PHYSMRV: PHYSICAL MEMORY RETRIEVAL AND VERIFICATION FOR PHYSICAL PLAUSIBILITY REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-language models (VLMs) have achieved remarkable performance on traditional visual question answering tasks, yet they remain unreliable in reasoning about physical plausibility, where understanding object interactions, causal dynamics, and fundamental physical principles is essential. This limitation is particularly evident on challenging physical reasoning benchmarks, revealing a persistent gap in this field. To address this challenge, we propose PhysMRV, a training-free physical memory and verification framework for physical plausibility reasoning. PhysMRV constructs a Hierarchical Memory Bank of structured physical knowledge comprising three complementary levels: scene descriptions capturing general visual information, event graphs modeling object interactions and state transitions, and physics-rule summaries distilling reusable physical rules. During inference, PhysMRV retrieves physically relevant memories and leverages their structured physical evidence to guide the VLM in verifying physical plausibility without requiring parameter updates. We evaluate PhysMRV with multiple advanced VLMs across three physical plausibility reasoning benchmarks: ImplausiBench, IntPhys2, and GRASP Level 2. Experimental results demonstrate consistent improvements across diverse VLM backbones and strong baselines, highlighting the effectiveness of structured physical memory and evidence-guided verification framework for physical plausibility reasoning.

1 INTRODUCTION

Recent VLMs have made substantial progress in VQA tasks, demonstrating strong performance on visual perception as well as mathematical, spatial, and commonsense reasoning (Bai et al., 2025; NVIDIA, 2026; Wang et al., 2025; An et al., 2026; Lu et al., 2024; Chen et al., 2024; Schwenk et al., 2022; Xin et al., 2025).However, they still struggle to understand and predict real-world dynamics and to reason about the underlying physical mechanisms (Ballout et al., 2025). In this work, we focus on physical plausibility reasoning, which aims to determine whether the input video is consistent with physical reality. For example, judging whether an object remaining suspended after its support is removed is physically plausible requires reasoning about support relations and temporal dynamics, rather than merely identifying the object. Such judgments require connecting visual observations to the underlying physical principles governing object interactions and state transitions.

Prior benchmarks, including ImplausiBench (Motamed et al., 2025), IntPhys2 (Bordes et al., 2025), and GRASP (Jassim et al., 2023), have revealed limitations of current VLMs distinguishing between physically plausible and implausible events. Related findings from InPhyRe (Sreekumar & Boddeti, 2026) show that VLMs can exhibit language bias and overlook visual inputs when reasoning about unfamiliar physical dynamics. Existing approaches have provided valuable insights into this challenge. Physics Context Builders (Balazadeh et al., 2025) generates structured scene descriptions, while TRAVL (Motamed et al., 2025) fine-tunes lightweight spatial and trajectory-aware temporal attention modules on physics-focused video data. Furthermore, retrieval and memory mechanisms have been introduced to augment models with auxiliary information, spanning video comprehension (Luo et al., 2025), long-term memory (Zhong et al., 2024), and embodied task planning (Ma et al., 2026).

In this work, we focus on three challenges in physical plausibility reasoning: (1) **insufficient perception**, where models may overlook critical visual details as videos unfold over time. TRAVL (Motamed et al., 2025) shows that current models remain limited in capturing motion continuity and

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.

identifying spatiotemporal inconsistencies; (2) **inadequate physical event representation**, where models may fail to capture explicit interactions and state transitions, such as object contact and changes in motion following a collision. Physics Context Builders (Balazadeh et al., 2025) demonstrates the value of detailed physical descriptions for downstream reasoning; and (3) **failure to apply rules to specific observations**, where knowing a physical principle does not ensure its correct application to an observed event. InPhyRe (Sreekumar & Boddeti, 2026)reports that models can possess parametric knowledge of conservation laws yet struggle to apply it when reasoning about collision outcomes.

To directly address these challenges, we propose PHYSMRV, a training-free framework for physical plausibility reasoning that integrates hierarchical physical memory, physically grounded retrieval, and evidence-guided verification. At its core, PHYSMRV constructs a *Hierarchical Physical Memory Bank* offline from source data, organizing each memory instance into three complementary layers: (1) a **Scene Context Layer**, capturing the global visual context through a scene caption; (2) a **Physical Event Layer**, representing physical interactions and state transitions through an event graph; and (3) a **Physical Rule Layer**, summarizing relevant physical rules. This hierarchy connects scene context to physical event structure and reusable physical knowledge, providing structured evidence for retrieval and verification.

During inference, PHYSMRV consists of two stages: **physics-aware retrieval** and **evidence-guided verification**. The retrieval stage first selects candidate memories through semantic matching and then reranks them through event-graph alignment, prioritizing memories with relevant physical interactions and state transitions. In the verification stage, the framework combines retrieved questions, explanations, and physical rules with the event graph extracted from the query video to construct structured physical evidence. The query graph represents the current visual cues, while the retrieved memory instances and rules provide references for evaluating their plausibility. This explicit organization guides the VLM to evaluate whether the observed interactions and state transitions satisfy the relevant physical constraints, rather than relying solely on semantic associations or model priors.

We evaluate PHYSMRV on ImplausiBench, IntPhys2, and GRASP Level 2 using multiple VLM backbones, including Qwen3-VL, Cosmos-Reason2, InternVL3.5, and LLaVA-OneVision-2. Experimental results consistently show improvements in physical plausibility reasoning without updating model parameters. Ablation studies further show that both physics-aware retrieval and structured physical evidence contribute to the performance gains.

2 RELATED WORK

2.1 PHYSICAL REASONING IN VISION-LANGUAGE MODELS

Benchmarks such as ImplausiBench (Motamed et al., 2025), IntPhys2 (Bordes et al., 2025), GRASP (Jassim et al., 2023), and PhysBench (Chow et al., 2025) reveal persistent limitations of VLMs in understanding physical interactions and dynamics. Existing methods generally fall into two categories: training-based approaches that adapt models for physical reasoning, and context-based approaches that provide additional physical information during inference. TRAVL (Motamed et al., 2025) combines trajectory-aware attention with physics-focused data fine-tuning for implausibility detection, while PhysVLM (Cao et al., 2024) uses instruction tuning and preference optimization to improve the recognition of physical commonsense understanding in gameplay videos. PhyDetEx (Wang et al., 2026b) fine-tunes VLMs to detect and explain physical violations, and PhyAR (Zhao et al., 2026) combines visually grounded reasoning with pair-wise optimization to reduce reliance on semantic priors. Physics Context Builders (Balazadeh et al., 2025) generate physics-enriched scene descriptions, while PhysAgent (Chow et al., 2025) leverages specialized vision models to provide additional physical knowledge for downstream reasoning. These approaches strengthen physical perception, provide reasoning context, or adapt models to physics-focused data. Unlike methods that rely on task-specific fine-tuning, PHYSMRV can be applied across different VLM backbones without parameter updates.

2.2 MEMORY AND RETRIEVAL

Memory and retrieval mechanisms augment model inference with task-relevant information from stored experiences or external sources. MemoryBank (Zhong et al., 2024) maintains conversational memories to support long-term interactions with LLMs, while Video-RAG (Luo et al., 2025) sup-

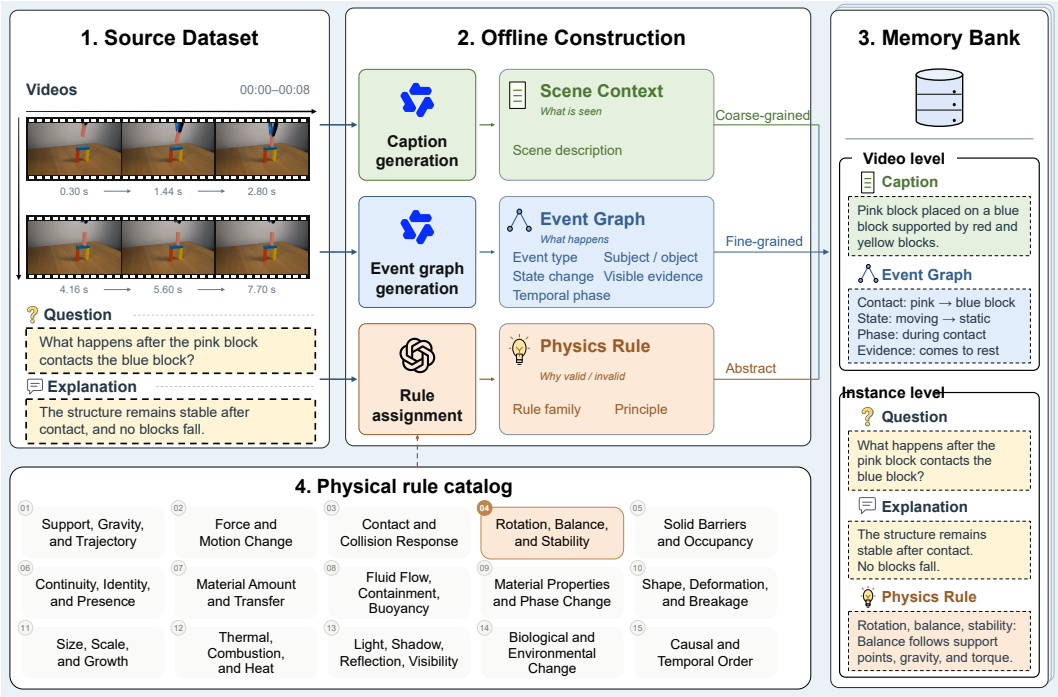


Figure 1: Overview of the hierarchical memory bank in PHYSMRV. The bank is constructed offline from source videos and their associated questions and explanations. Each memory instance links a question and its explanation to three complementary representations: a scene caption capturing visual context, a physical event graph recording interactions and state transitions, and a physical rule summarizing relevant physical constraints.

plements video inputs with query-relevant auxiliary text derived from optical character recognition, speech recognition, and object detection. In embodied tasks, BrainMem (Ma et al., 2026) organizes interaction histories into knowledge graphs and reusable planning guidelines, and PhysMem (Li et al., 2026) learns physical principles from test-time interaction by generating and verifying hypotheses before using the resulting knowledge to guide embodied planning. For physical reasoning, PhysNote (Zhang et al., 2026) organizes self-generated knowledge into hierarchical notes and uses a context selection agent to select task-relevant notes based on the question and visual inputs. In contrast, PHYSMRV constructs a hierarchical physical memory bank once offline, linking prior video instances to physical event structures and reusable rules. It combines semantic retrieval with event-graph alignment to select relevant memories for physical plausibility verification. The same memory bank is reused across all evaluated benchmarks.

2.3 MULTIMODAL REASONING

General-purpose VLMs such as Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025) support diverse visual understanding and reasoning tasks. Beyond backbone improvements, multimodal chain-of-thought methods (Lu et al., 2022; Zhang et al., 2024; Zheng et al., 2023; Feng et al., 2025; Wu et al., 2026) enhance reasoning through intermediate rationales that guide answer prediction. CCoT (Mitra et al., 2024) makes object attributes and relations explicit through scene graphs, while Video-of-Thought (Fei et al., 2024)and recent grounded reasoning methods (Wang et al., 2026a) incorporate structured spatiotemporal or visual evidence into the reasoning process. Together, these methods highlight the value of explicit intermediate representations. Physical plausibility reasoning further requires relating observed interactions and state transitions to physical constraints, motivating our use of physical event graphs and relevant physical knowledge.

3 METHOD

Overview. PHYSMRV is a training-free framework for physical plausibility reasoning that combines a shared **Hierarchical Memory Bank** as shown in 1 with two-stage inference: **physics-aware retrieval** and **evidence-guided verification**, as shown in 2. The memory bank is constructed once

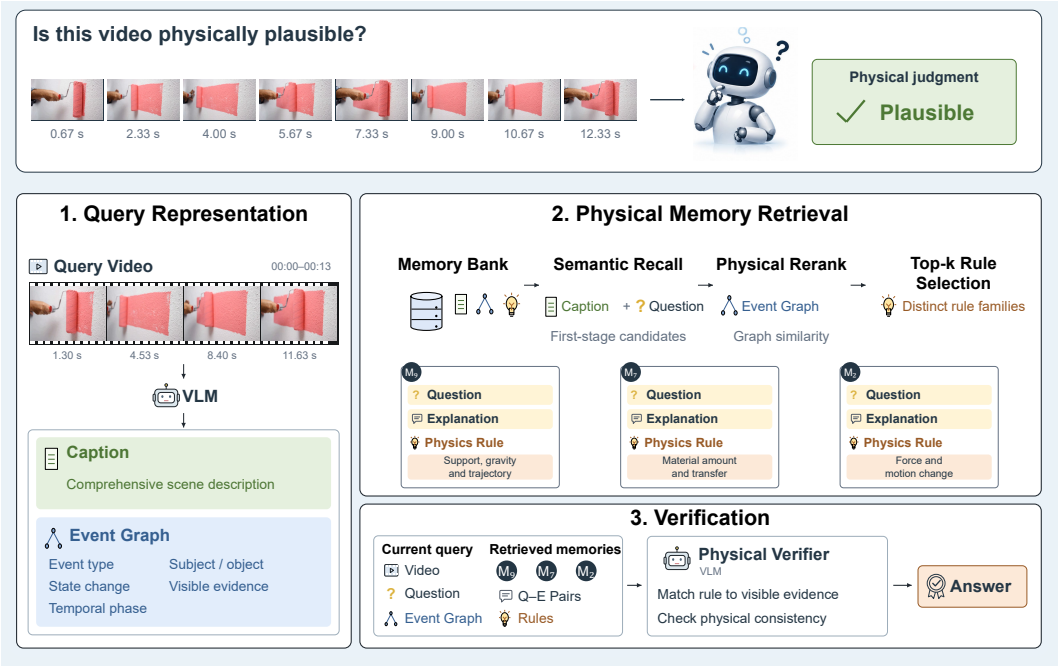


Figure 2: **Inference procedure of PHYSMRV.** At inference time, PHYSMRV operates on the hierarchical physical memory bank constructed offline and follows a two-stage pipeline: physics-aware memory retrieval and evidence-guided verification. In the retrieval stage, caption–question similarity recalls candidate memories, which are then reranked by physical-event similarity to select the top three memory instances. In the verification stage, the VLM evaluates the query video using query event graph together with the retrieved questions, explanations, and physical rules.

offline from existing physical-reasoning video data and can be reused across benchmarks and VLM backbones. Each memory instance links a question and its explanation to three complementary layers: a scene caption describing the global visual context, an event graph representing physical interactions and state transitions, and physical rules specifying relevant physical constraints. This hierarchy connects concrete visual observations to event structure and reusable physical knowledge.

During inference, given a query video, the coarse-to-fine retrieval stage first selects candidate memories through semantic matching and then reranks them through physical event-graph alignment to prioritize physically relevant cases. The verification stage combines the query event graph with retrieved questions, explanations, and physical rules to form structured physical evidence. Together with the query video and question, this evidence guides the VLM to determine whether the observed events satisfy the relevant physical constraints.

3.1 PRELIMINARIES

Given a video v_i, a question q_i, and labeled candidate options $\mathcal{A}_i$ (when provided), the task is to predict an answer $\hat{y}_i \in \mathcal{Y}_i$, where $\mathcal{Y}_i$ is a finite, task-specific answer space. This space consists of the option labels for multiple-choice instances or the prescribed binary labels, such as $\{\text{yes}, \text{no}\}$ or $\{0, 1\}$. We define the observable input as $x_i = (v_i, q_i, \mathcal{A}_i)$, with $\mathcal{A}_i = \varnothing$ when no explicit option list is provided. PHYSMRV augments this input with a fixed memory bank $\mathcal{M}$ constructed offline from source data and uses a frozen VLM F_θ for answer generation. All pretrained components are used without additional parameter updates.

3.2 HIERARCHICAL PHYSICAL MEMORY BANK

We construct a Hierarchical Physical Memory Bank from existing physical-reasoning video data. Each instance links an original question from source data and its explanation to three complementary levels of physical representation: scene context, observed events, and general principles. Formally,

$$\mathcal{M} = \{m_j\}_{j=1}^{N}, \qquad m_j = (c_j, g_j, q_j, e_j, r_j), \tag{1}$$

where c_j is the scene caption, g_j is the physical event graph, q_j and e_j are the original question and explanation, and r_j is the associated physical rule. Scene context and event graphs are generated by a pretrained vision-language model and shared across memory instances from the same video. Physical rules are assigned to each instance by a vision-language model according to a predefined taxonomy of physical rule families derived from the source data. This hierarchy connects scene observations to physical event structure and reusable physical knowledge.

Scene Context Layer. Caption describes the video's global scene context, covering salient objects, their activities, and visible interactions such as contact, support, and containment. It also summarizes motion and temporal changes to provide a coherent account of what occurs in the video.

Physical Event Layer. A physical event graph captures the key interactions and state transitions in a video. The VLM first identifies the primary subject and the entities involved in its interaction, such as a supporting surface, contacted object, or container. It then characterizes the interaction and the subject's states before, during, and after the event. Distinct physical stages are represented separately when they provide additional evidence. Each event includes its type, participants, temporal phase, observed states, state change, and visible evidence. The state-change component specifies the affected attribute, type of change, and initial and final states. Event types and temporal phases follow predefined vocabularies to maintain consistent descriptions across videos.

This representation preserves physical details that a global scene description may leave unspecified: what interacts with what, when an interaction occurs, and how the object's state changes. For example, when a support is removed, the graph records whether the object begins to fall or remains stationary, together with the visible evidence of its motion and support relation. Videos containing similar objects can therefore be distinguished by their state transitions, while events in different scenes can be matched through shared physical dynamics. These event records provide the basis for physical-event reranking and retain the observations used in subsequent verification.

Physical Rule Layer. Each instance is assigned to a primary family from a fixed catalog of fifteen physical-rule and inherits the corresponding general principle. These families cover phenomena such as support and gravity, force and motion, contact and collision, object continuity, and material transfer. Offline model-based assignment uses the question, its explanation, and video input to select the relevant family.

3.3 Physics-Aware Memory Retrieval

We adopt a coarse-to-fine retrieval strategy that combines semantic recall with physical-event reranking. Scene context and question semantics provide a broad relevance signal, but videos with similar objects and activities may involve different physical interactions and state transitions. We therefore first retrieve semantically related candidates, then compare their event graphs to prioritize memories with similar physical dynamics. This hierarchical approach also bypasses the computational cost of exhaustively matching detailed events across the entire memory bank. Building upon our predefined representation schemas, we extract a query caption c_i^{query} and a query event graph $g_i^{\text{query}} = \Gamma(v_i, q_i)$, where Γ denotes the VLM-based event-graph extraction function.

Semantic Recall. We first construct a textual representation for the query and each memory instance by concatenating the corresponding caption and question:

$$\tau_i^{\text{query}} = c_i^{\text{query}} \oplus q_i, \qquad \tau_j = c_j \oplus q_j, \tag{2}$$

where i indexes query instances, j indexes memory instances and $\oplus$ denotes text concatenation with field labels. All-mpnet-base-v2 text encoder (Reimers & Gurevych, 2019) is then used to project each text sequence into an ℓ_2-normalized embedding $\phi(\tau)$. Semantic similarity is computed as

$$S_{\text{sem}}(i, j) = \phi(\tau_i^{\text{query}})^\top \phi(\tau_j). \tag{3}$$

We retain the top 50 memory instances by semantic similarity as the candidate set $\mathcal{C}_i^{50}$.

Physical-Event Reranking. We rerank the candidates in $\mathcal{C}_i^{50}$ by comparing their structured events with those of the query. Let E_i^{query} and E_j denote the event sets of g_i^{query} and g_j for $j \in \mathcal{C}_i^{50}$, respectively. Events are compared across five distinct dimensions, $\mathcal{P} = (\text{type}, \text{participants}, \text{state}, \text{evidence}, \text{phase})$, corresponding to event type, subject and interacting counterpart, state change, visible evidence, and temporal phase. For each component p, let $\mathbf{u}_p(e)$

denote the ℓ_2-normalized embedding of its textual representation in event e, obtained using the same frozen text encoder. The similarity between two events is

$$M(e, e') = \sum_{p \in \mathcal{P}} \alpha_p \, \mathbf{u}_p(e)^\top \mathbf{u}_p(e'), \tag{4}$$

with fixed weights $(0.35, 0.20, 0.25, 0.15, 0.05)$ in the component order above, emphasizing event type and state change. The query-to-memory graph score is then defined as

$$S_{\text{graph}}(g_i^{\text{query}}, g_j) = \frac{1}{|E_i^{\text{query}}|} \sum_{e \in E_i^{\text{query}}} \max_{e' \in E_j} M(e, e'). \tag{5}$$

Each query event is matched to its most similar memory event, and the resulting similarities are averaged, yielding a directional measure of how well the memory covers the query events. Candidates are then reranked by this graph score. We finally select the top three memory instances in descending graph-score order, denoted by $\mathcal{M}_i^3$.

3.4 Evidence-Guided Verification

The verifier combines the query video and event graph with the selected memory instances $\mathcal{M}_i^3$ to evaluate the video. For each instance $m_j \in \mathcal{M}_i^3$, we retain its question q_j, explanation e_j, and physical rule r_j. The selected question–explanation–rule tuples form an ordered list $\mathcal{V}_i$:

$$\mathcal{V}_i = \big((q_j, e_j, r_j)\big)_{m_j \in \mathcal{M}_i^3}, \tag{6}$$

where $\mathcal{V}_i$ preserves the selection order. Retaining these three fields contextualizes the extracted physical constraints: the question identifies the core physical issue, the explanation interprets the behavior, and the rule states the underlying constraint.

During verification, the query event graph g_i^{query} provides a structured representation of the interactions and state transitions extracted from the current video, while the original video v_i remains available as the visual input. The retrieved records in $\mathcal{V}_i$ provide reference cases and physical rules for interpreting the current observation. The verifier is instructed to evaluate the applicability of the retrieved rules to the observed interactions and to use the relevant rules when judging whether the event is physically plausible.

We formulate the final inference as

$$P_i = \rho\left(q_i, \mathcal{A}_i, g_i^{\text{query}}, \mathcal{V}_i\right), \qquad \hat{y}_i = F_\theta(v_i, P_i), \tag{7}$$

where ρ formats the current question, available candidate options, query event graph, ordered memory records, and verification instructions into the textual prompt P_i. The original video v_i is provided separately to the VLM F_θ, which generates the final prediction $\hat{y}_i$ in the task-specific answer space. Complete prompt templates are provided in the appendix.

4 Experiments

We evaluate PhysMRV on three physical plausibility reasoning benchmarks across multiple VLM backbones. We compare it with general reasoning, retrieval-based, and physics-specific baselines to evaluate its effectiveness and generalization across models and benchmarks. We further conduct ablation studies to analyze the contributions of the memory construction, physics-aware retrieval, and evidence-guided verification, and examine whether the performance gains arise from improved physical discrimination rather than answer-polarity bias.

4.1 Experimental Setup

Benchmarks and Metrics. We evaluate our approach on three video benchmarks targeting physical plausibility. (1) **ImplausiBench** (Motamed et al., 2025) contains 300 videos paired into 150 plausible-implausible scenarios. We retain the original multiple-choice format and report exact option-matching accuracy. (2) **IntPhys2** (Bordes et al., 2025) consists of 1,012 synthetic videos in its main evaluation set, testing object permanence, immutability, spatio-temporal continuity, and solidity. We report accuracy over the balanced possible and impossible videos. (3) **GRASP Level 2** (Jassim et al., 2023) features 4,096 synthetic videos with concept-specific binary (yes/no) questions. Following the official

Table 1: Same-backbone results on three physical-reasoning benchmarks (%). CoT: Vanilla CoT; CLIP: CLIP MeanP retrieval; IV2: InternVideo2 retrieval. We report answer accuracy on ImplausiBench, IntPhys2, and GRASP L2. Bold indicates the best score within each backbone and benchmark.

		General Reasoning		Retrieval Baselines		Physics Specific Methods	
Benchmark	Direct	CoT	CCoT	CLIP	IV2	PhysNote	**PhysMRV**
Qwen3-VL-8B-Instruct.							
ImplausiBench	56.00	50.00	52.33	60.33	58.00	49.00	**60.67**
IntPhys2	49.90	50.79	50.40	51.19	52.27	52.17	**53.75**
GRASP L2	48.56	51.10	48.93	52.25	51.32	51.66	**55.20**
Cosmos-Reason2-8B							
ImplausiBench	53.00	45.67	47.33	57.33	56.67	45.00	**58.33**
IntPhys2	49.80	50.49	49.90	51.78	52.37	49.90	**53.36**
GRASP L2	49.98	49.73	50.98	51.66	51.93	53.27	**55.71**
InternVL3.5-8B							
ImplausiBench	56.33	53.67	42.00	58.33	**59.67**	43.67	**59.67**
IntPhys2	49.41	50.99	49.90	50.49	49.80	50.69	**52.08**
GRASP L2	49.07	52.00	51.42	51.25	52.05	51.03	**54.76**
LLaVA-OneVision-2-8B							
ImplausiBench	34.00	26.33	31.00	34.33	34.67	**36.33**	35.67
IntPhys2	48.42	49.80	42.00	48.62	47.63	**51.88**	50.30
GRASP L2	48.90	50.00	49.83	50.20	51.03	49.76	**51.25**

answer-parsing scheme, we report accuracy over all videos, counting a response as correct when its parsed 'yes' or 'no' answer matches the ground-truth label and as incorrect otherwise.

Baselines. We benchmark PHYSMRV across four advanced vision-language models: Qwen3-VL (Bai et al., 2025), Cosmos-Reason2 (NVIDIA, 2026), InternVL3.5 (Wang et al., 2025), and LLaVA-OneVision-2 (An et al., 2026). Our baselines include standard direct prompting (Direct) and zero-shot Chain-of-Thought (Vanilla CoT) (Kojima et al., 2022). We also adapt recent compositional reasoning methods to our video setting, specifically CCoT (Mitra et al., 2024) and PhysNote (Zhang et al., 2026). To evaluate the effectiveness of our two-stage coarse-to-fine retrieval strategy, we further compare against visual-similarity retrieval baselines using CLIP MeanP (Radford et al., 2021) and InternVideo2 (Wang et al., 2024). CLIP MeanP averages frame-level CLIP features into a video representation, while InternVideo2 directly provides video-level features. The subsequent verification procedure remains unchanged.

Implementation Details. We use Qwen3-VL-8B-Instruct to generate video captions with deterministic decoding (temperature = 0, max_new_tokens = 160) and physical event graphs (max_new_tokens = 1280, video_max_pixel = 151,200). We use the GPT-5.5 API for physical rule assignment based on a predefined rule-family taxonomy iteratively developed through GPT-5.5-assisted generation and manual refinement. Video inputs are uniformly decoded at 1 FPS using TorchCodec. For final answer generation, we use deterministic decoding with a maximum of 256 new tokens for ImplausiBench and IntPhys2 and 32 for GRASP. Experiments are run on a server with 8 NVIDIA A100 GPUs.

4.2 MAIN RESULTS

Improvements over direct inference. Table 1 reports results on three physical plausibility benchmarks with four VLM backbones. PHYSMRV outperforms Direct in all twelve backbone–benchmark combinations, with gains ranging from 1.67 to 6.64 percentage points (pp). For Qwen3-VL, performance increases from 56.00 to 60.67 on ImplausiBench, 49.90 to 53.75 on IntPhys2, and 48.56 to 55.20 on GRASP L2. The largest gains consistently occur on GRASP L2, reaching 6.64, 5.73, 5.69, and 2.35 pp across the four backbones. Overall, these results show that PHYSMRV consistently improve physical plausibility reasoning over direct inference without updating the VLM parameters.

Comparison with competing baselines. Compared with general reasoning methods, PHYSMRV consistently outperforms Vanilla CoT and CCoT across all twelve backbone–benchmark combinations, showing that explicitly retrieving and verifying physical evidence provides more reliable gains than generic reasoning strategies. Against retrieval baselines, PHYSMRV matches or surpasses both

Table 2: Comparasion of PHYSMRV with different Physics Specific Models

Method	ImplausiBench	IntPhys2	GRASP L2
TRAVL (reproduced)	31.33	33.50	49.34
VideoPhy2 (QA-adapted)	–	50.49	52.78
PhysNote(Qwen3-VL-8B-Instruct)	49.00	52.17	51.66
PhysMRV (Qwen3-VL-8B-Instruct)	**60.67**	**53.75**	**55.20**

CLIP MeanP and InternVideo2 in every setting, demonstrating the benefit of retrieving memories based on physical event structure rather than visual similarity alone. Compared with the physics-specific PhysNote baseline, PHYSMRV achieves higher accuracy in ten of the twelve settings. The advantage is particularly clear on GRASP L2, where PHYSMRV obtains the highest score with all four backbones and exceeds the strongest competing baseline by 2.95, 2.44, and 2.71 pp with Qwen3-VL, Cosmos-Reason2, and InternVL3.5, respectively. Overall, PHYSMRV achieves the best or tied-best result in ten of the twelve backbone–benchmark combinations, with the only exceptions occurring for LLaVA-OneVision-2 on ImplausiBench and IntPhys2.

Generalization across benchmarks and backbones. The same physical memory bank is constructed once offline and reused across all three benchmarks and four VLM backbones, without updating the backbone parameters. Across all three benchmarks, PHYSMRV consistently improves over direct inference, covering both the multiple-choice setting of ImplausiBench and the binary plausibility judgments of IntPhys2 and GRASP L2. More importantly, the improvements remain consistent across all four VLM backbones, with substantial gains observed for models with different architectures and baseline capabilities. Overall, the results show that a single shared physical memory bank can effectively transfer across both benchmarks and backbones.

4.3 PHYSICS-SPECIFIC MODELS

To further evaluate PHYSMRV against methods specifically designed for physical reasoning, we compare it with TRAVL (Motamed et al., 2025) and VideoPhy2 (Bansal et al., 2025). All comparisons use Qwen3-VL-8B-Instruct as the backbone for PHYSMRV.

For TRAVL, we follow its original training procedure for five epochs and evaluate the final checkpoint. For VideoPhy2, we use the released checkpoint and adapt its input–output interface to our benchmark QA format. On GRASP L2, its 0/1 predictions are mapped to no/yes, respectively. We do not evaluate VideoPhy2 on ImplausiBench because it performs binary plausibility prediction rather than multiple-choice reasoning.

Table 2 shows that PHYSMRV consistently outperforms TRAVL across all three benchmarks. The improvement is particularly substantial on ImplausiBench, where PHYSMRV exceeds TRAVL by 29.34 percentage points, while also maintaining higher accuracy on IntPhys2 and GRASP L2. Compared with VideoPhy2, PHYSMRV achieves gains of 3.26 percentage points on IntPhys2 and 2.42 percentage points on GRASP L2. These results show that the advantage of PHYSMRV extends across both multiple-choice and binary plausibility settings. Notably, these improvements are achieved without task-specific fine-tuning, demonstrating that retrieving and verifying relevant physical memories can provide competitive gains over models explicitly trained for physical reasoning.

4.4 ABLATION STUDIES

Table 3: Component ablations of PHYSMRV with Qwen3-VL-8B-Instruct. Scores are percentages.

Variant	Query Graph	Memory	Verification	ImplausiBench	IntPhys2	GRASP L2
w/o Query Graph	×	✓	✓	58.00	50.00	53.44
w/o Memory	✓	×	✓	57.67	52.08	52.56
w/o Verification	✓	✓	×	57.00	49.51	53.05
Full (PHYSMRV)	✓	✓	✓	**60.67**	**53.75**	**55.20**

We examine component contributions, retrieval strategies, and memory content using Qwen3-VL-8B-Instruct on ImplausiBench, IntPhys2, and GRASP L2.

Contributions of core components. Table 3 evaluates the query event graph, retrieved memories, and evidence-guided verification. Removing the query graph from the verifier input, without changing retrieval, lowers performance on all three benchmarks, as does removing retrieved memories. Among

Table 5: Direct inference versus PHYSMRV on three physical plausibility benchmarks. P/Y and I/N denote class-wise accuracy on physically plausible and implausible samples, respectively. Metric denotes original-option accuracy on ImplausiBench, balanced accuracy on IntPhys2, and accuracy on GRASP L2. Scores are reported as percentages.

Model	Method	ImplausiBench			IntPhys2			GRASP L2		
		P/Y	I/N	Metric	P/Y	I/N	Metric	P/Y	I/N	Metric
Qwen3-VL-8B-Instruct.	Direct	87.33	24.67	56.00	75.89	23.91	49.90	94.87	2.25	48.56
Qwen3-VL-8B-Instruct.	PHYSMRV	88.00	33.33	**60.67**	80.43	27.08	**53.75**	62.45	47.95	**55.20**
InternVL3.5-8B	Direct	84.00	28.67	56.33	97.23	1.58	49.41	96.39	1.76	49.07
InternVL3.5-8B	PHYSMRV	87.33	32.00	**59.67**	91.70	12.45	**52.08**	36.52	73.00	**54.76**

these component ablations, removing verification while retaining the graph and memories causes the largest drops on ImplausiBench and IntPhys2 (3.67 and 4.24 pp, respectively), supporting the contribution of evidence-guided verification.

Retrieval and matching. Table 4 compares retrieval variants with verification unchanged. Removing physical reranking reduces scores by 2.29–2.76 pp across the three benchmarks, while graph-only retrieval also underperforms the full method on all three. Replacing structured event-graph alignment with flat-text event matching lowers scores on ImplausiBench and IntPhys2. Together, these results support combining scene-level semantic retrieval with structured event matching.

Table 4: Ablations on memory construction and retrieval.

Variant	Implausi.	IntPhys2	GRASP
Retrieval and matching			
w/o Physical Reranking	58.33	50.99	52.91
Graph-only Retrieval	58.67	50.49	51.59
Flat-text Event Matching	57.33	50.69	53.44
Memory content			
Rule Only	58.00	51.28	52.51
w/o Rule	59.00	51.19	52.32
Shuffled Memory	56.67	50.69	52.56
Full (PHYSMRV)	**60.67**	**53.75**	**55.20**

Memory content and relevance. With the query graph verification unchanged, *Rule Only* retains only the retrieved physical rules, whereas *w/o Rule* retains only the retrieved questions and explanations. Both variants underperform the full method on all three benchmarks, supporting the joint use of these memory contents. Shuffling intact memory blocks across queries also lowers performance on all three benchmarks and does not outperform the no-memory variant on any benchmark. This suggests that memory relevance, rather than additional context alone, contributes to the observed gains.

4.5 BIAS IN PLAUSIBILITY JUDGMENTS

Table 5 examine PHYSMRV improves discrimination between plausible and implausible events rather than merely shifting predictions toward implausibility. We compare class-wise accuracy between direct inference and PHYSMRV using Qwen3-VL and InternVL3.5 across three benchmarks.

Plausible events. Direct inference already performs strongly on plausible instances in several settings, especially on GRASP L2, where both backbones exceed 94% accuracy. These events are often consistent with common visual and physical patterns learned during pretraining. PHYSMRV largely preserves this capability, while further improving plausible-instance accuracy in several settings.

Implausible events. Detecting implausible cases requires identifying visual evidence that violates expected physical behavior and relating it to the corresponding physical constraints. PHYSMRV improves implausible-instance accuracy in every setting, including from 2.25% to 47.95% for Qwen3-VL and from 1.76% to 73.00% for InternVL3.5 on GRASP L2. These results indicate that explicit physical-event representations and retrieved physical rules help the model detect violations.

5 CONCLUSION

We introduced PHYSMRV, a physical memory retrieval and verification framework for frozen video-language models. It combines semantic recall, physical-event reranking, and verification against training questions, explanations, and principles. Across three benchmarks, it improves all four evaluated backbones over direct inference. Ablations favor the two-staged retrieved evidence and the current-video graph. The results suggest that retrieved case context and physical rules provide complementary information, with their joint use outperforming either component alone.

Reproducibility Statement

Implementation details, model configurations, retrieval settings, prompt templates, and evaluation protocols are provided in the main paper and appendix. We will release the code and evaluation scripts upon publication.

AI use statement

In this work, we used generative AI tools to assist with several tasks. In particular, GPT-5.5 was used to support the construction and refinement of the physical-rule taxonomy and to assist with rule assignment during memory construction. Generative AI tools were also used to assist with code development and debugging. The physical-rule taxonomy was iteratively reviewed and revised by the authors, and all AI-assisted research outputs, code, experimental analyses, and manuscript content were reviewed and verified by the authors. The authors take full responsibility for the methodology, experimental results, claims, and final content of this work.

References

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. Llava-onevision-2: Towards next-generation perceptual intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Shuai Bai et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. doi: 10.1109/ICCV51701.2025.00687. URL https://arxiv.org/abs/2412.08619.

Mohamad Ballout, Serwan Jassim, and Elia Bruni. Pixels to principles: Probing intuitive physics understanding in multimodal language models, 2025. URL https://arxiv.org/abs/2507.16572.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation, 2025. URL https://arxiv.org/abs/2503.06800.

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849.

Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. Physgame: Uncovering physical commonsense violations in gameplay videos, 2024. URL https://arxiv.org/abs/2412.01800.

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

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q6a9W6kzv5.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition, 2024. URL https://arxiv.org/abs/2501.03230.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 99114–99137. Curran Associates, Inc., 2025. doi: 10.52202/085713-3310. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8eb3976840e08b80dda9667562574246-Paper-Conference.pdf.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. GRASP: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models, 2023. URL https://arxiv.org/abs/2311.09048.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. doi: 10.52202/068431-1613. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Haoyang Li, Yang You, Hao Su, and Leonidas Guibas. PhysMem: Scaling test-time memory for embodied physical reasoning, 2026. URL https://arxiv.org/abs/2602.20323.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2411.13093.

Xiaoyu Ma, Lianyu Hu, Wenbing Tang, Zixuan Hu, Zeqin Liao, Zhizhen Wu, and Yang Liu. BrainMem: Brain-inspired evolving memory for embodied agent task planning, 2026. URL https://arxiv.org/abs/2604.16331.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mitra_Compositional_Chain-of-Thought_Prompting_for_Large_Multimodal_Models_CVPR_2024_paper.html.

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. TRAVL: A recipe for making video-language models better judges of physics implausibility, 2025. URL https://arxiv.org/abs/2510.07550.

NVIDIA. Cosmos-Reason2-8B. https://huggingface.co/nvidia/Cosmos-Reason2-8B, 2026. Hugging Face model card. Initial release: December 19, 2025; updated March 10, 2026. Accessed: 2026-06-30.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022.

Gautam Sreekumar and Vishnu Naresh Boddeti. Inphyre discovers: Large multimodal models struggle in inductive physical reasoning, 2026. URL https://arxiv.org/abs/2509.12263.

Peiyao Wang, Haotian Xu, Noranart Vesdapunt, Rui Hou, Jingyi Zhang, Haibin Ling, Oleksandr Obiednikov, Ning Zhou, and Kah Kuen Fu. Reinforcing structured chain-of-thought for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9794–9803, June 2026a.

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*, 2025.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024*, 2024. URL https://arxiv.org/abs/2403.15377.

Zeqing Wang, Keze Wang, and Lei Zhang. Phydetex: Detecting and explaining the physical plausibility of t2v models. In *European Conference on Computer Vision*, pp. 168–186. Springer, 2026b.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33577–33587, June 2026.

Yue Xin, Wenyuan Wang, Rui Pan, Ruida Wang, Howard Meng, Renjie Pi, Shizhe Diao, and Tong Zhang. Generalizable geometric image caption synthesis, 2025. URL https://arxiv.org/abs/2509.15217.

Sinin Zhang, Yunfei Xie, Yuxuan Cheng, Haoyu Zhang, and Tong Zhang. Physnote: Self-knowledge notes for evolvable physical reasoning in vision-language model, 2026. URL https://arxiv.org/abs/2604.24443.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024. URL https://arxiv.org/abs/2302.00923.

Zicheng Zhao, Chaofan Gan, Shijie Li, and Weiyao Lin. From priors to perception: Grounding video-llms in physical reality, 2026. URL https://arxiv.org/abs/2605.04515.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5168–5191. Curran Associates, Inc., 2023. doi: 10.52202/075280-0228. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/108030643e640ac050e0ed5e6aace48f-Paper-Conference.pdf.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

A Data and Evaluation Protocol

This appendix specifies the memory representation, retrieval procedure, prompt interfaces, baseline adaptations, and evidence-removal controls used by PhysMRV. It also presents a saved qualitative example with its complete retrieved context.

A.1 Memory source and evaluation inputs

The memory contains 19,708 question–explanation entries associated with 3,482 videos from the TRAVL tuning split. Each memory entry references one video-level caption and one video-level event graph. Thus, multiple questions about the same video share visual representations but remain distinct retrieval candidates. The explanation is the original training annotation; the principle is the neutral principle associated with the assigned physical-rule family.

The evaluation inputs comprise 300 ImplausiBench videos, 1,012 IntPhys2 Main videos, and 4,096 GRASP Level 2 videos. ImplausiBench retains the original multiple-choice question and all original options. IntPhys2 uses the possible/impossible judgment encoded as `1`/`0`. GRASP retains each concept-specific yes/no question: its wording, including any assumptions about the objects, is part of the task rather than a generic plausibility question substituted for every example.

Information separation. Caption generation takes the video; evaluation graph generation takes the video and its question/options. Semantic retrieval uses the caption and question, and event matching uses the generated graph. Neither stage uses evaluation labels or reference explanations. The answer model receives the current video, original task, current event graph, and retrieved training questions, explanations, and principles. The evaluation caption is used for retrieval but is absent from the final verifier text.

A.2 Metrics and answer validity

Let a_i be the answer returned by the applicable parser, with $a_i = \perp$ for an invalid response. For option accuracy and original-answer accuracy,

$$\mathrm{Acc} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[a_i = y_i]. \tag{8}$$

Invalid responses remain in the denominator and contribute zero. IntPhys2 uses the mean of the two class recalls,

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\sum_{i:y_i=0}\mathbf{1}[a_i=0]}{n_0} + \frac{\sum_{i:y_i=1}\mathbf{1}[a_i=1]}{n_1}\right). \tag{9}$$

The selected IntPhys2 set has 506 examples per class, so balanced accuracy equals all-example accuracy under this fixed coverage. Class-wise ImplausiBench accuracy measures correct selection of the original option within each plausibility class.

B Physical Memory Representation

B.1 Stored entries and event fields

Each logical entry is $m_j = (c_j, g_j, q_j, e_j, r_j)$, using the notation of the main text. The three representation layers play different roles: the caption supplies scene context for semantic recall, the graph supplies event components for reranking, and the principle supplies a reusable physical constraint for verification. The question and explanation connect these representations to the particular training example.

The event graph is a JSON object containing an ordered list of event records. The extraction prompt favors one or two core events and permits at most three, rather than enumerating every object or every frame. An event's fields are listed in Table 6. The counterpart may be empty if none is visible; uncertain states are explicitly represented as `unknown`. A mostly static scene can still contain support, resting, or persistence relations.

Table 6: Event-record fields. State snapshots describe the visible subject; the counterpart supplies its interaction context.

Field	Content
`event_id`	A local event identifier.
`event_type`	One value from the event vocabulary, such as contact, falling, occlusion, or static persistence.
`subject`, `object`	Main moving/changing object and visible counterpart, such as floor, barrier, support, container, or occluder.
`phase`	`start`, `during`, `end`, or `unknown`.
`observed_states`	Before/during/after snapshots of motion, quantity, color, and visibility.
`state_change`	Attribute, initial and final values, and change type.
`visible_evidence`	A short description of the supporting visual observation.
`uncertainty`	`low`, `medium`, or `high`.

C Retrieval and Memory Selection

C.1 Text construction and event matching

For each event, the five component texts and weights are given in Table 7. Missing event types use `observed_object`; missing state-change attributes, types, endpoints, and phases use `unknown`. Subject and counterpart strings retain their available text. The visible-evidence component uses its text directly, without an extra field prefix.

Table 7: Textualization of the five event components. Bracketed values are substituted from a graph record.

Component	Encoded text	Weight
Type	`event type: <event_type>`	0.35
Participants	`subject: <subject>; object: <object>`	0.20
State change	`state change: <attribute> <change_type> from <from> to <to>`	0.25
Evidence	`<visible_evidence>`	0.15
Phase	`phase: <phase>`	0.05

C.2 Deduplication, diversity, and fallback

Candidates are ordered by descending graph score, descending semantic score for graph-score ties, and the stable QA key for remaining ties. A first pass retains only the highest-ranked candidate for each training video. A second pass greedily selects candidates with distinct rule families. If fewer than three families are available, the remaining positions are filled in ranked order from the unused, video-deduplicated candidates. Thus, video uniqueness is retained, while family diversity is a preference with a defined fill policy rather than an unconditional guarantee.

The preserved evaluation inputs include eleven IntPhys2 cases marked as having a missing or unusable evaluation graph, with preserved semantic-fallback selections. Their existing Top3 identities and order are retained rather than regenerating an event graph or dropping the cases. The same fallback selections are retained in the corresponding retrieval controls. An absent training graph receives the sentinel graph score -10^9, below all valid normalized-similarity scores; it is not replaced by a graph inferred from the answer label.

Algorithm 1: Physical-memory retrieval and verification

Input: current video, original task, frozen memory and encoder; preserved query representations and retrieval selections where available.

1. If this is a preserved fallback case, load its ordered Top3 and available saved query representation, then jump to Step 6. Otherwise, obtain the query caption and question-conditioned event graph.
2. Encode the caption–question document; recall 50 training QA entries.
3. Compute weighted event similarities and the directional graph score.
4. Order candidates by graph score, semantic score, and stable key; deduplicate training videos.
5. Select up to three distinct families, then fill any remaining positions from unused candidates in ranked order.
6. Render the selected questions, explanations, and principles in selection order, together with the query graph and original task.
7. Attach the current video, generate the benchmark-format answer, and apply the benchmark's parser.

Figure 3: Inference procedure. Dataset labels are used only after prediction for scoring. The fallback branch uses the saved selection rather than a new retrieval or graph-generation call.

Computational structure. Memory text and event-component embeddings can be precomputed. After semantic recall, reranking compares the query events only with events in 50 candidates. For $P = 5$ components, embedding dimension d, and event counts $|E_i|$ and $|E_j|$, the event dot-product work is proportional to $Pd|E_i|\sum_{j\in\mathcal{C}_i^{50}}|E_j|$. This describes the computation, not an end-to-end latency measurement: caption/graph construction and answer generation are separate costs.

D Baseline Adaptations and Ablation Controls

D.1 Backbones and generation settings

The compared answer backbones are Qwen3-VL-8B-Instruct, Cosmos-Reason2-8B, InternVL3_5-8B, and LLaVA-OneVision-2-8B-Instruct. Answer generation uses frozen weights, BF16, greedy decoding, and 1-FPS video sampling. The PhysMRV answer budgets are 256 tokens on ImplausiBench and IntPhys2 and 32 on GRASP. The selected Qwen and Cosmos adapters use TorchCodec. InternVL and LLaVA use Decord frames; the selected LLaVA adapter sets the maximum pixel budget to 200,704. Each adapter retains its model-native visual preprocessing. Generation budgets are upper bounds, not requirements to produce an explanation; the final-answer instructions still request the benchmark-format answer.

The MPNet retrieval runtime records PyTorch 2.8.0, Transformers 4.57.3, TorchCodec 0.7.0, and sentence-transformers 5.2.0. This specifies that retrieval/Qwen environment rather than implying that all baseline adapters share one software stack. Model-native adapters and baseline-specific stages are retained.

D.2 Compared methods

Direct. Direct receives the current video and original task/options, without retrieved examples, a generated evaluation event graph, or the PhysMRV verifier instruction. For IntPhys2, the complete binary task is placed in the user message, as in the adopted PhysMRV interface.

Vanilla CoT. CoT uses a single reasoning-augmented generation with a 4,096-token budget and an answer-block instruction. Its compatible extractor first accepts the strict answer block, then considers

explicitly stated answers with cosmetic formatting differences. A valid strict answer block takes precedence. Only when that strict check fails does the compatible fallback collect explicit selections and reject conflicting candidates. This extractor does not select an arbitrary number from the reasoning and is not applied to PHYSMRV.

CCoT. The video adaptation first generates a question-conditioned scene graph describing visible objects, attributes, and relations, with a 1,024-token budget. It then passes the raw graph text, video, and original task to the answer stage, using the 256/256/32 answer budgets. This graph is generated by the CCoT pipeline; it is not the shared PHYSMRV physical-event graph. Graph formatting or truncation is recorded without silently replacing the graph with a repaired output. CCoT receives no physical memory.

PhysNote. The adaptation constructs a three-level external note hierarchy from a frozen set of 500 training examples using Qwen, then reuses those notes for the four evaluation backbones. Note construction does not update backbone weights. The note memory is constructed from training questions, reference explanations, and video, without importing PHYSMRV principles or event graphs. At evaluation, `all-MiniLM-L6-v2` orders the note directories and the evaluation backbone selects a note node; all nodes remain available and selection may return null. The procedure allows up to three hypothesis, observation, and validation rounds. The selected evaluation configuration allocates 256 tokens to short stages, 2,048 to long stages, and 256/256/32 to the final answer. The 500 distinct training videos are sampled without replacement from sorted video IDs using Python `Random(42)`; one QA is then randomly chosen from each video's entries sorted by QA index and ID. The saved selection is included in the source bundle. An invalid node selection produces an empty final response. An invalid hypothesis, observation, or validation ends only that sample's current round, permitting a later round; a valid selection still reaches final-answer generation. The notes remain fixed during evaluation. This is a video-benchmark adaptation with its own memory construction and reasoning loop, not an exact rerun of the original paper's model/data setting.

CLIP MeanP and InternVideo2. These baselines select training videos using visual similarity rather than the caption–question/physical-event pipeline. They retain the PHYSMRV evaluation graph and answer verifier and use the fixed first QA for each retrieved video. CLIP uses ViT-B/32: each real frame embedding is normalized, the frame embeddings are averaged, and the resulting video vector is normalized again. InternVideo2 uses Stage2 1B at 224p with four-frame clips and its trained projection. Consecutive groups of four 1-FPS frames form clips; the final partial group repeats its last frame. Normalized clip embeddings are averaged with weights equal to the number of real frames in each group, then the video vector is normalized. Both methods sample the entire video without a total-frame cap and rank training videos by cosine similarity, breaking ties by video ID. They select Top3 without rule-family diversity. The representative QA has the smallest `qa_index`, then the smallest QA ID. The InternVideo2 visual implementation is taken from the preserved VideoICL source at revision `bde72fef0770`. Consequently, their comparison includes both visual ranking and the choice of QA evidence; it does not isolate a text-encoder-versus-video-encoder substitution with all memory identities fixed.

Method-specific parsing. InternVL Direct on ImplausiBench has a leading-hyphen normalization before the original option parser. CoT uses its compatible extractor. These exceptions are local to those methods. Original-parser accuracy and strict return-only compliance are distinct diagnostics, and the methods' different stage counts and token budgets should be retained when interpreting the comparisons.

Supplementary implementation map. The source bundle provides `appendix_assets/REPRODUCIBILITY.md` and a machine-readable manifest mapping each baseline to its selected configuration, exact prompt-building and parser source, and frozen-resource hashes. The PhysNote selection, notes and index are included, along with the MiniLM model reference and pinned VideoICL source link. This documents the selected interfaces and resources without treating historical outputs as a verification of every main-text score.

D.3 What each ablation changes

The seven controls use Qwen3-VL and preserve the original video, task, verifier wording, and benchmark answer constraints. Table 8 states the intervention independently of its numerical result. In particular, deleting an evidence block does not optimize a new prompt for the reduced input.

Table 8: Precise evidence and retrieval interventions. QA means the retrieved question and explanation.

Variant	Intervention relative to Full
w/o Memory	Delete the retrieved QA/principle block and its heading; keep the current event graph.
w/o Physical Reranking	Sort the same saved Top50 by semantic score, then apply video deduplication and family diversity/fill.
Shuffled Memory	Transfer complete three-memory blocks between different queries within the same benchmark; keep the recipient's task, video, and graph.
w/o Eval Graph	Hide only the current graph in the verifier; keep graph-based retrieval and the selected memories.
Rule Only	Keep the principles of the original selected Top3; remove their questions and explanations.
w/o Rule	Keep the original Top3 questions and explanations; remove the principles.
Graph-only Retrieval	Score all 19,708 QA candidates by event-graph similarity, without semantic Top50 filtering; keep QA and principles in verification.

The shuffled-memory control uses a within-benchmark derangement with seed 42. Each query receives another query's intact memory block, preserving the marginal distribution of blocks and the order of memories inside each block. The derangement depends on sample keys, not labels. In graph-only retrieval, graph-score ties use the stable QA key rather than semantic similarity; same-video QA entries therefore share a graph score without gaining QA-semantic discrimination. Video deduplication and family diversity/fill remain in place. The eleven IntPhys2 fallback selections are retained in the semantic-ranking and graph-only controls.