VideoMem: Enhancing Ultra-Long Video Understanding via Adaptive Memory Management

Anonymous ACL submission

Abstract

Ultra-long video understanding remains an open challenge, as existing vision language models (VLMs) falter on such content due to limited context length and inefficient long-term memory retention. To address this, recent works have attempted to construct external knowledge bases and corresponding retrieval agumented generation (RAG) systems, yet these incur enormous storage and computational overhead. In this paper, we propose **VideoMem**, a novel framework that pioneers models long video understanding as a sequential generation task via adaptive memory management. Specifically, VideoMem dynamically updates a global memory buffer, which adaptively retains critical information while discarding redundant content across the video timeline. To efficiently train VLMs for such long-term tasks, VideoMem integrates the Progressive Grouped Relative Policy Optimization (PRPO) algorithm, equipped with two core modules: Progressive State Propagation (PSP) adaptively retains valid current states, propagates them to the next rollout step, and gradually narrows the model's exploration space. Temporal Cascading Reward (TCR) further alleviates reward sparsity, improving sample utilization and accelerating convergence. Extensive experiments demonstrate that VideoMem significantly outperforms existing open-source models across diverse benchmarks for ultra-long video understanding tasks.

1 Introduction

Nowadays, Vision-Language Models (VLMs; (Liu et al., 2023; Li et al., 2023; Alayrac et al., 2022; Li et al., 2024a)) have attained remarkable advancements across a wide range of image-centric downstream tasks (Bai et al., 2025; Wang et al., 2025c). However, processing video content, particularly ultra-long videos, remains a prominent challenge for their core architectures (Wang et al., 2025b; Fu et al., 2025). This phenomenon stems from two fundamental bottlenecks: (i) **limited context length** (Chen et al.; Liu et al., 2024b), restricted by the quadratic computational cost of attention mechanisms; (ii) **inefficient memory modeling** (He et al., 2024; Song et al., 2024; Wang et al., 2025e), which results in catastrophic forgetting of early video content. These constraints impede the development of critical applications such as long-form video summarization and question-answering over extended timelines, both of which demand comprehensive and holistic video comprehension.

Existing long-video understanding approaches can be broadly categorized into three paradigms. Compression-based methods reduce input dimensionality through frame sampling or feature compression (Jin et al., 2024; Li et al., 2024b; Shen et al.; Wang et al., 2025a; Wu et al., 2025; Jin et al., 2026a), yet inevitably discard fine-grained temporal details. Static memory mechanisms employ pre-defined rules to manage long-term memory (Song et al., 2024; Zhang et al., 2024a; Wang et al., 2025d; He et al., 2024; Qian et al., 2024; Wang et al., 2025e), but lack adaptability to accommodate diverse video content and task-specfic requirements. RAG-based methods (Luo et al., 2024; Tian et al., 2025; Long et al., 2025) construct external knowledge bases to store full video information and design corresponding retrieval-augmented generation (RAG) systems, which however incur substantial storage and computational overhead.

To address these challenges, we propose VideoMem, a framework for ultra-long video understanding via adaptive memory management. This core component dynamically retains critical video details to mitigate forgetting while discarding redundant content to control resource consumption. In Figure 1, VideoMem pioneers framing long-video understanding as a sequential generation task conditioned on the query. Specifically, the long video is first split into segments and are then processed iteratively: at each step, the model

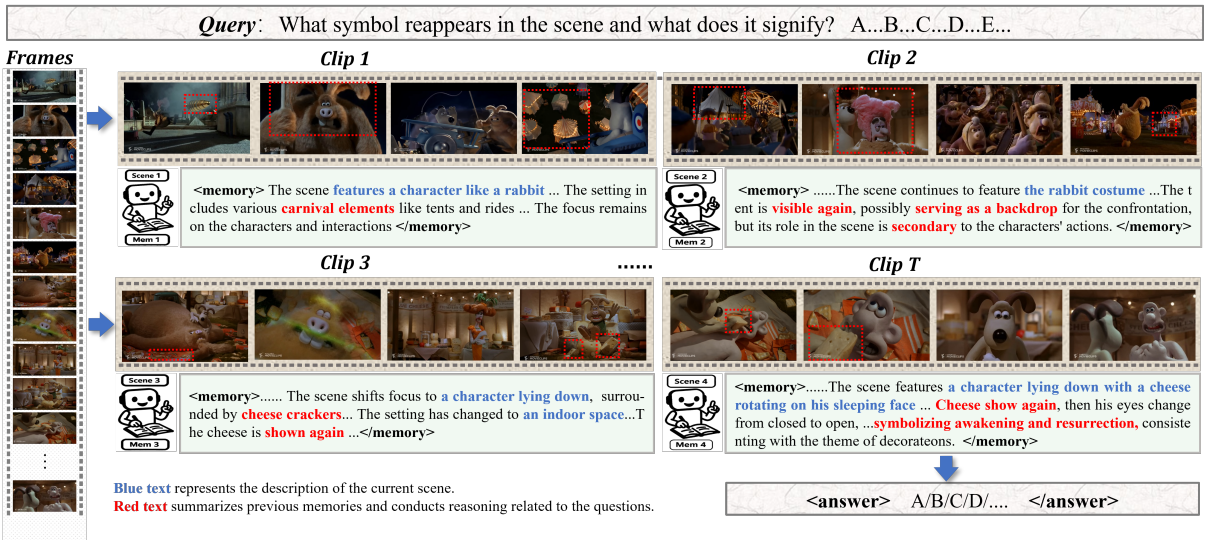

Figure 1: VideoMem demonstrates amazing ability to capture key information and perform long-term memory.

dynamically updates a global memory buffer using the current segment, while reserving key temporal details from prior segments. After processing the entire long-video, the model synthesizes all accumulated memories to generate the final answer.

To train this framework, reinforcement learning (RL) is a natural fit for such sequential tasks. However, directly applying prior methods (Shao et al., 2024; Jin et al., 2026b,c) to ultra-long videos processing is hindered by sparse and delayed reward signals, as well as prohibitive training computational costs. We thus propose Progressive Grouped Relative Policy Optimization (PRPO), an algorithm that enable efficient and effective RL training for this setting. In detail, PRPO reduces training overhead by progressively propagating valid states and narrowing the exploration space, namely progressive state propagation (PSP). Furthermore, it incorporates a temporal cascading rewards (TCR) mechanism to provides denser, stage-aware feedback, effectively mitigating the sparse reward problem inherent to long-term tasks.

To extensively validate our proposed framework, we conduct experiments on a comprehensive suite of standard and challenging long-video benchmarks, including VideoMME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2024), LVBench (Wang et al., 2025b), LongTimeScope (Zohar et al., 2025b). On one hand, leveraging adaptive memory management, VideoMem achieves state-of-the-art (SOTA) performance among 17 comparable open-source models across all long-video benchmark. On the other hand, PRPO excels in training efficiency: it outperforms GRPO by 3.1× in computational efficiency, cuts convergence steps by 30% (i.e., 0.7× the original), and yields notably improved training stability.

2 Related Work

Long video understanding poses challenges that stem from highly redundant visual information and inefficient temporal dependency modeling. Expanding context length is an intuitive solution, which directly enlarges model's temporal receptive field. Methods such as Gemini(Team et al., 2023), Kangaroo(Liu et al., 2024b), and LongVILA (Chen et al.) achieve this via efficient positional encoding or sliding-window. Token compression approaches such as VideoAgent (Fan et al., 2024), LongVU (Shen et al.), LLaVA-onevision (Li et al., 2024a), PLLaVA (Xu et al., 2024) and VideoXL-2 (Qin et al., 2025) attempt to reduce the sequence length via pooling mechanism or similarity computation, while tend to lose fine-grained details. Recent studies focus on building RAG systems for long videos. Ego-R1 (Tian et al., 2025) constructs a knowledge base in a multi-level hierarchy, and trains VLMs to retrieve from it. M3-Agent (Long et al., 2025) trains two models separately: one learns to store while the other learns to retrieve. Different from the above methods, we enhance long video understanding via adaptive memory management without external databases or handcrafted selection rules. More related work is described in the appendix A.

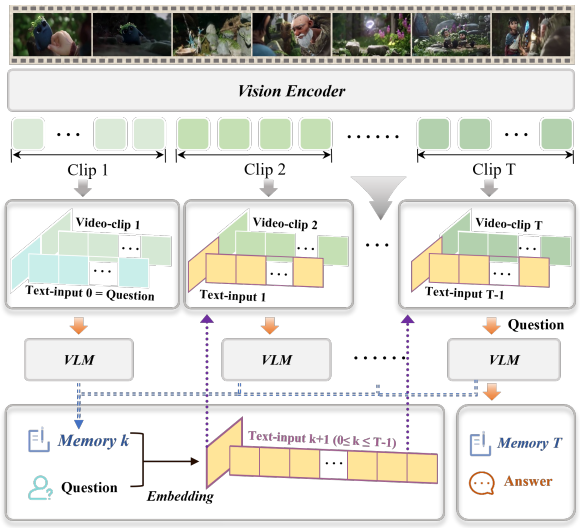

Figure 2: An overview of VideoMem framework. The video are divided into multiple clips and fed into the model iteratively. The question, video clips are concatenated with the memory of each turn's model output as input for the next turn.

3 VideoMem

In this section, we elaborate on our proposed framework VideoMem. We formulate ultra-long video understanding task as a sequential generation task and outline adaptive memory management, as shown in Figure 2. Then we design an automated pipeline to construct chain-of-memory data and perform cold start training for enhanced stability, as illustrated in Figure 3. Finally, we propose a novel RL framework, Progressive Grouped Relative Policy Optimization (PRPO), which comprises two core components: temporal cascading reward (TCR) and progressive state propagation (PSP).

3.1 Adaptive Memory Management

Ultra-long video understanding grapples with two inherently conflicting bottlenecks: redundant temporal information in ultra-long content demands compression to save computation yet risks losing critical details, while inefficient long-term memory modeling requires continuously retaining key information for holistic comprehension yet exacerbates redundancy and computational burden. To address these issues, we frame ultra-long video understanding as a query-conditioned sequential generation task augmented with adaptive memory management. This method dynamically retains key information, discards redundancy, and accumulates temporal context across video segments, enabling the model to holistically comprehend ultra-long videos without catastrophic forgetting or prohibitive costs.

Algorithm 1 Progressive Grouped Relative Policy Optimization (PRPO)

Require: Policy π_θ, video segments $V = \{V_1, \ldots, V_T\}$, question Q, ground-truth answer Ans, Group count G, token budget $L_{\max}$, coefficients α, β

Ensure: Updated policy π_θ

Initialize memory $M \leftarrow$ ""
Initialize trajectory buffer $\mathcal{B} \leftarrow \emptyset$
for each example in dataset **do**
  $Q, V, \text{Ans} \leftarrow$ example
  for $t = 1$ **to** T **do**
    $\{O_i\}_{i=1}^{G} \leftarrow \text{Rollout}(\pi_\theta, Q, V_t, M)$ {Generate G trajectories}
    $\{\hat{a}_i\}_{i=1}^{G} \leftarrow \text{ExtractAnswer}(Q, \{O_i\}_{i=1}^{G})$ {Provisional answer from <answer>}
    $\text{ConsReward}_i \leftarrow \mathbb{I}[\hat{a}_i = \text{Ans}]$ {1 if matches Ans, 0 otherwise}
    $\text{FmtReward}_i \leftarrow \mathbb{I}[\text{CheckFormat}(O_i)]$ {1 if <memory> $\cdots$ </memory> < answer> $\cdots$ </answer>, 0 else}
    $\text{MemPenalty}_i \leftarrow \max(0, \|\text{ExtractMemory}(O_i)\|_{\text{tokens}} - L_{\max})$
    $R_i \leftarrow \alpha \cdot \text{ConsReward}_i + \text{FmtReward}_i - \beta \cdot \text{MemPenalty}_i$
    $\{A_i\}_{i=1}^{G} \leftarrow \text{ADV}(\{R_i\}_{i=1}^{G})$ {Group-relative advantage}
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{(O_i, A_i)\}_{i=1}^{G}$ {Reserved for policy updates}
    if $t < T$ **then**
      $M \leftarrow \text{ExtractBestMemory}(\{O_i\}_{i=1}^{G}, \{A_i\}_{i=1}^{G})$ {Sample $m_j \sim p_j \propto A_j^{\text{rel}}$}
    end if
  end for
  Update π_θ via GRPO-style objective using $\mathcal{B}$
end for

In Figure 2, we first split a given ultra-long video into a sequence of equally long clips $\{V_1, V_2, ..., V_T\}$. We then iteratively feed these clips to the VLM one on at a time, processing clip V_t at step t. At each step t, the VLM takes three types of inputs: the current video clip V_t, the user's query Q, and a global memory buffer M_{t-1}. Given these inputs and the prompt detailed in Appendix C, the VLM retains critical temporal details (e.g., re-

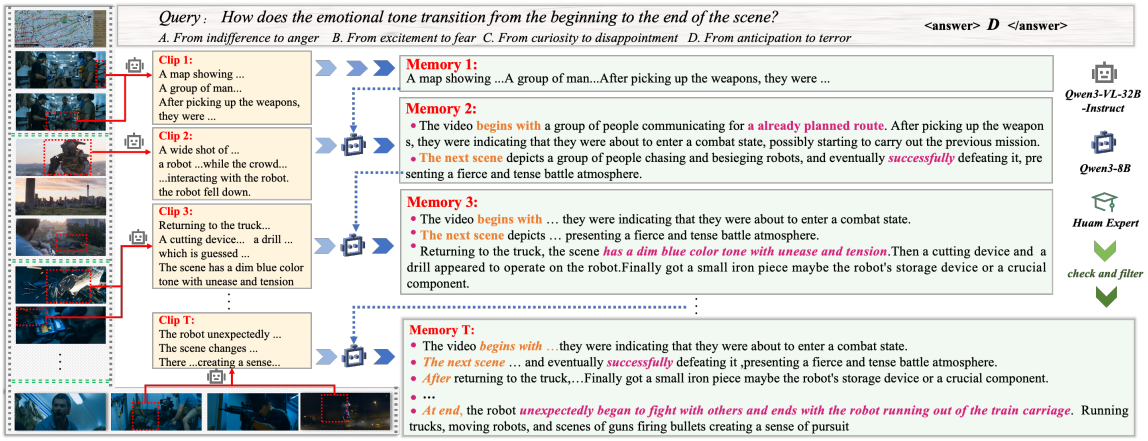

Figure 3: Chain-of-memory data construction pipeline. By combining multiple models, multi-stage generation, and human supervision, we ensure high-quality of video chain-of-memory pairs.

curring symbols, plot transitions) and prunes redundant content (e.g., static backgrounds) to form the updated global memory buffer M_t. This adaptive memory management balances current information with prior context, resolving the "forgetting vs. redundancy" dilemma. Finally, the global memory buffer M_T adaptively accumulates key information from all clips $\{V_1, ..., V_T\}$, enhancing the accuracy of the final answer.

This sequential generation task can be naturally formulated as a reinforcement learning (RL) problem, where the policy π_θ is optimized to generate high-quality trajectories. However, challenges arise in two key aspects: first, the ultra-long video setting exacerbates inherent reward sparsity and demands improved training efficiency; second, the model requires robust foundational memory capabilities to avoid unproductive exploration during RL training. Therefore, we detail two tailored training components in the following subsections: a cold-start stage and the Progressive Grouped Relative Policy Optimization algorithm.

3.2 Cold Start Stage

Following Deepseek-R1 (Shao et al., 2024), we introduce a cold start stage prior to RL to equip the base model with fundamental capabilities for compressing ultra-long videos and retaining their key information. By pre-aligning the model with basic memory management heuristics, it further accelerates the RL process, enabling more efficient discovery of high-value policies—thereby stabilizing and enhancing the subsequent training pipeline.

To automatically construct our chain-of-memory data, we first sample video–question–answer pairs from VideoMarathon (Lin et al., 2025) dataset. VideoMarathon is a comprehensive open-source long video training dataset, covering a wide range of video distributions with 1,573k multiple-choice questions and 1,726k open-ended questions. Each video is split into multiple equal long clips, which are sequentially fed into `Qwen3-VL-32B-Instruct` to generate semantic summaries. These summaries capture core narrative or factual content according to the context of specific video clips. Then, the clip-wise summaries are fed to `Qwen3-8B` (Yang et al., 2025) iteratively to filter out irrelevant information and generate coherent memories with manual checking. This yields a high-quality chain-of-memory dataset. Prompt templates are illurated in Appendix C.

Finally, we fine-tune the base model on these high-quality chain-of-memory instances using standard supervised instruction tuning to initialize its memory management ability. This cold start phase equips the model with the fundamental capability to construct coherent and concise memory representations.

3.3 Progressive Grouped Relative Policy Optimization

While Grouped Relative Policy Optimization (GRPO; (Shao et al., 2024)) lays a solid foundation for policy optimization without step-wise reward modeling, it encounters notable challenges when adapted to long-term tasks. The primary issues are the exponential expansion of the exploration space and exacerbated sparse reward signals inherent to extended temporal sequences. To mitigate these limitations, as illustrated in Figure 4, we propose

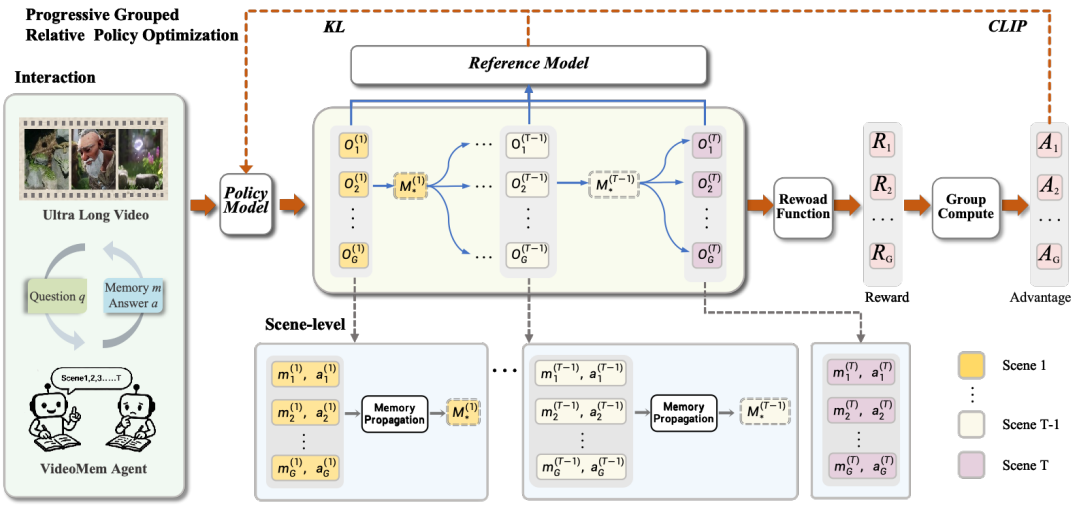

Figure 4: rollout and policy optimization pipeline of PRPO

Progressive Grouped Relative Policy Optimization (PRPO). This algorithm is integrated with two key innovations: temporal cascading reward (TCR) and progressive state propagation (PSP).

3.3.1 Temporal Cascading Reward

A major challenge in long-term RL is the sparsity and delay of reward signals: A reward for correctly answering a question grounded in early video content may only be observed after processing the entire sequence, rendering credit assignment highly non-trival. To address this, the temporal cascading reward (TCR) mechanism decomposes sparse final rewards into denser intermediate rewards at each step, delivering more frequent feedback to guide memory management. Specifically, at each step t, the model generates a provisional answer $\hat{a}_t$ conditioned on the current video clip V_t, memory buffer M_{t-1} and query Q. The immediate reward $\hat{R}_t$ for each trajectory is then computed as the sum of three components:

$$\hat{R}_t = \alpha \cdot R_{\text{cons}} + R_{\text{format}} - \beta \cdot \text{MemPenalty} \quad (1)$$

(i) $R_{\text{cons}} = \mathbb{I}[\hat{a}_t = a]$: For multiple-choice QA tasks, consistency reward is 1 if $\hat{a}_t$ exactly matches the ground-truth answer a and 0 otherwise. (ii) R_{format}: A binary reward, where the reward is equal to 1 if the output conforms to the required format $\langle\texttt{memory}\rangle \cdots \langle\texttt{memory}\rangle\langle\texttt{answer}\rangle \cdots \langle\texttt{answer}\rangle$, 0 otherwise. (iii) MemPenalty $= \max(0, \|M_t\| - L_{\max})$: Penalizes the length of memory buffer exceeding threshold $L_{\max}$ to prevent bloat while preserving efficiency.

Hyperparameters α and β control the relative importance of consistency and memory efficiency, respectively. These rewards are used to compute group-relative advantages $\hat{A}_t^{\text{rel}}$ following the same procedure as GRPO for policy update, detailed in Appendix E. This TCR mechanism can alleviate reward sparsity, improve sample utilization, and accelerate convergence.

3.3.2 Progressive State Propagation

Moreover, another key challenge in long-term RL is the exponential expansion of the exploration space as the sequence accumulates. Directly applying GRPO would incur prohibitive computational costs, especially for ultra-long video sequences. To mitigate this cost barrier, the Progressive State Propagation (PSP) mechanism acts as a memory-aware trajectory selection strategy that adaptively narrows the exploration space while retaining high-reward paths.

Let $M_t = \{m_1, m_2, \ldots, m_G\}$ denotes the set of global memory buffers extracted from the G trajectories at step t. Instead of retaining all G states, PRPO selects a single memory buffer M_t^* to propagate to the next turn generation, conditioned on the group-relative advantage $\hat{A}_i^{\text{rel}}$ of each trajectory:

$$M_t^* = \text{Sample}(M_t,\, p_j \propto \hat{A}_j^{\text{rel}}) \quad (2)$$

The sampling probability is softened with a temperature τ_t that decreases over time:

$$p_j = \frac{\exp(\hat{A}_j^{\text{rel}}/\tau_t)}{\sum_{k=1}^{G} \exp(\hat{A}_k^{\text{rel}}/\tau_t)} \quad (3)$$

where $\tau_t = \tau_0 \cdot \gamma^t$ ($\gamma < 1$) ensures high exploration early and focused exploitation later. The selected state M_t^* is concatenated with the query Q and the

upcoming clip V_{t+1} to form the input for the next turn rollout. This progressive narrowing yields two key benefits: (i) Computational efficiency, as prefilling complexity drops from $O(N \cdot G)$ to $O(N)$. Since only one memory path is cached per step, explained in detail in Section 4.8. (ii) Training stability, where early noise is filtered out and only trajectories with strong provisional consistency and format compliance propagate, thereby preventing error accumulation.

4 Experiment

4.1 Implementation Details

We choose Qwen3-VL-8B as our base model. During training, we divide each video into four equal segments, and extract 32 frames from each segment with max pixels not exceeding $128 * 32 * 32$. The number of rollouts G is set to 8. For temporal cascading reward, the hyperparameters are set to $\alpha = 1.0$, $\beta = 0.005$ and $L_{\max} = 1024$. We set global batch size to 16 and use the AdamW optimizer. All experiments are conducted on 16x PPU-ZW810E with 96GB memory. The complete training process requires 4800 PPU hours or 3200 A100 GPU hours. During evaluation, the input frame number is all set to 64, and the number of video segments in VideoMem is 4.

As for data, we sample 26k records from VideoMarathon dataset, and label them with chain-of-memory as the cold start training set. During RL training stage, we sample 147k multi choice QA pairs from VideoMarathon and 49k from LLaVA-Video-178K (Zhang et al., 2025c) dataset as training dataset, emphasis on long videos while also considering short videos. We adopt this data ratio, following the best results described in VideoMarathon (Lin et al., 2025).

4.2 Main Results

Table 1 presents our main results across all benchmarks. VideoMem significantly outperforms various strong baselines, achieving new state-of-the-art (SOTA) performance among comparable open-source models on diverse long-video benchmarks. Notably, our model delivers the largest performance gains on the most challenging long-form video benchmarks–LVBench, VideoMME and LongTimeScope–outperforming the base model Qwen3-VL-8B by 7.9%, 5.7% and 8.9%, respectively. These results are particularly revealing. While the base model exhibits strong static visual comprehension capabilities, it is not inherently optimized for the long-term reasoning required by ultra-long videos. It processes extended context as a flat sequence, struggling to effectively retain salient information over time. In contrast, VideoMem is explicitly trained via PRPO to act as a stateful agent: it learns an adaptive memory management policy that decides what critical information from the current clip V_t to integrate into memory M_t and what to discard. The substantial performance gains provide direct evidence that our adaptive memory mechanism successfully overcomes this core limitation of baseline models.

These findings offer a key insight: rather than relying solely on expanding context windows, which faces computational scaling challenges, a more efficient and effective paradigm is to reframe long-video understanding as a sequential generation problem. The fundamental reason for VideoMem's success lies in this shift—from training a model for static, one-shot comprehension to training an agent with a dynamic policy for information compression and propagation. Our work demonstrates that VLMs can be effectively optimized for long-term tasks using tailored RL strategies like PRPO, which successfully manage the sparse-reward and large-exploration-space challenges inherent to this formulation.

4.3 Core Components Analysis

We ablate the core components of VideoMem in Table 2. Removing progressive state propagation (PSP) drops performance by 2.1%, confirming its role in stabilizing long-term training. Removing temporal cascading reward (TCR) causes even larger degradation (4.2-4.9)%, leading to severe reward sparsity problem. And removing the memory penalty will also lead to a certain performance degradation.

4.4 Reward Mechanism Analysis

Table 3 further analyzes the impact of temporal cascading reward (TCR) compared to the single terminal reward (TR) used in GRPO. We observe significantly denser and more stable reward feedback when applying TCR. PRPO's temporal reward decomposition increases reward frequency by 4.0×, resulting in smoother policy gradients and faster convergence. We also notice that temporal cascading rewards stabilize training and prevent premature policy collapse. Empirically, models trained with TCR converge in about 70% of the

Models	Size	VideoMME		LongVideoBench	MLVU	LVBench	LongTimeScope
		Long	Overall				
Duration		30∼60 min	1∼60 min	23sec∼60 min	3∼120 min	30∼90 min	300min+
Proprietary Models							
GPT-4V	-	56.9	60.7	59.1	49.2	-	-
GPT-4o	-	65.3	71.9	66.7	64.6	64.4	-
Gemini-2.5-Pro	-	-	87.0	-	81.2	69.2	-
Seed1.5-VL	8B	-	77.9	74.4	82.1	64.6	-
Eagle2.5	8B	-	72.4	66.4	77.6	-	-
Open Source Models							
Video-LLaVA (Lin et al., 2024)	7B	38.1	40.4	39.1	47.3	-	17.6
ShareGPT4Video (Chen et al., 2024)	8B	37.9	43.6	39.7	46.4	-	-
LLaVA-Next-Video (Liu et al., 2024a)	7B	-	46.5	50.5	39.3	-	-
VideoLLaMA2 (Cheng et al., 2024)	7B	43.8	46.6	-	48.5	-	-
LongVA (Zhang et al., 2024b)	7B	47.6	54.3	-	56.3	36.2	-
LLaVA-Onevision (Li et al., 2024a)	7B	46.7	58.2	56.4	64.7	26.9	30.2
LLaVA-video (Zhang et al., 2025c)	7B	50.6	62.6	58.2	70.8	41.5	34.0
Qwen2.5-VL (Bai et al., 2025)	7B	51.6	65.1	54.7	70.2	45.3	40.7
VideoNSA (Song et al., 2025)	7B	-	-	60.0	51.8	-	44.4
Apollo (Zohar et al., 2025a)	7B	-	61.3	58.5	70.9	-	-
NVILA (Liu et al., 2025b)	8B	54.8	64.2	57.7	70.1	44.0	-
Flow4Agent(Liu et al., 2025a)	7B	54.2	64.7	60.4	71.4	-	-
VideoLLaMA3 (Zhang et al., 2025a)	7B	54.1	66.2	59.8	73.0	45.3	39.1
Qwen2.5-VL	72B	61.2	73.3	-	74.6	47.3	-
InternVL3.5 (Wang et al., 2025c)	8B	-	66.0	62.1	70.2	-	-
Video-XL2 (Qin et al., 2025)	8B	-	66.6	61.0	74.8	48.4	-
Qwen3-VL*	8B	58.6	67.9	59.1	71.6	50.6	36.2
VideoMem*	8B	**64.2**	**73.6**	**63.3**	**77.4**	**58.5**	**45.1**

Table 1: Evaluate results on various long video benchmarks, including VideoMME, LongVideoBench, MLVU, LVBench, LongTimeScope. All benchmarks adopt percentage accuracy as the evaluation metric. *Scores from our evaluation using 64 uniform frames, to ensure fair comparisons with same context length. All other results are from their original paper, with at least 64 frames of input.

Method	PSP	TCR	MP	MME	LTS
VideoMem	✓	✓	✓	**73.6**	**45.1**
w/o PSP.		✓	✓	71.5	43.0
w/o TCR	✓		✓	69.4	40.3
w/o MP	✓	✓		72.0	43.9
SFT				67.5	38.2

Table 2: Ablation of VideoMem core components.

steps required by TR, highlighting the efficiency of our dense feedback design.

Reward Type	Reward Density	Convergence Steps	LVBench
TR	1.0x	1.0x	54.6
TCR	4.0x	0.7x	58.5

Table 3: Ablation of TCR mechanism.

4.5 Per-Segment Accuracy Analysis

To verify the performance improvement of VideoMem truly stems from long-term memory, rather than a "general guess" based on early segments, we conducted the following experiments on LVBench.

We utilize the Temporal Cascaded Reward (TCR) mechanism to generate a provisional answer $\hat{a}_t$ at each step t, and record the corresponding "provisional accuracy". As shown in Table 4, the accuracy of the baseline model decreased from 50.3% to 44.8% with the addition of segments. This demonstrates catastrophic forgetting, where the model fails to integrate new information to correct early error. Conversely, VideoMem's accuracy showed non-trivially increases with each new segment added, steadily increasing from 51.2% to 58.5%. This strongly demonstrates that VideoMem continuously and proactively utilizes new segments to refine and improve its understanding, confirming the substantial contribution of later fragments.

Method	Seg1-1	Seg1-2	Seg1-3	Seg1-4
Qwen3-VL	50.3	48.5	46.2	44.8
VideoMem	51.2	55.4	57.1	58.5

Table 4: Seg1-N represents the accuracy of the model's prediction after N turns of input.

4.6 Cold Start Effiency

To verify the actual contribution of the cold start phase, we conducted corresponding ablation experiments in Table 5. We conducted a comparison on three representative long video bench-

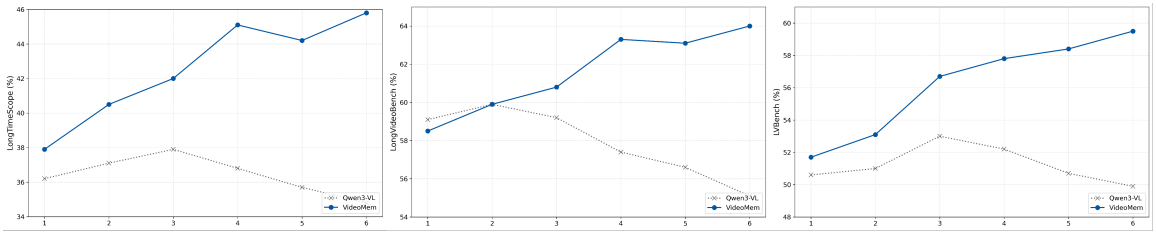

Figure 5: Inference Scalability

marks: VideoMME (long subset), LVBench, and LongVideoBench. All other training hyperparameters (batch size, number of frames, token budget, etc.) are consistent with the main experiment. Results show that cold start is a necessary training phase that can accelerate model convergence and avoid invalid rollouts.

Method	VideoMME	LVBench	LVB
Qwen3-VL	58.6	50.6	59.1
SFT	58.1	51.7	58.5
PRPO-only	59.8	53.4	59.5
ColdStart-only	61.9	54.4	60.7
VideoMem	**64.2**	**58.5**	**63.3**

Table 5: Cold Start Efficiency and Key Baseline Comparison. We compared five different experimental setups: (1) Qwen3-VL (untrained); (2) SFT (fine-tuned on general video instruction data); (3) PRPO-only (trained using only PRPO); (4) ColdStart-only (SFT training only on our chain-of-memory data); (5) VideoMem (trained with both cold start and PRPO).

4.7 Inference Scalability

While the number of video segments during training is fixed to $T = 4$, to ensure stable RL convergence and balanced reward density, the learned memory policy generalizes seamlessly to arbitrary segment counts during inference.

To verify this scalability, we conducted analysis on LongTimeScope, LongVideoBench and LVBench, shown in Figure 5. Fix the number of segments to 4 during training while evaluate its performance during inference with different numbers of segments $T = 1, 2, 3, 4, 5, 6$. Simultaneously, we compared the performance of the baseline model (Qwen3-VL) under the same settings.

The performance of baseline quickly reaches a bottleneck around $T = 3$, and shows a rapid decline, indicating that it cannot effectively utilize long term memory context. In contrast, VideoMem exhibits clear positive correlation scalability: as the number of segments increases, VideoMem is able to integrate more contextual information, and its accuracy steadily improves accordingly.

4.8 Computational Efficiency

VideoMem achieves a 3.1x speedup in training compared to vanilla GRPO while maintaining superior performance, shown in Table 6. The key lies in **PSP**: at each turn, a memory state is selected based on relative strengths to serve as the starting state for next turn, avoiding the expensive "prefill" operation for all G paths. This reduces the number of pre-filling calls from $O(N * G)$ to $O(N)$ (only once per segment). While decoding (i.e., token-level generation for each candidate trajectory) still requires $O(N * G)$ times, its computational overhead is far less than pre-filling. The evaluation results did not decrease but even improved, making our approach practical for large-scale deployment.

Method	Train Speed ↑	VideoMME	LVBench
Qwen3-VL	-	67.9	50.6
GRPO	1.0x	72.0	56.7
PRPO	3.1x	73.6	58.5

Table 6: Computational Efficiency

5 Conclusion

In this paper, we introduced VideoMem, a novel framework designed to address the significant challenges of ultra-long video understanding. We pioneeringly model this problem as a sequential generation task, and train VLMs to progressively compress video information and maintain effective long-term memory. We also propose Progressive Grouped Relative Policy Optimization (PRPO) algorithm, a novel reinforcement learning strategy tailored for this complex task. Experiments demonstrate that VideoMem achieves new state-of-the-art performance among comparable open-source models on a wide range of challenging long-video benchmarks, showing particularly strong gains on hour-long tasks.

Limitations

While VideoMem demonstrates state-of-the-art performance in ultra-long video understanding, we acknowledge certain limitations that present opportunities for future work. First, the framework's upper bound is inherently tied to the zero-shot spatial-temporal perception of the base VLM. In instances requiring deep implicit reasoning—such as inferring psychological states from subtle facial expressions—initial perception failures may lead to the discarding of critical cues during memory updates. However, VideoMem's highly modular architecture allows for the seamless integration of more powerful base models to naturally alleviate this bottleneck. Additionally, while our uniform video segmentation poses a slight risk of truncating continuous actions, the Progressive State Propagation (PSP) mechanism robustly bridges most temporal gaps, though future work could explore dynamic, event-driven chunking to further enhance semantic continuity.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, and 1 others. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, and 1 others. 2025. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, and 1 others. Longvila: Scaling long-context visual language models for long videos. In *The Thirteenth International Conference on Learning Representations*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

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

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. 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*, pages 24108–24118.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.

Hongbo Jin, Kuanwei Lin, Wenhao Zhang, Yichen Jin, and Ge Li. 2025. Videocurl: Video curriculum reinforcement learning with orthogonal difficulty decomposition. *arXiv preprint arXiv:2601.00887*.

Hongbo Jin, Siyi Xie, Jiayu Ding, Kuanwei Lin, and Ge Li. 2026a. Tir-flow: Active video search and reasoning with frozen vlms. *arXiv preprint arXiv:2601.06176*.

Hongbo Jin, Rongpeng Zhu, Jiayu Ding, Wenhao Zhang, and Ge Li. 2026b. Himac: Hierarchical macro-micro learning for long-horizon llm agents. *arXiv preprint arXiv:2603.00977*.

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

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13700–13710.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024b. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984.

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

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and Emad Barsoum. 2025. Unleashing hour-scale video training for long video-language understanding. *Preprint*, arXiv:2506.05332.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024b. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

Ruyang Liu, Shangkun Sun, Haoran Tang, Wei Gao, and Ge Li. 2025a. Flow4agent: Long-form video understanding via motion prior from optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23817–23827.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, and 1 others. 2025b. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. 2025. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. 2025. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. *arXiv preprint arXiv:2506.19225*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Forty-second International Conference on Machine Learning*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Enxin Song, Wenhao Chai, Shusheng Yang, Ethan Armand, Xiaojun Shan, Haiyang Xu, Jianwen Xie, and Zhuowen Tu. 2025. Videonsa: Native sparse attention scales video understanding. *arXiv preprint arXiv:2510.02295*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Han Wang, Yuxiang Nie, Yongjie Ye, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. 2025a. Dynamic-vlm: Simple dynamic visual token compression for videollm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20812–20823.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025b. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025c. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yun Wang, Long Zhang, Jingren Liu, Jiaqi Yan, Zhanjie Zhang, Jiahao Zheng, Xun Yang, Dapeng Wu, Xiangyu Chen, and Xuelong Li. 2025d. Episodic memory representation for long-form video understanding. *arXiv preprint arXiv:2508.09486*.

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

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

Peiran Wu, Zhuorui Yu, Yunze Liu, Chi-Hao Wu, Enmin Zhou, and Junxiao Shen. 2025. Marc: Memory-augmented rl token compression for efficient video understanding. *arXiv preprint arXiv:2510.07915*.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024a. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025b. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *Preprint*, arXiv:2504.09641.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025c. Llava-video: Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pages arXiv–2406.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, and 1 others. 2025a. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901.

Orr Zohar, Xiaohan Wang, Rui Li, Andrés Marafioti, Miquel Farré, Merve Noyan, Leandro von Werra, Serena Yeung-Levy, and Thomas Wolf. 2025b. Apollo2: Exploring the long-video frontier of large multimodal models.

A Extended Related Works

RL has become a cornerstone for LLMs to align with human preferences and enhance reasoning capabilities (Ouyang et al., 2022; Jin et al., 2026c). PPO (Schulman et al., 2017) optimizes policies through actor-critic frameworks, balancing exploration and exploitation to improve generation quality. Recent GRPO (Shao et al., 2024) further reduces computation overhead by eliminating critic model and estimating advantages relative to grouped samples. In video domain, Video-R1 (Feng et al., 2025) explores the R1 paradigm (Guo et al., 2025) to incentivize video reasoning. LongVILA-R1 (Chen et al., 2025) introduces a full-stack framework that scales up reasoning in long videos. TinyLLaVA-Video-R1 (Zhang et al., 2025b) explored the reasoning potential of small-scale VLMs. However, standard RL approaches struggle with exponential exploration space and reward sparsity in ultra-long video sequences (Jin et al., 2025; Lin et al., 2026). Our work incorporates temporal cascading rewards and progressive state propagation specifically designed for such challenges.

B Discussion

Generalizability of VideoMem to Open-ended Tasks. A key avenue for future work is to generalize our framework to open-ended QA tasks. The core sequential processing and adaptive memory management of VideoMem are task-agnostic. We hypothesize that by replacing the binary reward R_{cons} within TCR with a score from an LLM-as-a-judge, PRPO can effectively train VLMs to generate high-quality, free-form answers, confirming that VideoMem learns a deep, compressible understanding of video content.

Generalizability of PRPO to Other Long-term Tasks. PRPO algorithm is a general-purpose optimization strategy for sequential decision-making under sparse rewards. Its core components, PSP and TCR, address fundamental challenges in long-term tasks, independent of input modality. We believe PRPO holds significant promise for other domains, such as long-document QA, where text chunks replace video clips V_t. Furthermore, PRPO could be applied to multi-step agents, where PSP prunes the exponential exploration space of reasoning paths, and TCR provides denser feedback on intermediate sub-goals, effectively extending its application far beyond video understanding.

C Prompt Template

This section details the prompt templates used during VideoMem training and data construction.

Figure 6 shows the prompt template used for PRPO reinforcement learning training. This template instructs the model to perform two tasks after receiving the current video clip, the question, and previous memory.

Update Memory: Carefully watch the video clip, updating the memory with new information while retaining all relevant details from the previous memory.

Answer Question: Answer the question using only a single option letter (e.g., A, B, C, D, etc.) within the <answer> tag.

Figure 7 shows the prompt template used for data annotation in the Cold Start phase. As described in Section 3.2 of the main paper, this phase uses two models:

Qwen3-VL-32B-Instruct: As a multimodal model, it is responsible for watching the video clip and reading the question, then summarizing the main plot points relevant to the question.

Qwen3-8B: As a pure text model, it receives current content and historical memory from the multimodal model, and is responsible for piecing the two together into a complete memory while removing irrelevant information.

D Case Study

D.1 Good Case

VideoMem possesses global semantic awareness capabilities for videos several hours long and exhibits a certain degree of robustness against interference, as shown in Figure 8.

It performs exceptionally well in scene and environment recognition, person and object tracking, anomaly detection/event classification, and scene sequence/temporal understanding, thus validating the effectiveness of its temporal understanding and memory mechanisms.

D.2 Bad Case

The model exhibits errors on implicit reasoning tasks, such as inferring causal logic or mental states from subtle cues like facial expressions or third-party hints.

During long-term memory iterations, low-relevance estimations may cause the model to discard crucial implicit information, resulting in inference gaps.

User:
<vision start> <video pad> <vision end>
You are presented with a **question**, a **video** clip, and some **memory** (between the `<memory>` `</memory>` tags) about previous clips from the same video.
Please watch the video clip carefully and update the memory with new information that helps to answer the question, while retaining all relevant details from the previous memory.
<question> {question} </question>
<memory> {memory} </memory>
Update previous memory within the <memory> </memory> tags, then answer the question with only single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.
Assistant:

Figure 6: Prompt template for PRPO training

Specific reasons include:

Facial expression misinterpretation: The model fails to accurately infer a subject's personality or emotional state from specific facial cues, and also fails to understand the implied meaning behind expressions (e.g., recognizing that a smile may sometimes imply malice, not just happiness).

Implicit inference failure: The model fails to capture implicit narrative reasoning. For example, although video clips subtly foreshadow subsequent events through subtle cues (such as a dog entering a room, a man brandishing a gun, or a violent conflict between women), the model fails to capture this.

E Group Computation

In reinforcement learning, standard policy optimization algorithms (such as PPO (Schulman et al., 2017)) typically rely on a "critic" model to estimate the value of a state or the advantage of an action. However, training an accurate critic model for long sequence generation tasks is challenging in itself and incurs additional computational overhead.

Grouped Relative Policy Optimization (GRPO) (Shao et al., 2024) provides an efficient alternative. The core idea of GRPO is to "eliminate the critic model". It achieves this by generating a "group" of G candidate trajectories for the same input. It eschews the need for an accurate step-wise reward model by learning from relative advantage. Given a set of G trajectories $\{T_i\}_{i=1}^{G}$ generated from the policy, the GRPO objective is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{[q,\{o_i\}_{i=1}^{G}]} \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \Big(r_{i,t}, \text{clip}(r_{i,t}, 1-\varepsilon, 1+\varepsilon) \Big) \hat{A}_{i,t} - \beta \mathbb{D}_{KL} \big[\pi_\theta \| \pi_{\text{ref}} \big] \right\}, \tag{4}$$

where $r_{i,t}$ is the probability ratio:

$$r_{i,t} = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})} \tag{5}$$

and $\hat{A}_{i,t}$ denotes the group-relative advantage:

$$\hat{A}_{i,t} = \frac{r_{i,t} - mean(\mathbf{r}_t)}{std(\mathbf{r}_t)} \tag{6}$$

E.1 Memory Length Analysis

No Cold Start (Blue Line) 10: The memory length shows a clear decreasing trend across all chunks (chunks 0, 1, 2). This indicates that the model's memory was excessively large in the early stages of training (e.g., close to 600 tokens in chunk 2), and its learning process largely involved compressing information to avoid MemPenalty.

With cold start (orange line): the memory length trend is more stable and shows a gentle upward trend. This indicates that the model has mastered how to concisely represent information through the SFT cold start phase. Therefore, during the RL training phase, it can confidently add more relevant contextual information to memory without triggering penalties. This reflects a healthier and more efficient memory management strategy.

Qwen3-VL-32B-Instruct

User:
<vision start> <video pad> <vision end>
Watch this video carefully and read the question about it.
Then summarize the main plot in the video that relates to the question.
<question> {question} </question>
Assistant:

Qwen3-8B

User:
<memory> {memory} </memory>
<content> {content} </content>
<question> {question} </question>
You are given a memory of a past video clip, summary about the current video clip, and a question about the entire video.
Please piece the current information together with your past memory to form a complete recollection.
This memory should retain the main information from the video to answer the question, while removing irrelevant information.
Assistant:

Figure 7: Prompt template for cold start annotation

F Benchmark

We extensively tested the performance of VideoMem on various benchmarks, These benchmarks comprehensively evaluate whether VideoMem can effectively preserve video memory from highly redundant information.

VideoMME (Fu et al., 2025) includes 900 videos of varying lengths and 2,700 manually annotated multiple-choice questions. The video lengths include short (2min), medium (2-30min), and long (30-60min). It distinguishes from prior benchmarks through four key features: diversity in video types, duration in temporal dimension, breadth in data modalities and quality in annotations.

LVBench (Wang et al., 2025b) is a benchmark specifically designed to assess the ability of VLMs to understand ultra-long videos. It contains 1549 question-answer pairs with an average duration of approximately 68 minutes. It is designed to challenge multimodal models to demonstrate long-term memory and extended comprehension capabilities.

LongVideoBench (Wu et al., 2024) is a benchmark for long video retrieval and reasoning, containing 6,678 manually annotated multiple-choice questions and 17 fine-grained categories.

LongTimeScope (Zohar et al., 2025b) is an open-source benchmark designed to measure how well vision-language models understand long videos, by adding short “needle” clips into videos ranging from 1 minute to 8 hours.

MLVU (Zhou et al., 2024) is also a benchmark for long video understanding, containing nine different categories with video lengths ranging from 3 minutes to 2 hours. It presents the following critical values: the substantial and flexible extension of video lengths, the inclusion of various video genres, and the development of diversified evaluation tasks.

G Manual Screening in Cold Start

In our data building process, the standard operating procedure for human expert review aims to ensure high-quality "chain-of-memory" data. Our human experts rigorously verify that the memories generated by Qwen3-8B simultaneously meet four criteria:

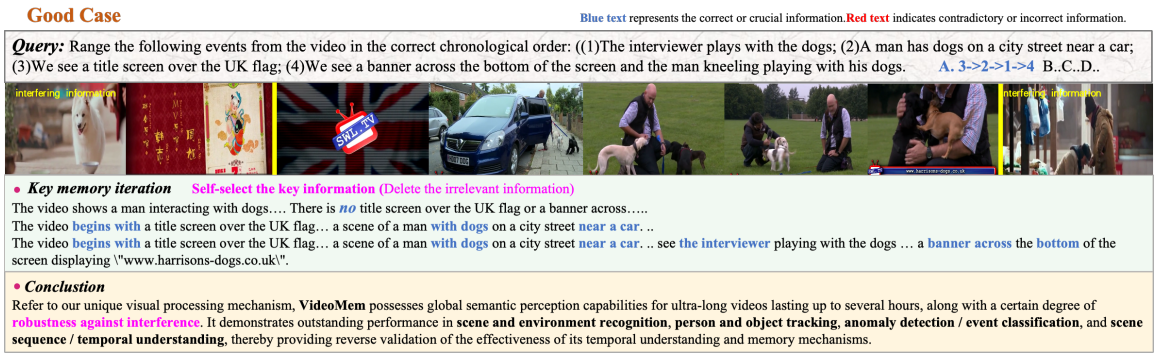

Figure 8: Good Case

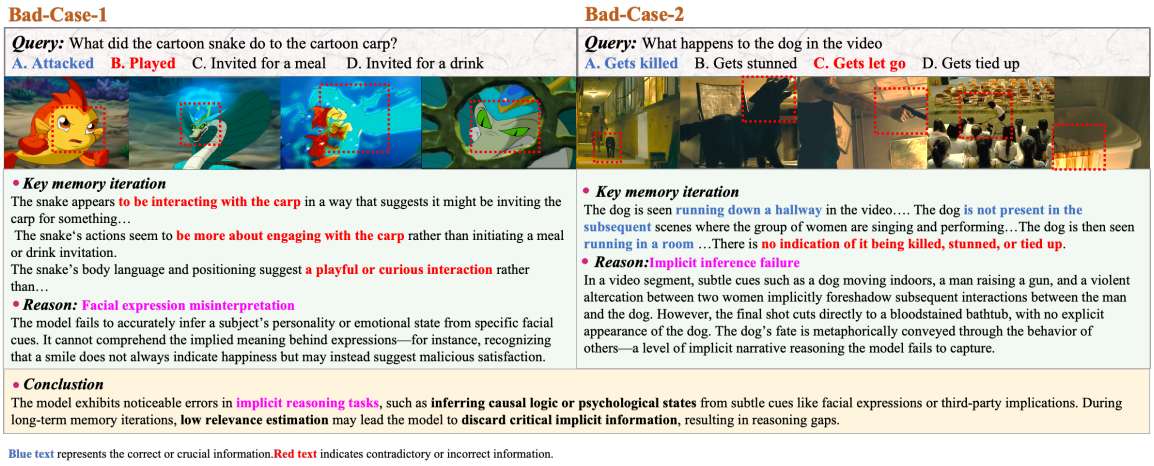

Figure 9: Bad Case

`Factual Consistency`: Ensure the integrated memories faithfully reflect the video clip summaries provided by Qwen3-VL-32B-Instruct, without distorting or omitting key facts.

`Query Relevance`: Verify that the retained memories closely revolve around the "key information needed to answer the question," filtering out irrelevant details.

`Conciseness (Low Redundancy)`: Ensure the model successfully removes irrelevant information or redundant content such as static backgrounds, retaining only core information.

`Coherence`: Ensure that the logical and temporal order of memories is "coherent" when iterating across segments, forming a complete context.

H Ethical Considerations

We acknowledge that the long video datasets used to train and evaluate this model (e.g., the Video-Marathon dataset) may contain copyrighted content. Our research focuses primarily on the technical challenges of developing efficient video understanding and adaptive memory management; the model itself is not intended to generate or reproduce copyrighted material.

However, we recognize that the potential for using such techniques on unauthorized copyrighted content is a significant societal issue. We advocate for prioritizing the use of datasets that have been properly licensed or belong to the public domain in future research. Furthermore, exploring technologies or norms to ensure "fair use" and content attribution in practical applications is also a direction worthy of future attention.

I Open Source Plan

This research plan is to release complete open-source resources on GitHub (to be released within one month of acceptance) and Hugging Face Hub (to be released within 1 week of acceptance) after the paper is officially accepted: the GitHub repository contains the core code (adaptive memory management, PRPO algorithm, training/evaluation scripts), supporting documentation (environment configuration yml, step-by-step reproduction guide, low-memory solutions), and sample data; the Hug-

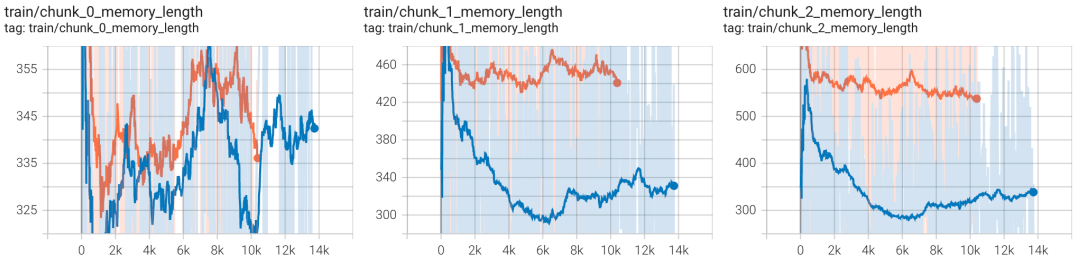

Figure 10: Memory length across segments

ging Face Hub repository will release the cold start model and the final VideoMem model weights (Safetensors format); optimization content will be added within one month of acceptance, and issues will be maintained weekly for three months (later converted to monthly). The code is licensed under the MIT license, and the model follows the Qwen3-VL license. The open-source resources are for academic research only, and unauthorized copyright or privacy-related misuse is prohibited, aiming to ensure the reproducibility and reusability of the research.