the user query.

Existing Paradigm (Joint Encoding): Current video understanding paradigms formulate the task as estimating the conditional probability of answer a given the full video V and query q. In this setting, model F acts as a joint encoder:

$$\hat{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} P(a \mid \mathcal{V}, q) \approx \mathcal{F}(\text{Attn}(q, \mathcal{V})) \tag{1}$$

This formulation allows the model to use query q as a cue to retroactively extract specific frames from V. This query dependency bypasses the need to maintain global information, degenerating memory into on-demand retrieval.

Proposed Paradigm (Decoupled Encoding and Decoding): To evaluate genuine memory, we introduce a strict decoupling between perception and interrogation. We formulate video memory as

a two-stage process guided by a latent memory state $\mathbf{M}$: *(1) Query-Agnostic Memorization.* The model must compress the video stream $\mathcal{V}$ into a memory state $\mathbf{M}$ without access to the query q. This enforces conditional independence ($\mathbf{M} \perp q$) between the memory construction process and the query:

$$\mathbf{M} = \mathcal{E}_{\text{enc}}(\mathcal{V}), \quad \text{s.t. } |\mathbf{M}| \ll |\mathcal{V}| \tag{2}$$

where $|\cdot|$ denotes information capacity, such as token count. This constraint introduces an entropy bottleneck, compelling the model to perform lossy compression instead of simple global caching. *(2) Memory-Based Recall.* Answer generation depends solely on the compressed memory $\mathbf{M}$ and the query q, where the original video $\mathcal{V}$ remains inaccessible:

$$\hat{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} P(a \mid \mathbf{M}, q; \phi) \tag{3}$$

Axioms. Under this definition, an effective video memory system must satisfy the following three axioms: **(1) Causal Unidirectionality:** Memory updates must strictly follow chronological order. The model can only construct the memory state $\mathbf{M}_t$ based on current and historical observations, without using global attention to access future frames. **(2) Query Decoupling:** Memory construction must be independent of subsequent queries. The model must autonomously filter and retain key content from the video stream without prior knowledge of the specific task. **(3) Constructive Compression:** Memory is not a simple sampling or caching of original visual frames, but a semantic reconstruction. Under a strict capacity bottleneck ($|\mathbf{M}| \ll |\mathcal{V}|$), the model must distill semantics and object evolution patterns from the video.

3.2 A Cognitive Taxonomy of Video Memory

Existing research lacks a systematic taxonomy for the memory capabilities of video models. Inspired by Tulving's taxonomy of consciousness (Tulving, 2002) in cognitive neuroscience, we categorize the evolution of video memory systems into four progressive cognitive levels. Grounded in psychological concepts, we define video memory as follows:

Level 1: Semantic Perception. This level corresponds to basic visual feature extraction and entity recognition. The model identifies objects, attributes, or basic actions from individual frames without relying on cross-frame temporal information. Here, the memory system acts as a static semantic mapping, primarily processing spatial visual features without temporal awareness.

Level 2: Episodic Event Perception. This level corresponds to recognizing short-term temporal actions and state changes. The model integrates consecutive frames to capture local temporal dynamics, such as opening a door. The memory system operates as a short-term context dependency, often based on a sliding window. It processes basic action continuity but lacks a global spatiotemporal structure independent of visual inputs.

Level 3: Spatiotemporal Episodic Reconstruction. This level corresponds to autonoetic consciousness in psychology, defined as the ability to reconstruct specific experiences without relying on current environmental stimuli. The model overcomes local frame dependency to build global scene state representations in the latent space. The core objective is to achieve object permanence and spatial consistency. Even under occlusion or out of view, the model maintains the existence and location of objects. For instance, it can distinguish a specific key placed in a drawer from the generic concept of a key. This requires the system to integrate discontinuous observations, maintain a spatiotemporally continuous state tracker, and ensure memory representations remain stable despite viewpoint shifts or short-term information loss.

Level 4: Predictive World Simulation. This level corresponds to prospective memory in cognitive science, representing an advanced memory form for autonomous agents. Instead of statically maintaining historical states, the model constructs a dynamically evolving world model. It utilizes historical memory states to deduce future states, enabling counterfactual reasoning. For example, upon observing a glass falling, the model predicts it will break based on physical priors. If subsequent visual inputs contradict this prediction, the model detects a causal anomaly. At this stage, memory empowers the model to perform causal inference in unobserved scenarios, facilitating a deeper understanding of the physical world.

Table 1: Overview of the Cognitive Taxonomy in VideoMem-Bench.

Cognitive Level	Core Focus & Included Tasks
Level 1: Semantic Perception *(Cognitive Basis: Sensory Memory)*	**Focus:** Static feature binding and pre-semantic processing from individual frames without temporal awareness. **Tasks:** Object & Attribute Recognition; Video OCR / Text Spotting.
Level 2: Episodic Event Perception *(Cognitive Basis: Working Memory)*	**Focus:** Capturing local temporal dynamics and action continuity within a short-term sliding window. **Tasks:** Action & State Recognition; Speed & Trajectory; Cross-Modal Associative Memory.
Level 3: Spatiotemporal Episodic Reconstruction *(Cognitive Basis: Episodic Memory)*	**Focus:** Building a global spatiotemporal index and maintaining object permanence independent of current visual stimuli. **Dimensions & Tasks: Entity Constancy** (Counting, State Tracking); Information Granularity (Scene Synopsis, Visual Needle-in-a-Haystack (Lu et al., 2025; Li et al., 2026b)); Temporal Dimension (Action Sequencing, Duration Estimation); Spatial Dimension (Dynamic Spatial Relationships, Egocentric Navigation, 3D Layout Inference).
Level 4: Predictive World Simulation *(Cognitive Basis: Prospective Memory)*	**Focus:** Constructing a dynamically evolving world model for causal inference and counterfactual reasoning based on priors. **Dimensions & Tasks:** Physical Simulation (Explanatory Reasoning, Physical Commonsense, Feasibility); Social Simulation (Intent Tracking, Social Norms, Pragmatic Deception Detection); Spatiotemporal Deduction (Counterfactuals, Hypothetical Scenarios, Anticipation).
Other: Reality Monitoring	**Tasks:** Negative Memory & Hallucination Detection (testing the ability to reject nonexistent events).

4 VIDEOMEM-BENCH

4.1 TASK DESIGN AND TAXONOMY

To evaluate long-term memory, we introduce a multi-dimensional QA system based on our four-level cognitive framework. Inspired by Tulving's taxonomy (Tulving, 2002), the tasks progress from basic semantic and event perception to advanced spatiotemporal reconstruction and predictive simulation (Figure 2b). We also incorporate negative samples for hallucination detection (Gao et al., 2025) to probe model boundaries. Supported by diverse video lengths and cognitive levels (Figure 2a), this progressive design systematically benchmarks model capabilities across different memory stages. Table 1 details the complete taxonomy.

4.2 DATA COLLECTION AND ANNOTATION

We adopt a hybrid construction strategy that combines curating open-source data and synthesizing scarce high-order data. The data pipeline consists of three core stages: reusing and cleaning existing datasets, generating high-order spatiotemporal and causal data, and conducting human validation.

4.2.1 REUSING AND CLEANING EXISTING DATASETS

We aggregate mainstream video understanding benchmarks covering diverse video lengths from 5 seconds to 50 hours. These include Video-MME (Fu et al., 2025), MVBench (Li et al., 2024), LongVideoBench (Wu et al., 2024) and VSI-Super (Yang et al., 2025a). Because existing datasets vary in difficulty and their reliance on genuine memory, they often contain pseudo-memory questions answerable using only a single frame or text-based commonsense. To address this, we design a four-stage gating and routing mechanism to automatically filter and categorize the raw data.

Gate 1: Static Knowledge Separation. This stage discards pseudo-visual samples answerable using only the internal parametric knowledge of language models. We use a language model without visual inputs to compute the normalized conditional log-probability for each candidate option o_k:

$$S_{\text{text}}(o_k) = \frac{1}{|o_k|} \log P_\theta(o_k \mid q) \tag{4}$$

The option with the highest log-probability is selected as the prediction $\hat{a}_{\text{text}} = \text{argmax}_k \, S_{\text{text}}(o_k)$. If the prediction matches the ground truth a_{gt} and its score exceeds the second-best option by a margin δ (i.e., $(\hat{a}_{\text{text}} == a_{\text{gt}}) \wedge (S_{\text{text}}(\hat{a}_{\text{text}}) - S_{\text{text}}(a_{\text{second}}) > \delta)$), the sample is deemed text-dependent and directly discarded.

Table 2: Comparison between VideoMem-Bench and mainstream video understanding benchmarks.

Benchmark	Videos	QAs	Decoupled Eval.	Multi-Level Cognitive Assessment				Hallucination Det.
				L1 Semantic	L2 Episodic	L3 Recon.	L4 Sim.	
Ego4D (Episodic) (Grauman et al., 2022)	Massive	-	×	✓	✓	✓	×	×
EgoSchema (Mangalam et al., 2023)	5,031	5,031	×	✓	✓	✓	×	×
MVBench (Li et al., 2024)	3,655	4,000	×	✓	✓	✓	✓	×
VideoMME (Fu et al., 2025)	900	2,700	×	✓	✓	✓	×	×
LongVideoBench (Wu et al., 2024)	3,763	6,678	×	✓	✓	×	×	×
VSI-Super (Yang et al., 2025a)	700	700	×	✓	✓	✓	×	×
LVBench (Wang et al., 2025b)	103	1,549	×	✓	✓	✓	×	×
OVBench (Huang et al., 2025)	1,463	7,090	×	✓	✓	✓	✓	✓
EgoMem (Zuo et al., 2026)	42	504	×	✓	✓	✓	×	×
VideoMem-Bench (Ours)	**3,433**	**5,360**	✓	✓	✓	✓	✓	✓

Gate 2: Sensory Buffer Separation. For samples passing the first gate, we introduce an oracle frame validation to identify questions relying solely on momentary perception. We compute the cosine similarity between the CLIP (Radford et al., 2021) embeddings of the question E_q and all video frames $\{E_{f_t}\}$ to select the most relevant single frame f^*. We then feed f^*, the question q, and the options $\{o_k\}$ into a vision-language model. If $\hat{a}_{\text{img}}$ matches a_{gt}, the question does not require temporal memory and is categorized as Level 1.

Gate 3: Visual Window Determination. After excluding language priors and single-frame dependencies, we leverage the chain-of-thought reasoning (Wei et al., 2022; Xu et al., 2024) of large language models to analyze the logical dependencies of the QA pairs based on text, without reprocessing frames. We input the question q, the ground truth a_{gt}, and a brief video summary into the model to infer the required visual time window. If the visual evidence is confined to a single continuous shot or a short temporal window (typically under 10 seconds) without needing long-range cross-event references, the sample is assigned to Level 2.

Gate 4: Spatiotemporal and Causal Routing. For the remaining long-range dependent samples, we use reasoning paths generated by large video models (Han et al., 2025; Wang et al., 2025c; Lin et al., 2026) to determine their core cognitive attributes. We input video features or detailed text descriptions into the model, prompting it to generate solution steps and summarize the core dependency. If the model correctly predicts a_{gt} and its reasoning relies on cross-frame tracking, entity counting, or temporal sequencing, the sample is routed to Level 3. If the reasoning involves intent inference, physical mechanism judgment, or counterfactual hypotheses, the sample is routed to Level 4. Throughout this cascaded pipeline, data with ambiguous decisions or low confidence scores are compiled into an anomaly list and discarded.

Category Assignment. To ensure reliable cognitive category labels for the filtered QA pairs, we use a cross-validation mechanism combining an independent LLM with human verification. Given only the question stem, candidate options, and ground-truth answer, Qwen-Max-latest assigns each sample to a fine-grained cognitive category (such as explanatory reasoning). These fine-grained categories are then mapped to cognitive levels (Level 0 to Level 4) and compared against the labels from the earlier model gating mechanism. Samples with matching levels are automatically kept. Otherwise, samples with discrepancies are flagged and sent to human experts for final resolution based on multi-source evidence. To maintain objectivity during manual review, the interface displays all relevant fields: the gating-assigned level, LLM-assigned level, confidence score, machine rationale, 23 cognitive labels, visibility check, memory dependency check, logical uniqueness check, and annotator remarks. Annotators use this information to determine the final fine-grained category and cognitive level for each sample. Two annotators independently review each sample in this phase; any sample with inter-annotator disagreement is discarded. Through this validation pipeline, we build a high-quality QA dataset with accurate cognitive annotations from existing datasets.

4.2.2 Synthesizing Scarce High-Order Data

After data cleaning, we observed a scarcity of Level 3 and Level 4 QA pairs in existing datasets, along with temporal distribution gaps. To address this, we use unlabeled videos to fill the temporal gaps and design a top-down data generation pipeline.

Hierarchical Description Generation. To provide a high-resolution text foundation for subsequent QA generation, we sample videos at one frame per second and construct a three-tier description system. (1) Level 1 Frame and Clip Descriptions: The system generates local detail descriptions

Table 3: Comparative evaluation of baseline models on VideoMem-Bench. Accuracy (%) is reported across four cognitive levels and negative hallucination queries.

Model	L1	L2	L3	L4	Neg.	Overall
Token Compression & Memory Bank						
MovieChat+ (Song et al., 2025)	35.11	32.06	30.49	31.96	6.09	30.90
MovieChat (Song et al., 2024)	33.84	32.82	30.55	31.96	7.31	30.91
BIMBA (Islam et al., 2025)	54.45	81.17	53.57	64.67	68.29	58.34
Memory Agents						
Ego-R1 (Tian et al., 2025)	30.28	33.84	23.93	21.67	15.85	24.50
M3-Agent (Long et al., 2025)	28.50	33.33	33.76	40.13	43.90	34.91
WorldMM (Yeo et al., 2025)	16.54	37.66	30.76	37.94	90.24	32.72
VideoAgent (Fan et al., 2024)	26.72	53.44	30.10	33.64	62.20	32.84
LongVideo-R1 (Feng et al., 2025)	17.56	42.49	22.47	24.28	50.00	24.40
VideoTree (Wang et al., 2025e)	12.72	40.46	16.12	24.62	42.68	19.94
VideoLucy (Zuo et al., 2026)	30.03	63.36	38.60	45.95	64.63	41.81
DVD (Zhang et al., 2026)	30.28	64.63	40.38	42.16	50.00	41.96

every 10 seconds. Using current frames and historical context, it records micro-changes in actions, objects, and scenes. (2) Level 2 Stage Summaries: The system generates stage summaries every 30 seconds by aggregating the recent three Level 1 descriptions and the previous Level 2 summary, compressing dynamic event progressions into a rolling context memory. (3) Level 3 Global Details: Upon video completion, the system integrates the latest Level 2 summary and remaining clips to output a comprehensive description of the entire video.

Fine-Grained QA Generation. After obtaining the global description, we input predefined Level 3 categories (e.g., cross-frame counting) and Level 4 categories (e.g., physical feasibility, social intent) into a large language model (Han et al., 2025; Zhang et al., 2025). We construct prompts containing task definitions, three hard negative examples, and the global video description to guide the model in generating category-specific multiple-choice questions. We then use Sentence-Transformers (Reimers & Gurevych, 2019) to extract features and remove semantic duplicates. Additionally, we use regular expressions to strictly filter out uninformative options containing phrases like "not explicitly stated" or "not shown in the video" to ensure data quality.

Human Validation and Annotation. For the candidate datasets from the automated pipeline, we introduce a final human validation stage performed by professionals trained in cognitive logic using visual annotation tools. Annotators verify each QA pair against three strict criteria; samples are included in the final dataset only if all criteria are met. Each question is independently annotated by two individuals. Any data with inconsistent annotations is discarded to ensure accuracy.

4.3 Dataset Statistics

VideoMem-Bench contains 5,360 QA pairs. By cognitive level, the distribution is: Level 1 and 2 (7.3% each), Level 3 (61.7%), Level 4 (22.1%), and Negatives (1.5%). Over 83% of samples belong to Levels 3 and 4, emphasizing higher-order memory capabilities like long-term dependencies and causal reasoning over basic recognition. Video lengths span from short clips to over an hour: $<$1 min (53.3%), 1 to 5 min (7.0%), 5 to 30 min (9.7%), 30 to 60 min (23.3%), and $>$1 hour (6.7%). This distribution supports analysis of memory degradation and retention in video understanding.

4.4 Comparison with Existing Benchmarks

Table 2 compares VideoMem-Bench with existing long video understanding benchmarks. Existing benchmarks (e.g., EgoSchema (Mangalam et al., 2023)) typically employ joint vision-language encoding, allowing models to retrieve video content conditioned on the queries. In contrast, our benchmark adopts a query-agnostic, decoupled evaluation paradigm to rigorously assess actual information compression and memory retrieval. Furthermore, while prior works largely focus on shallow perception, our benchmark evaluates higher-order predictive cognition. Finally, we introduce negative samples to detect hallucinations caused by memory decay, providing an objective standard for assessing the long-term memory capabilities of multimodal models.

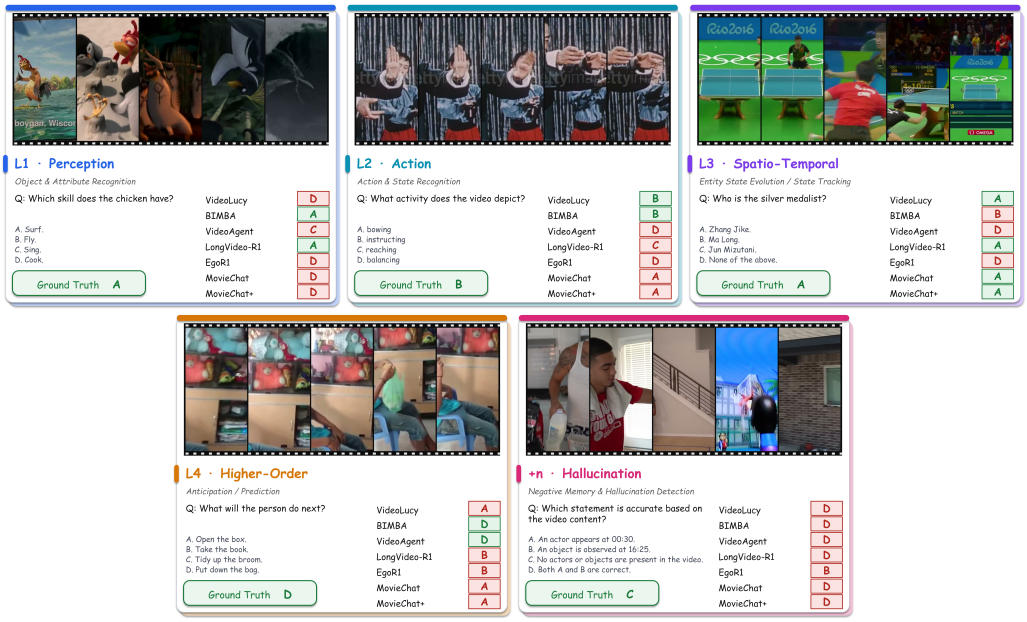


Figure 3: Qualitative comparison of representative baselines on VideoMem-Bench. Each row corresponds to one cognitive level (L1–L4) or a negative-memory probe (+n).

5 Experiments

5.1 Baselines and Adaptation

Baselines. We evaluate two representative categories of baselines: *(1) Memory Agents* (Ego-R1 (Tian et al., 2025), Seeing, Listening, Remembering, and Reasoning (M3-Agent) (Long et al., 2025), WorldMM (Yeo et al., 2025), VideoAgent (Fan et al., 2024), LongVideo-R1 (Feng et al., 2025), VideoTree (Wang et al., 2025e), DVD (Zhang et al., 2026), and VideoLucy (Zuo et al., 2026)). To enforce our decoupled paradigm, we prohibit these agents from accessing original videos or invoking external VLMs for frame retrieval during the recall phase. *(2) Token Compression and Memory Banks* (MovieChat+ (Song et al., 2025), MovieChat (Song et al., 2024), BIMBA (Islam et al., 2025)). These methods encode long videos via selective scanning, cross-modal reduction, or query-aware sparse memory construction.

Standardized Setup. To ensure fair comparisons, we use Qwen2.5-VL-7b (Bai et al., 2025) as the base model for all Memory Agent baselines during both memory construction and inference. When retrieval models are required, we consistently use VLM2Vec V2. We retain the original internal logic, memory update mechanisms, and default hyperparameters of each method without introducing any task-specific tuning. Models used for dataset construction are strictly isolated from those used for evaluation. Detailed adaptation protocols are provided in Appendix A. For Memory Bank models, replacing the base model degrades performance, so we adopt their original configurations. Specifically, MovieChat and MovieChat+ use EVA-CLIP ViT-G/14 with Vicuna-7B, and BIMBA uses CLIP ViT-L/14 with Vicuna-7B.

5.2 Comparative Analysis

We conduct a decoupled evaluation of the aforementioned baselines on the full VideoMem-Bench test set, with results shown in Table 3.

Cognitive Level Analysis. Based on our four-level cognitive taxonomy, we analyze the memory capabilities of the baselines. As shown in Table 3, most models achieve their peak performance at Level 2 (episodic event perception), suggesting that current architectures are relatively adept at localized event tracking and feature matching. However, a consistent performance drop occurs at

Table 4: Resource overhead and memory capacity per query for each model. Tokens include both textual and visual tokens. VRAM and RAM are reported in GB, and time is reported as average inference seconds per query. ↓ indicates lower resource consumption or inference cost.

Model	Tokens (Avg./Q)↓	Time (s/Q)↓	Peak VRAM↓	Avg. VRAM↓	Peak RAM↓	Avg. RAM↓
Token Compression & Memory Bank						
MovieChat+	166.30	0.306	21.14	16.36	3.73	2.64
MovieChat	166.30	0.292	30.78	16.16	3.51	3.18
BIMBA	118.87	0.783	30.30	30.15	2.52	2.48
Memory Agents						
Ego-R1	8,150.97	4.667	28.76	28.67	17.57	13.99
M3-Agent	2,329.71	0.818	48.75	48.31	19.51	14.62
WorldMM	27,502.74	15.039	64.01	26.75	8.16	4.14
VideoAgent	2,096.00	31.012	27.13	27.11	17.00	15.51
LongVideo-R1	121,759.00	39.222	23.59	23.03	15.80	15.79
VideoTree	435.57	1.421	16.47	16.46	2.34	2.33
VideoLucy	2,827.00	13.911	9.50	15.70	5.90	2.40
DVD	475.53	8.362	23.41	21.53	4.60	1.25

Level 3 (global spatiotemporal reconstruction). For instance, LongVideo-R1 drops from 42.49% on L2 to 22.47% on L3, and VideoTree drops from 40.46% to 16.12%. While some models exhibit a slight performance recovery at Level 4 (predictive world simulation), this likely relies on the internal reasoning priors of their LLMs rather than visual memory. Ultimately, the overall degradation on L3 and L4 demonstrates that existing models still struggle to construct coherent long-term spatiotemporal dependencies and causal logic in their latent spaces.

Negative Memory and Hallucination Analysis. Robustness to nonexistent visual content is important for evaluating memory mechanisms (Gao et al., 2025). As shown in Table 3, performance on negative queries varies widely. MovieChat and MovieChat+ achieve only 7.31% and 6.09%, suggesting strong memory hallucination when no video frames are available to verify false textual premises. In contrast, Ego-R1 and BIMBA are more robust, while VideoAgent, LongVideo-R1, and DVD show intermediate performance. Overall, without visual anchors, weaker models rely more on language priors and are more likely to produce incorrect but premise-consistent answers.

Qualitative Analysis. Figure 3 complements Table 3 with examples spanning the L1 to L4 cognitive hierarchy and a negative memory probe (+n). Among the baselines shown, no single model dominates. Success varies across L1 perception, L2 action, and L3 spatiotemporal tracking. Only BIMBA and VideoAgent succeed at L4, highlighting the difficulty of predictive reasoning. Notably, all models fail the +n probe. These cases show that overall accuracy obscures distinct failure modes across cognitive levels and a shared susceptibility to hallucination.

Efficiency and Memory Footprint Analysis. To separate memory quality from memory size, we report the average tokens per query and inference overhead of all baselines, including time and peak/average VRAM and RAM, in Table 4. Average token usage combines textual and visual tokens as a unified measure of memory capacity. Token-compression methods use fewer than 170 tokens per query, while memory agents range from 435.57 tokens for VideoTree to 121,759.00 for LongVideo-R1. Comparing accuracy with token usage therefore helps assess whether a model extracts useful information efficiently or mainly depends on a larger memory buffer.

6 Conclusion

We introduce VideoMem-Bench, a benchmark that decouples video encoding from question answering to evaluate internalized video memory. With a four-level cognitive taxonomy, VideoMem-Bench measures memory across semantic perception, episodic recall, long-term spatiotemporal reasoning, and causal prediction. Our results show that current multimodal models degrade on high-level tracking and causal questions. They also remain vulnerable to false textual premises when direct visual evidence is unavailable, leading to memory hallucinations. These findings suggest that long-video understanding cannot be solved by simply extending context length. Future work should develop memory architectures with spatiotemporal indexing, causal simulation, and reality monitoring.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Jiayu Ding, Xinpeng Liu, Zhiyi Pan, Shiqiang Long, and Ge Li. Extrinsplat: Decoupling geometry and semantics for open-vocabulary understanding in 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31019–31028, 2026a.

Jiayu Ding, Haoran Tang, Hongbo Jin, Wei Gao, and Ge Li. 3d instruction ambiguity detection. *arXiv preprint arXiv:2601.05991*, 2026b.

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.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025.

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 Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26181–26191, 2025.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3328–3338, 2025.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29096–29107, 2025.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *arXiv preprint arXiv:2512.04540*, 2025.

Hongbo Jin, Rongpeng Zhu, Zhongjing Du, Xu Jiang, Jingqi Tian, Qiaoman Zhang, and Jiayu Ding. Dgpo: Distribution guided policy optimization for fine grained credit assignment. *arXiv preprint arXiv:2605.03327*, 2026.

Chaoyue Li, Yongxue Xu, Jie Feng, and Jiayu Ding. Lmm-track4d: Eliciting 4d dynamic reasoning in lmms via trajectory-grounded dialogue. *arXiv preprint arXiv:2605.19390*, 2026a.

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.

Miaoyu Li, Qin Chao, and Boyang Li. Two causally related needles in a video haystack. *Advances in Neural Information Processing Systems*, 38, 2026b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Hao Lin, Kunyang Lv, Xu Jiang, Jingqi Tian, Zhongjing Du, Jiayu Ding, Qiaoman Zhang, and Hongbo Jin. Visd: Enhancing video reasoning via structured self-distillation. *arXiv preprint arXiv:2605.06094*, 2026.

Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. Pave: Patching and adapting video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3306–3317, 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.

Haoyu Lu, Yuqi Huo, Yifan Du, Zijia Zhao, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms. In *International Conference on Learning Representations*, volume 2025, pp. 99750–99782, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Jaap MJ Murre and Joeri Dros. Replication and analysis of ebbinghaus' forgetting curve. *PloS one*, 10(7):e0120644, 2015.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, et al. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

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*, 2025.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

Endel Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25, 2002.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025a.

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, 2025b.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, William Wang, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey, 2025c.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24170–24181, 2025d.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3272–3283, 2025e.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024.

Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding, 2025.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2025a.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. *arXiv preprint arXiv:2512.02425*, 2025.

Huaying Yuan, Jian Ni, Zheng Liu, Yueze Wang, Junjie Zhou, Zhengyang Liang, Bo Zhao, Zhao Cao, Ji-Rong Wen, and Zhicheng Dou. Momentseeker: A task-oriented benchmark for long-video moment retrieval. *Advances in Neural Information Processing Systems*, 38, 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49.

Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. *arXiv preprint arXiv:2506.08817*, 2025.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026.

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. *Advances in Neural Information Processing Systems*, 38:25660–25691, 2026.

A Extended Implementation Details

In this section, we provide extended details regarding our dataset construction pipeline, rigorous quality assurance protocols, decontamination procedures, and the faithful adaptation of baseline models.

A.1 Detailed Model Specifications and Versioning

To ensure reproducibility and clarify the automated components of our pipeline, we detail the specific models and versions invoked at each dataset construction stage. We deliberately utilized state-of-the-art models to maintain high filtering and synthesis quality: **Gate 1 (Static Knowledge Separation):** We utilized `Llama-3-70B-Instruct` as the language-only baseline to compute text-conditional log-probabilities and filter out language-prior shortcuts.

Gate 2 (Sensory Buffer Separation): We employed `CLIP-ViT-L/14` to retrieve the most relevant frame, followed by `Qwen2-VL-72B-Instruct` to verify if the question could be answered from this static frame alone.

Gates 3 and 4 (Visual Window Determination and Routing): We used `GPT-4o` (version `gpt-4o-2024-05-13`) to generate reasoning paths and execute logical dependency analysis for temporal routing.

Category Assignment Validation: We employed `Qwen-Max` (version `qwen-max-0428`) as the independent evaluator to assign fine-grained cognitive labels.

Hierarchical Description and QA Generation: For synthesizing high-order data, we utilized `Gemini-1.5-Pro` due to its extensive context window capabilities, enabling the integration of rolling summaries and the generation of accurate multiple-choice questions based on the global video context.

A.2 Quality Assurance and Human Validation Protocols

While our pipeline leverages advanced models for scalable generation and initial routing, we implemented strict human-in-the-loop protocols to audit evidence correctness, mitigate hallucinations, and resolve boundary ambiguities.

Auditing Rationale Correctness and Synthetic Evidence. A correct multiple-choice prediction alone is insufficient to guarantee data quality. During human validation, annotators explicitly evaluated the logical validity of the model-generated reasoning paths. If a model predicted the correct answer but relied on hallucinated evidence or flawed logic, the sample was immediately discarded. Furthermore, for synthesized data, annotators cross-referenced the model-generated global text descriptions with the video playback. This ensures that the underlying evidence used for QA generation accurately reflected genuine visual events and was free from text-hallucinated artifacts.

Guardrails Against Spurious Shortcuts. Beyond standard regular expression filtering for uninformative options, we implemented an adversarial baseline check. We fed the generated questions and candidate options into a text-only LLM without any video context. If the model exhibited a disproportionately high probability for the correct answer, indicating a linguistic artifact in the option design rather than a necessity for visual reasoning, the sample was removed.

Disambiguating Level 3 and Level 4 Boundaries. To systematically resolve ambiguities between spatiotemporal reconstruction (Level 3) and predictive simulation (Level 4), annotators adhered to an observability criterion. Level 3 strictly evaluates observable state changes over time (e.g., "What sequence of actions occurred?"). Conversely, Level 4 requires inferring unobservable dynamics, such as social intents, physical mechanisms, or counterfactuals (e.g., "If the person had not intervened, what would have happened?"). Samples mixing these characteristics were resolved by expert consensus or discarded.

A.3 Robustness Validation of Automated Cognitive Leveling

Our dataset construction relies on a language model pipeline for cognitive level assignment and reasoning generation. To validate its reliability, we randomly sampled 400 automatically categorized questions for independent human verification. Crucially, experts evaluated both the accuracy of the assigned cognitive category and the factual correctness of the generated reasoning against the visual evidence.

Table 5: Confusion matrix on 400 random samples, with errors occurring only between adjacent cognitive levels.

Predicted ↓	L1	L2	L3	L4	Total
Level 1	**28**	2	0	0	30
Level 2	0	**28**	2	0	30
Level 3	0	4	**240**	6	250
Level 4	0	0	6	**84**	90

Out of the 400 samples, the automated pipeline correctly categorized and provided valid evidence for 380 cases, achieving a 95.0% agreement rate with expert human judgments. As detailed in Table 5, the 20 misclassified cases (5.0%) occurred exclusively between adjacent categories. Specifically, discrepancies arose between Level 2 and Level 3, and between Level 3 and Level 4, where the boundaries for temporal window tracking and causal inference are inherently subjective. Severe errors crossing distant cognitive stages (e.g., Level 1 vs. Level 4) are completely absent. This error distribution confirms that our automated process introduces minimal hierarchical bias and faithfully reflects visual evidence, ensuring the structural validity of VideoMem-Bench.

A.4 Decontamination and Leakage Prevention

The credibility of our benchmark relies on preventing data contamination, ensuring that the test samples were not exposed during the pre-training or fine-tuning phases of the evaluated baseline models. We executed a rigorous decontamination audit comprising the following steps:

Step 1: URL Hashing and Source Filtering. We extracted and hashed the source URLs for all videos. We cross-referenced these hashes against mainstream large-scale pre-training datasets (e.g., WebVid, HowTo100M). Identified overlaps were permanently excluded.

Step 2: Textual Overlap Removal. To prevent language contamination, we extracted the 13-grams of all generated QA pairs and matched them against major open-source text corpora (e.g., C4, CommonCrawl subsets). Exact string matches were stripped from the dataset.

Step 3: Temporal Dynamics Guarantee. The synthesized high-order questions (Levels 3 and 4) isolate specific chronological intersections and fine-grained state changes. By design, these localized temporal dynamics cannot be inferred from static priors acquired during LLM pre-training, ensuring the tasks strictly evaluate active episodic memory rather than parametric knowledge.

A.5 Faithfulness in Baseline Adaptation

To ensure fair comparisons while enforcing our decoupled evaluation protocol, we standardized the base model of all Memory Agent baselines to Qwen2.5-VL-7B for both memory generation and inference, while Memory Bank methods retained their original backbones. We detail how we maintained the architectural integrity of each method to prevent unexpected behavior or information leakage.

Base Model Substitution without Logic Alteration. For all evaluated models, we strictly replaced the underlying vision-language model with Qwen2.5-VL. We implemented this by wrapping the Qwen2.5-VL API to match the input and output formats expected by each baseline's official codebase. We did not alter any internal reasoning logic, temporal chunking strategies, or memory update mechanisms. All baselines utilized their default hyperparameters as specified in their original publications (e.g., node limits in VideoTree, summary frequency in M3-Agent). We performed no custom engineering or task-specific tuning.

Faithful Memory Construction for Agents. Memory agents typically process videos through iterative retrieval and captioning. In our decoupled framework, their memory construction phase operates entirely query-agnostic. We executed their original scripts to encode the video into structured textual memory before exposing any questions. We explicitly did not introduce query-aware tuning, task distribution adjustments, or coarse frame previews. The generated captions and event

logs are derived exclusively from the raw video sequence according to the original design of each agent.

Strict Retrieval Disabling During Recall. During the question answering phase, the raw video files are completely inaccessible. For agent baselines that normally invoke frame retrieval tools dynamically based on the query, we safely mocked these API functions (e.g., `get_frame(t)`) to return a disabled state. This intervention forces the agents to rely strictly on the textual memory constructed in the first phase, without breaking their internal reading pathways. We guarantee there are no hidden gradient access scenarios, thumbnail previews, or bypassed leakage channels available during this phase. This ensures that the results genuinely reflect the capacity of the constructed memory limit rather than covert visual retrieval.

B Additional Results

B.1 Human Baseline and Solvability.

We evaluate human performance to verify the solvability of VideoMem-Bench. We sample 380 higher-level questions, including 178 L3, 120 L4, and 82 negative questions. For each question, two annotators watch the video once, answer without notes, and cannot revisit it. We report their average accuracy. This setting follows a limited-memory protocol related to the Ebbinghaus forgetting effect (Murre & Dros, 2015), while using zero-delay testing to avoid measuring long-term human forgetting. As shown in Table 6, humans achieve 93.18% overall accuracy, confirming that the questions are answerable from the video content. The gap to BIMBA at 65.97% and Ego-R1 at 58.95% shows that current models still struggle with long-term video memory and causal reasoning.

Table 6: Human solvability on a 380-question subset (L3, L4, and negatives).

Method	L3	L4	Neg.	Overall
Token Compression & Memory Bank				
MovieChat+	30.34	51.67	6.09	31.84
MovieChat	30.34	51.67	7.31	32.11
BIMBA	53.63	82.64	68.29	65.97
Memory Agents				
Ego-R1	40.45	76.67	73.17	58.95
M3-Agent	33.71	69.17	39.02	46.05
WorldMM	22.47	55.00	41.46	36.84
VideoAgent	35.39	68.33	62.20	51.58
LongVideo-R1	25.84	42.50	50.00	36.32
VideoTree	16.85	54.84	42.68	32.86
VideoLucy	50.56	74.17	64.63	61.05
DVD	43.82	75.27	50.00	53.54
Human	92.74	93.33	93.90	93.18

B.2 Temporal Decay Analysis

To investigate memory retention across different temporal spans, Table 7 reports baseline accuracy across five video duration intervals. The data shows that as the temporal span increases, different architectures exhibit distinct memory degradation patterns:

Typical Memory Decay. Some models experience memory loss in long videos. For example, LongVideo-R1 achieves 26.11% accuracy on videos under 1 minute, but its performance declines to 19.27% on videos over 1 hour. This reflects a context capacity bottleneck that emerges as video length increases. When the video exceeds the model's effective memory window, its long-range spatiotemporal retrieval largely fails.

Performance Rebound in Long Videos. Notably, performance does not strictly decrease for all models. For instance, M3-Agent and VideoTree perform better on extremely long videos (over 30 minutes or 1 hour) than on short videos. We attribute this anomaly to two factors. First, extremely long videos (such as movies or meetings) often have sparse information density, and their questions lean toward macro-level summarization and causal inference (Levels 3 and 4). Second, models with strong LLM backbones (such as Qwen3-32B or GPT-4) can leverage their internal language priors and logical reasoning to offset the lack of fine-grained visual memory in these tasks. However, this shortcut based on language commonsense is not equivalent to true multimodal episodic memory.

C Limitations

Our benchmark has several limitations. First, while multiple-choice questions facilitate standardized evaluation, they cannot assess open-ended generation or interactive recall. Second, despite domain diversity, potential biases in video styles and cultural contexts suggest that benchmark performance

Table 7: Accuracy (%) across different video durations. Models show different levels of performance change as the temporal span increases, reflecting differences in long-term memory capability.

Model	<1 min	1–5 min	5–30 min	30–60 min	>1 hour	Overall
Token Compression & Memory Bank						
MovieChat+ (Song et al., 2025)	33.23	24.33	26.68	29.26	31.01	30.90
MovieChat (Song et al., 2024)	33.26	24.87	26.10	29.50	30.45	30.91
BIMBA (Islam et al., 2025)	62.71	58.51	49.81	55.61	63.97	58.34
Memory Agents						
Ego-R1 (Tian et al., 2025)	31.40	22.52	27.83	13.27	5.87	24.50
M3-Agent (Long et al., 2025)	32.13	24.93	31.67	41.17	50.28	34.91
WorldMM (Yeo et al., 2025)	28.70	24.93	28.41	41.97	46.93	32.72
VideoAgent (Fan et al., 2024)	33.50	21.18	26.30	36.37	36.87	32.84
LongVideo-R1 (Feng et al., 2025)	26.11	23.59	20.35	23.90	19.27	24.40
VideoTree (Wang et al., 2025e)	18.27	9.36	17.66	23.44	35.47	19.94
VideoLucy (Zuo et al., 2026)	40.25	35.39	37.81	44.84	56.15	41.81
DVD (Zhang et al., 2026)	43.66	39.30	36.66	38.77	50.00	41.96

may not perfectly generalize to all real-world applications. Third, the decoupled evaluation protocol deliberately prohibits video playback to isolate memory compression; therefore, it is unsuitable for tasks requiring precise temporal grounding or retrieval. Consequently, VideoMem-Bench should be viewed as a specialized stress test for memory-augmented agents rather than a universal metric for general video understanding.

D Failure Modes

Experiments and manual analysis reveal six primary failure modes for current video memory models evaluated on VideoMem-Bench:

Detail Loss via Memory Compression. During encoding, models often retain global semantic summaries while discarding infrequent but critical local details. Actions like briefly picking up and dropping an object, moving items to peripheral areas, or intermediate steps in an event sequence are easily lost. These errors are prominent in Level 3 tasks such as entity state tracking, cross-clip counting, and event sequencing.

Entity Binding Errors. When videos contain visually similar people, objects, or locations, models often confuse identities and misattribute state changes. For example, a model might remember a red cup placed on a table corner as a white cup, or attribute an early action to a different person appearing later in the video. This indicates a lack of stable object permanence and entity-centric memory representations.

Temporal Misordering. In long videos, models frequently recall key events but fail to maintain their chronological order. For instance, a model might reverse the sequence of opening a door and fetching an item, or confuse the exact occurrence of a repeated action. This suggests that current memory representations act more as unordered semantic sets rather than queryable temporal indices.

Causal Over-inference. In Level 4 tasks, models often rely on language priors rather than visual evidence. If a person picks up a tool, the model may automatically infer an intent to repair or clean, even without supporting visual context. Similarly, for physical predictions, models tend to default to commonsense outcomes, ignoring atypical or occluded events that actually occurred in the video.

Reality Monitoring Failures. When tested on negative queries, models must identify events or visual elements that never appeared. In our decoupled setting, without original video frames as grounding anchors, models are easily misled by false premises in the questions. They often hallucinate plausible but non-existent memories. This indicates a lack of robust reality monitoring mechanisms to distinguish between observed facts, reasonable inferences, and unoccurred events.

Query-Induced Memory Reconstruction. Although we use a decoupled evaluation where memory is built without seeing the questions, the query text still influences how models interpret their stored memory during inference. Given queries with strong suggestive entities or events, models may

reconstruct non-existent details from their compressed memory. This shows that current memory acts as a generative representation dynamically altered by language priors, rather than a static factual record.

E Dataset Composition

To evaluate video memory and cognitive abilities thoroughly, we decompose the four macro cognitive levels into 23 specific sub-task categories. The complete taxonomy and task definitions are in Table 8. Specifically, Levels 1 and 2 focus on basic visual binding and short-term state tracking at the perceptual level. Level 3, a core component of our benchmark, covers complex long-term episodic memory tasks, including entity constancy, information granularity extraction, and spatiotemporal reconstruction. Level 4 explores higher-order predictive capabilities such as physical world simulation, social intent inference, and counterfactual reasoning. Additionally, we include an independent "Reality Monitoring" category to test the model's ability to resist false visual premises and identify hallucinations. This fine-grained division helps researchers pinpoint specific flaws in the memory mechanisms of different models.

F Ethics, AI Use and Reproducibility

Ethics and Data License

VideoMem-Bench is constructed from public academic datasets and additional videos with accessible licenses. For reused datasets, we follow their original license terms and release only annotations, metadata, and evaluation scripts when raw-video redistribution is not permitted. The benchmark is intended for research use only. We will provide source-level license information in the released datasheet.

AI use statement

In this work, we used generative AI tools for dataset construction, including sample gating, cognitive-level assignment, hierarchical video description, and question generation, as specified in Appendix A. We have not used generative AI tools to fabricate experimental results, and other disclosure tasks not listed here are not applicable to this work. All AI-assisted dataset outputs were reviewed under the human validation protocols described in the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

VideoMem-Bench is constructed from public academic datasets and additional videos with accessible licenses. For reused datasets, we follow their original license terms and release only annotations, metadata, and evaluation scripts when raw-video redistribution is not permitted. The benchmark is intended for research use only. We will provide source-level license information in the released datasheet.

Reproducibility statement

We provide extended implementation details in the appendix, including model versions used in dataset construction (Appendix A), quality-assurance and human-validation protocols, decontamination procedures, and baseline adaptation settings (Appendix A.5). Prompts used for gating, relabeling, hierarchical description, and QA generation are included in Appendix G. Dataset composition and the full cognitive taxonomy are also reported in the appendix.

G Prompts

This section summarizes the prompt templates used in VideoMem-Bench. The prompts are designed around two goals. First, we use a set of gating prompts to filter out samples that can be solved

Table 8: Complete cognitive taxonomy and task definitions in VideoMem-Bench.

Task Category	Cognitive Basis & Task Definition
Level 1: Semantic Perception (Cognitive Basis: Sensory Memory / Feature Binding) Focuses on pre-semantic processing, requiring high-fidelity extraction of physical features from single frames.	
Object & Attribute Recognition	Identify specific entities and their visual properties, including color, material, shape, and state.
Video OCR / Text Spotting	Recognize and transcribe textual information appearing in video frames, such as subtitles and signs.
Level 2: Episodic Event Perception (Cognitive Basis: Working Memory / The Sliding Window) Focuses on capturing dynamic micro-evolutions within a limited context window without requiring long-range retrieval.	
Action & State Recognition	Distinguish action categories and track immediate state changes within a short temporal window.
Speed & Trajectory	Quantify physical movement states, including speed and directional trajectories.
Cross-Modal Associative Memory	Maintain visual-audio identity bindings within working memory when one modal cue temporarily disappears.
Level 3: Spatiotemporal Episodic Reconstruction (Cognitive Basis: Episodic Memory / Spatiotemporal Indexing) Focuses on autonoetic consciousness, requiring a global spatiotemporal index independent of the current view.	
Entity Constancy & Evolution *(Tracking who and what changed.)*	
Counting	Track objects across frames to avoid double-counting during occlusion or re-entry.
Entity State Tracking	Maintain the state evolution history of specific entities against proactive and retroactive interference.
Information Granularity *(Evaluating macro-compression and micro-retrieval.)*	
Scene & Synopsis	Perform semantic compression and macro-level summarization of the entire video.
Visual Needle-in-a-Haystack (Lu et al., 2025; Li et al., 2026b)	Pinpoint and extract microscopic visual details from extremely long interfering contexts.
The Temporal Dimension *(Reconstructing chronological logic.)*	
Action Sequencing	Reconstruct the logical chronological order of events and locate the exact start or end times offline.
Duration Estimation	Maintain attention over long spans to quantify the elapsed time of specific events.
The Spatial Dimension *(Building spatial maps and understanding structures.)*	
Dynamic Spatial Relationships	Track changing relative positions between objects over time.
Egocentric Navigation	Infer spatial structures and navigation instructions from first-person visual cues.
3D Layout Inference	Reconstruct 3D spatial topologies and environmental layouts from 2D video frames.

without genuine long-video memory. Second, we use hierarchical description and QA generation prompts to construct memory-oriented evidence and multiple-choice questions from grounded video descriptions. All prompts enforce constrained outputs, such as a single option letter, strict JSON, or the exact string `None`, to reduce parsing ambiguity and support automatic post-processing.

G.1 Gating Prompts for Sample Filtering

VideoMem-Bench uses four gates to remove shortcut samples. Gate 1 checks whether a question can be answered from language priors alone. Gate 2 tests whether a single frame is sufficient. Gate 3 estimates whether the required evidence window is short-term or long-term without exposing the model to the video. Gate 4 further separates temporal memory from causal or counterfactual reasoning based on the model-generated reasoning trace.

Gate 1: Text-Only Bias Check

Table 8: Complete cognitive taxonomy and task definitions in VideoMem-Bench (continued).

Task Category	Cognitive Basis & Task Definition
Level 4: Predictive World Simulation (Cognitive Basis: Prospective Memory / Mental Simulation) Focuses on building a dynamic world model to simulate unobserved states using physical and behavioral priors.	
Physical World Simulation *(Deducing physical laws and causality.)*	
Explanatory Reasoning (Why-QA)	Deduce the underlying causes behind observed events.
Physical Commonsense	Reason based on Newtonian mechanics and daily physical laws, such as collision causes and falling directions.
Feasibility Judgment	Assess whether a specific action is executable under current environmental constraints.
Social World Simulation *(Inferring mental states and social interactions.)*	
Intent & Belief Tracking	Track mental states, prior knowledge, and belief changes of characters to explain current reactions.
Social Norm Inference	Combine episodic and semantic memory to judge whether actions fit social contexts and predict consequences.
Pragmatic Deception Detection	Compare visual ground truth with verbal claims to detect lies, irony, or sarcasm.
Spatiotemporal Deduction *(Projecting alternative or future realities.)*	
Counterfactual Reasoning	Simulate alternative outcomes by altering historical conditions in what-if scenarios.
Hypothetical Scenarios	Infer sequence results under broad hypothetical environmental changes.
Anticipation & Prediction	Predict inevitable or probable future states based on current trajectories.
Other: Reality Monitoring (Cognitive Basis: Reality Monitoring & Discrimination)	
Negative Memory & Hallucination	Test the ability to reject non-existent visual elements and identify what did not happen in the video.

```
Question: {question}
Options:
{A/B/C/D options}
Answer directly with a single letter (A, B, C, or D).
```

Gate 2: Single-Frame Sufficiency Check

```
Question: {question}
Options:
{A/B/C/D/... options}
Based on the provided image, answer the question directly with a single
↪ letter
(A, B, C, or D).
```

Gate 3: Visual Window Routing

```
You are an expert Video Question Routing System.
You do NOT see the video. Infer the required visual window from question
↪ type
and answer.

{optional: Video Caption (for context only)}

Input:
- Question: {question}
- Correct Answer: {answer_gt}

Classification Criteria:
1) SHORT_TERM (Level 2)
- Instantaneous recognition or short continuous action/state.
- Evidence likely within a single shot or short window (<10 seconds).
- No need to bridge disjoint shots.

2) LONG_TERM (Level 3+)
- Requires sequence ordering, counting over time, global tracking,
or cross-shot integration.
- Evidence likely spans multiple shots/long timeline (>10 seconds).
```

```
Output STRICT JSON only:
{
        "reasoning": "Brief rationale in 2-4 sentences",
        "estimated_window": "< 10 seconds" or "> 10 seconds",
        "label": "SHORT_TERM" or "LONG_TERM"
}
```

Gate 4: Reasoning Trace and Dependency Classification

```
[Stage A: Video-Based Reasoning]

Input Question: {question}
Options:
{options_formatted}
Your task is to answer the question based on the video content.
Please strictly follow these steps:
1. Generate a detailed "Reasoning Trace" (Chain-of-Thought).
2. Select the best option letter as your "Prediction".

OUTPUT FORMAT:
You must output a single valid JSON object strictly matching this format.
Constraints: "prediction" must be a single uppercase letter (A, B, C,
↪  etc.).

Example:
{
        "reasoning_trace": "Analysis of the video content...",
        "prediction": "A"
}


[Stage B: Core Dependency Classification]

You are a logic analyzer. Classify the "Core Dependency" of the reasoning
↪  into:
TEMPORAL or CAUSAL.

Definitions:
1. TEMPORAL: Tracking objects over time, counting, ordering events,
recalling past attributes.
2. CAUSAL: Inferring intent, predicting future physics, explaining 'why',
counterfactuals.

Original Question: {question}

Reasoning Trace:
{reasoning_trace}

OUTPUT FORMAT:
Output a single valid JSON object.

Constraints:
- "dependency" must be exactly "TEMPORAL" or "CAUSAL".
- "evidence" must be a list of 1-2 short quotes
(max 25 words each) from the reasoning trace.

Example:
{
        "dependency": "TEMPORAL",
        "evidence": ["First quote from reasoning", "Second quote"]
}


[JSON Repair Prompt]

Error: {fail_reason}. Please strictly output a VALID JSON object
(no comments, no extra text).
```

G.2 Full-Sample Relabeling Prompt

In addition to the gating pipeline, we use a text-only relabeling prompt to assign each QA item to one of five cognitive levels. This relabeling step provides a unified level definition for downstream analysis. The prompt only uses the question, options, and ground-truth answer text, which prevents the relabeling model from introducing video-dependent cues that are not present in the final QA item.

Full-Sample Cognitive Level Relabeling

```
You are classifying a QA item into one cognitive level.

Definitions:
- LEVEL0: solvable by text prior/bias without visual evidence.
- LEVEL1: solvable by one key frame, no temporal memory required.
- LEVEL2: short-term temporal window (<10s) in one continuous shot.
- LEVEL3: long-term spatiotemporal accumulation
(tracking/order/count across time).
- LEVEL4: causal/predictive/counterfactual reasoning.

Return STRICT JSON ONLY:
{
        "pred_level": "LEVEL0|LEVEL1|LEVEL2|LEVEL3|LEVEL4",
        "pred_reasoning": "2-4 sentences",
        "pred_confidence": "HIGH|MEDIUM|LOW"
}

Question: {question}
Options:
{options_text}
Ground-truth answer text: {answer_text}


[JSON Repair Prompt]

Schema error: {err}. Return one valid JSON object only.
```

G.3 Hierarchical Video Description Prompts

VideoMem-Bench constructs memory-ready video representations through a three-level description hierarchy. Level 1 records local observable events within a short window. Level 2 compresses recent Level-1 descriptions into a rolling story-so-far summary. Level 3 aggregates the full video into a structured spatiotemporal memory, including event chronology, entity tracks, spatial relations, countable facts, causal hypotheses, negative facts, and uncertainty notes. This hierarchy supports query-agnostic memory construction because the descriptions are generated before downstream questions are produced.

Hierarchical Description System Prompt

```
You are a strict long-video benchmark data builder.
Ground responses in provided evidence only.
Return JSON only when requested.
```

Level-1 Local Window Description

```
Task: Generate one Level-1 local description for the current window.

Inputs:
- video_id: {video_id}
- window_start_sec: {window_start_sec}
- window_end_sec: {window_end_sec}
- sampled_frames: {sampled_points}
- latest_level2_summary: {latest_level2_summary}
- unsummarized_recent_level1: {unsummarized_recent_level1}

Requirements:
- Focus on observable events in this window.
- Keep precise temporal ordering.
- Include uncertainty notes for ambiguous details.

Output JSON only:...
```

Level-2 Rolling Summary

```
Task: Generate one Level-2 rolling summary at the current 30-second
↪  boundary.

Inputs:
```

```
- video_id: {video_id}
- boundary_sec: {boundary_sec}
- previous_level2_summary: {previous_level2_summary}
- recent_level1_segments: {recent_level1_segments}

Requirements:
- Summarize story-so-far up to boundary_sec.
- Compress while retaining key state transitions and timeline order.
- Keep memory-ready structure for later windows.
- If unresolved ambiguity exists, keep it in open_uncertainties.

Output JSON only:...
```

Level-3 Full-Video Description

```
Task: Generate the final Level-3 full-video detailed description at video
↪  end.

Inputs:
- video_id: {video_id}
- latest_level2_summary: {latest_level2_summary}
- tail_level1_segments_not_merged: {tail_level1_segments}
- optional_video_meta: {optional_video_meta}

Requirements:
- Build a global spatiotemporal index across the whole video.
- Reconstruct chronology and entity evolution.
- Include details useful for downstream L3 and L4 QA generation.
- Separate observed facts from inferred hypotheses.
- Use stable event IDs.

Output JSON only:...
```

G.4 Multiple-Choice QA Generation Prompts

Question generation is grounded in the final Level-3 description. The standard generator follows the predefined question type, cognitive basis, applicable conditions, and in-context examples. If the given type is not supported by the evidence, the model must output `None`. A relaxed generator is used as a supplement when the strict generator fails to produce enough valid samples. It keeps the same grounding requirement but allows a softened task definition while still forbidding unsupported distractors.

Standard MCQ Generation

```
You are a strict QA generator. Return JSON or None only.
Task: Generate at most ONE multiple-choice QA for the given type.

Target level: {qtype.level}
Type id: {qtype.type_id}
Type name: {qtype.name}
Variant idx: {variant_idx}
Task definition: {qtype.task_definition}
Cognitive basis: {qtype.cognitive_basis}
Applicable conditions: {json(qtype.applicable_conditions)}
Non-applicable conditions: {json(qtype.non_applicable_conditions)}
Invalid answer prefixes: {json(qtype.invalid_answer_prefixes)}
ICL examples: {first 3 of qtype.icl_examples as json}
Final video description: {json(row.level3_final)}

If this type is not applicable, output exactly: None

If applicable, output JSON only:...
```

Relaxed MCQ Generation for Pool Supplementation

```
[System Prompt]

You are a strict long-video QA generator.
Ground every question in the supplied Level-3 description.
Return JSON only or the exact string None.
```

```
[User Prompt]

Task: Generate at most ONE grounded multiple-choice QA for the given type.

Target level: {qtype.level}
Type id: {qtype.type_id}
Type name: {qtype.name}
Original task definition: {qtype.task_definition_original}
Relaxed task definition: {qtype.task_definition_relaxed}
Relaxation note: {qtype.relaxation_note}
Applicable conditions: {json(qtype.applicable_conditions_relaxed)}
Non-applicable conditions: {json(qtype.non_applicable_conditions_relaxed)}
Grounding rules: {json(qtype.grounding_rules)}
Invalid answer prefixes: {json(qtype.invalid_answer_prefixes)}
ICL examples: {first 3 of qtype.icl_examples as json}
Final video description (Level-3): {json(row.level3_final)}

Rules:
- Generate a question only if the Level-3 description gives enough
↪  grounded evidence.
- Keep the question faithful to the target type.
- Prefer conservative inference over speculative reasoning.
- Do not use 'not mentioned', 'not shown', or similar unsupported
↪  distractors.
- Provide four distinct answer options.
- Exactly one option must be best supported by the description.
- Keep rationale brief and grounded in the provided description.

If this type is still not applicable, output exactly: None

If applicable, output JSON only:...
```