MVP: Enhancing Video Large Language Models via Self-supervised Masked Video Prediction

Anonymous ACL submission

Abstract

Reinforcement learning based post-training paradigms for Video Large Language Models (VideoLLMs) have achieved significant success by optimizing for visual-semantic tasks such as captioning or VideoQA. However, while these approaches effectively enhance perception abilities, they primarily target holistic content understanding, often lacking explicit supervision for intrinsic temporal coherence and inter-frame correlations. This tendency limits the models' ability to capture intricate dynamics and fine-grained visual causality. To explicitly bridge this gap, we propose a novel post-training objective: Masked Video Prediction (MVP). By requiring the model to reconstruct a masked continuous segment from a set of challenging distractors, MVP forces the model to attend to the sequential logic and temporal context of events. To support scalable training, we introduce a scalable data synthesis pipeline capable of transforming arbitrary video corpora into MVP training samples, and further employ Group Relative Policy Optimization (GRPO) with a fine-grained reward function to enhance the model's understanding of video context and temporal properties. Comprehensive evaluations demonstrate that MVP enhances video reasoning capabilities by directly reinforcing temporal reasoning and causal understanding.

1 Introduction

Reinforcement learning with Verifiable Rewards (RLVR) has significantly enhanced the reasoning capabilities of Multimodal Large Language Models (MLLMs) (Wu et al., 2025; Xing et al., 2025; Feng et al., 2025). Motivated by this success, recent research has increasingly attempted to transfer RL paradigms to Video Large Language Models (VideoLLMs) (Yan et al., 2025; Park et al., 2025; Fu et al., 2025b). Currently, the approaches in the video domain involves leveraging algorithms like Group Relative Policy Optimization (GRPO) (Shao et al., 2024) paired with diverse reward functions. These methods typically conduct training on standard tasks such as video question answering (videoQA) (Fu et al., 2025a; Hu et al., 2025a), video captioning (Zhang et al., 2025a), and video grounding (Gao et al., 2017; Lei et al., 2021) to bolster the models' capabilities.

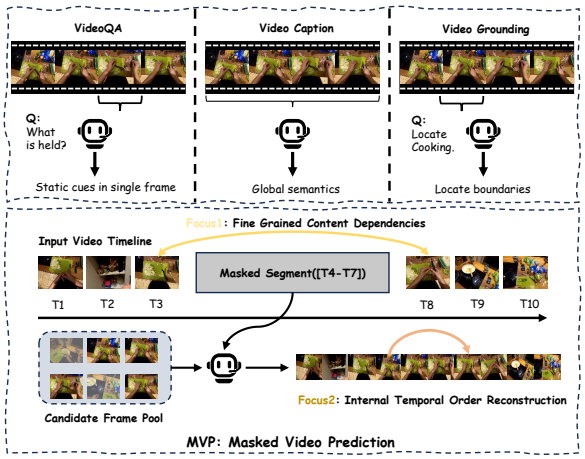


Figure 1: **Comparison between MVP and other tasks**. MVP compels the model to attend to both contextual content information within the video and the temporal relationships between frames.

Despite these improvements, a critical limitation persists: most existing methods lack explicit supervision for understanding video temporal properties and inter-frame correlations. This deficiency hinders the models from fully grasping the intrinsic temporal nature and frame-to-frame relationships within videos. Moreover, obtaining high-quality training data that effectively captures video dynamics is challenging. Prevalent training tasks, such as videoQA, video captioning and video grounding tend to prioritize object perception, largely neglecting the **dynamic temporal properties** and the **intricate dependencies** between frames.

To address these limitations, we propose a novel post-training objective: **Masked Video Prediction (MVP)**. Analogous to masked token prediction in BERT (Devlin et al., 2019) like models, MVP re-

quires the model to reconstruct a masked video segment, compelling it to explicitly attend to sequential logic and temporal context. We formulate this task as a multiple-choice problem where a continuous sequence of frames is masked and mixed with intra-video distractors. The MVP task essentially consists of two relatively independent subtasks: frame selection and temporal ordering. The model must select the correct frames and arrange them in the precise chronological order to fill the gap. Given the significant distributional shift between this task and standard pre-training data, direct Supervised Fine-Tuning (SFT) risks degrading the model's prior knowledge. Consequently, we adopt a Reinforcement Learning approach, utilizing Group Relative Policy Optimization (GRPO) with a fine-grained reward mechanism. Crucially, rather than relying on a simple binary accuracy metric, our reward structure differentiates between correct frame selection and correct temporal ordering. This dual-reward strategy incentivizes the model to simultaneously refine its understanding of both video content and intrinsic temporal properties. Additionally, we introduce a scalable MVP data synthesis pipeline capable of transforming arbitrary video corpora into training samples, enabling the generation of virtually unlimited data to meet large-scale training demands.

To rigorously validate the effectiveness and generalizability of the Masked Video Prediction task, we conduct comprehensive experiments across multiple base models and investigate the impact of data scaling by training with varying sample sizes. Our evaluation spanned six diverse and challenging benchmarks: VideoMME (Fu et al., 2025a), LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2025c), MLVU (Zhou et al., 2024), Video-Holmes (Cheng et al., 2025), and TempCompass (Liu et al., 2024). Empirical results consistently demonstrate that the MVP objective significantly enhances the models' video reasoning capabilities, proving its efficacy in reinforcing temporal reasoning and contextual grasp.

In summary, our main contributions are as follows:

- We propose **Masked Video Prediction**, a novel post-training objective that compels VideoLLMs to master temporal logic, accompanied by a scalable data synthesis pipeline that generates unlimited training samples from arbitrary video corpora.
- We design a **fine-grained visual supervision reward function** within the GRPO framework. This mechanism distinguishes between frame selection and temporal ordering, providing precise feedback that ensures stable and effective model training.
- Extensive experiments across multiple base models and six benchmarks demonstrate that MVP significantly enhances video reasoning capabilities, proving the effectiveness and generalizability of our approach.

2 Related Work

2.1 Video Large Language Models

Video Large Language Models (VideoLLMs) have emerged as powerful tools for video understanding (Bai et al., 2025a; Wang et al., 2025d; Zhang et al., 2025b; An et al., 2025; Team et al., 2025; Guo et al., 2025; Wang et al., 2025a; Yang et al., 2025), building upon the strong abilities of Large Language Models (LLMs). These models have achieved impressive performance in tasks like video question-answering (Fu et al., 2025a; Zhao et al., 2025; Hu et al., 2025b) and captioning (Zhang et al., 2025a) by enabling comprehensive content interpretation. Representative works have introduced various mechanisms to enhance this capability: BOLT (Liu et al., 2025) normalizes the similarity scores between video frames and the input question into a probability distribution, and then selects frames via inverse transform sampling to reduce the input length. SlowFast-LLaVA (Xu et al., 2025) fits the two-stream Slow-Fast mechanism into a streamlined training pipeline to minimize input token usage while maintaining the completeness of video information. These advancements highlight the potential of MLLMs in advancing video understanding.

2.2 Video Reinforcement Learning

Recent breakthroughs by OpenAI-o1 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025) have highlighted the efficacy of Reinforcement Learning (RL) in elevating the reasoning capabilities of Large Language Models (LLMs). Following this trend, RL techniques such as DPO (Rafailov et al., 2023) and GRPO (Shao et al., 2024) have been adapted for MLLMs and VideoLLMs to enhance visual reasoning through verifiable reward mechanisms (Yan et al., 2025; Feng et al., 2025; Xing et al., 2025; Park et al.,

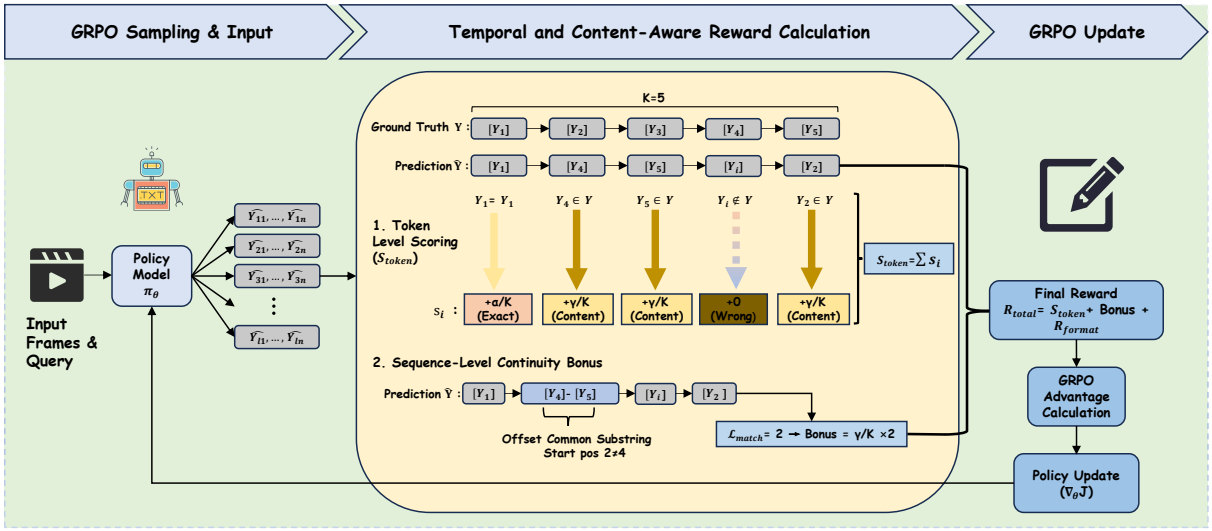


Figure 2: Illustration of GRPO with temporal and content-aware rewards.

2025; Wu et al., 2025; Tao et al., 2025; He et al., 2025; Wang et al., 2025b). Video-R1 (Feng et al., 2025) employs GRPO to improve implicit temporal and spatial reasoning by rewarding the model to utilize temporal information in vidoes. TSPO (Tang et al., 2025) trains an event-aware temporal agent through reinforcement learning to effectively select frames, thereby enhancing the model's understanding of long videos. Despite achieving a certain degree of effectiveness, these methods overlook the supervision required for the model to understand fine-grained temporal properties and inter-frame correlations in videos.

3 Method

In this section, we present the proposed approach in detail. We first provide a formal formulation of the Masked Video Prediction (MVP) task. Next, we describe the scalable data synthesis pipeline designed to transform arbitrary video sources into high-quality training samples, as illustrated in Fig. 3. Finally, we detail the reinforcement learning framework (Fig. 2), specifically focusing on the design of fine-grained reward functions tailored to capture temporal logic and sequential dependencies.

3.1 Masked Video Prediction

The Masked Video Prediction (MVP) task can be conceptualized as a "visual cloze test" for videos. It requires the model to reconstruct a missing video segment by selecting the correct frames from a candidate pool and arranging them in the correct chronological order based on the surrounding context. Formally, given a video sequence $V = \{f_1, f_2, ..., f_N\}$, a continuous segment V_{target} is masked, leaving the observable context $V_{context} = V \setminus V_{target}$. The masked segment is decomposed into K ordered key frames, forming the positive set $\mathcal{S}_{pos} = \{p_1, ..., p_K\}$. These are mixed with a set of hard negative distractors $\mathcal{S}_{neg}$ sampled from the same video to form a shuffled candidate pool $\mathcal{C} = \text{Shuffle}(\mathcal{S}_{pos} \cup \mathcal{S}_{neg})$. The objective is to predict a sequence of indices $Y = \{y_1, ..., y_K\}$ that selects the correct frames from $\mathcal{C}$ and arranges them in their original temporal order, such that the reconstructed sequence logically bridges the gap in $V_{context}$.

3.2 Scalable Data Synthesis Pipeline

To systematically investigate the impact of MVP on video understanding capabilities, we design a scalable data synthesis pipeline capable of generating MVP training samples from arbitrary video sources as shown in figure 3. Our objective is to mask semantically significant segments within continuous video streams, thereby necessitating the model to capture underlying temporal logic and long-range sequential dependencies.

Given the inherent redundancy in video data, where consecutive frames often exhibit high visual similarity, a naive frame selection strategy would result in information-poor samples. To mitigate this, we employ a visual de-duplication strategy. For a given video, we first load frames at 1 FPS. Starting from a randomly selected timestamp t_{start}, we aim to select a sequence of N distinct frames. Let $f_{current}$ denote the most recently selected frame. We iterate through subsequent frames

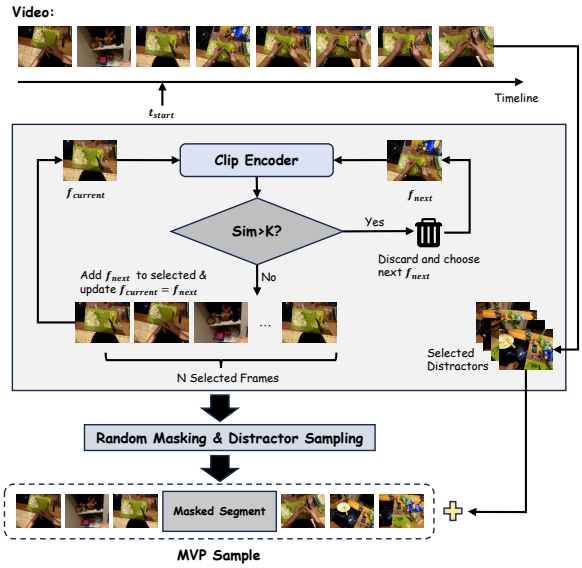


Figure 3: **Scalable Data Synthesis Pipeline.** We aim to ensure that the MVP samples contain minimal redundancy, enabling the model trained on this data to effectively learn temporal relationships in videos and dependencies between frames.

f_{next} and compute their visual similarity using a CLIP (Radford et al., 2021) encoder $\phi(\cdot)$:

$$s = \phi(f_{curr}) \cdot \phi(f_{next})^\top \tag{1}$$

If $s > \kappa$ (where κ is a pre-defined threshold), f_{next} is discarded as redundant. We continue this process until we find a frame where the similarity falls below κ, at which point it is added to our selected set, and the process repeats until N frames are collected.

From this selected sequence of N frames, we randomly mask m frames to serve as the prediction targets. To construct a challenging candidate pool, we randomly sample l distractor frames from the temporal vicinity (before or after) of the selected N-frame segment within the same video. These distractors are mixed with the m target frames to form the final shuffled candidate set.

This pipeline is universally applicable to any video corpus, allowing for the effortless generation of large-scale datasets. Furthermore, by varying the starting timestamp and random masking, a single video can yield multiple distinct training samples, maximizing data utilization and providing diverse temporal contexts

3.3 GRPO with Temporal and Content-Aware Rewards

3.3.1 Preliminary: GRPO

Group Relative Policy Optimization (GRPO) operates as a highly efficient reinforcement learning framework that explicitly eliminates the need for a separate value function critic, thereby significantly reducing both memory usage and computational overhead. For each input query q, GRPO samples a cohort of G outputs $\{o_1, \ldots, o_G\}$ from the current policy $\pi_{\theta_{\text{old}}}$. The advantage A_i for each individual output o_i is then derived by normalizing its reward r_i using the mean and standard deviation of the rewards within that group:

$$A_i = \frac{r_i - \text{mean}(\{r_j\})}{\text{std}(\{r_j\}) + \epsilon} \tag{2}$$

The objective function maximizes the clipped surrogate objective, regularized by a KL-divergence term:

$$\mathcal{J} = \frac{1}{G}\sum_{i=1}^{G}\left(\mathcal{L}_i^{clip} - \beta D_{KL}(\pi_\theta || \pi_{ref})\right) \tag{3}$$

where $\mathcal{L}_i^{clip} = \min(\rho_i A_i, \text{clip}(\rho_i, 1 \pm \epsilon) A_i)$ and $\rho_i = \pi_\theta(o_i|q)/\pi_{\theta_{old}}(o_i|q)$.

3.3.2 Reward Design

To precisely guide the model in mastering both visual content identification and temporal reasoning, we design a hierarchical reward function. Given the ground truth sequence $Y = \{y_1, \ldots, y_K\}$ and the predicted sequence $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_K\}$, the total correctness reward $R_{correct}$ is composed of token-level matching scores and sequence-level continuity bonuses.

Token-Level Scoring. First, we evaluate each predicted item $\hat{y}_i$ at position i to determine if the model has correctly identified the content and its temporal placement. The scoring function $s(i)$ assigns full credit for exact matches (content + position) and partial credit for content retrieval without correct ordering:

$$s(i) = \begin{cases} \alpha/K, & \hat{y}_i = y_i \\ \gamma/K, & \hat{y}_i \in Y \setminus \{y_i\} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Here, we set $\alpha > \gamma$ to strongly incentivize correct temporal ordering, and γ acknowledges the model's ability to recognize valid visual content even if the timestamp is misplaced. The total token score is $S_{token} = \sum_{i=1}^{K} s(i)$.

Sequence-Level Continuity Bonus. To further enforce temporal logic, we detect the **Common**

Substrings (length ≥ 2) between $\hat{Y}$ and Y. We apply a continuity bonus mechanism that rewards preserved temporal substructures. Specifically, for all matching substrings that do not start at the correct absolute position (i.e., offset matches), we provide an additional reward. This encourages the model to capture relative temporal order even when global alignment is imperfect. Let $\mathcal{L}_{match}$ be the sum of lengths of such valid substrings. The final reward combines the token-level accuracy with this structural bonus:

$$R_{correct} = S_{token} + \frac{\gamma}{K} \times \mathcal{L}_{match} \tag{5}$$

Through this granular reward design, we encourage the model to incrementally learn both content identification and temporal sequencing, ultimately fostering more robust video reasoning skills.

Format Reward. To encourage the model to perform explicit reasoning before generating the final answer, we introduce a format-based reward that constrains the structure of the model output. Specifically, the model is required to enclose its intermediate reasoning process within `<think>` tags and to present the final prediction within `<answer>` tags. Based on this requirement, the format reward R_{format} is defined as an indicator function that returns 1 if and only if the generated output strictly adheres to the prescribed structure, and 0 otherwise. By explicitly enforcing this structural constraint, the model is compelled to externalize its reasoning process prior to producing an answer, thereby facilitating a clear and interpretable *think-before-answer* mechanism.

We formulate the final training objective as a composite reward function that jointly accounts for both formatting compliance and task-specific correctness. The total reward is defined as a weighted sum of the format reward and the correctness reward:

$$R_{\text{total}} = \beta R_{\text{format}} + (1-\beta) R_{\text{correct}}, \tag{6}$$

where $\beta \in [0, 1]$ is a balancing coefficient that controls the relative importance of enforcing the reasoning structure versus optimizing task performance. This compound reward formulation jointly promotes task accuracy and structural compliance, resulting in outputs that are both accurate and interpretable.

3.3.3 Policy Optimization

We integrate the proposed fine-grained reward function R_{total} into the GRPO framework to optimize the VideoLLM policy π_θ. For each masked video query q, the model generates a group of G candidate sequences $\{\hat{Y}_1, \ldots, \hat{Y}_G\}$ from the current policy $\pi_{\theta_{old}}$. We compute the reward $r_i = R_{total}(\hat{Y}_i, Y)$ for each candidate using Eq 6, which explicitly values both frame identification and temporal sequencing.

The advantages A_i are then derived using the group statistics as defined in Eq 2. Finally, the policy is updated by maximizing the objective $\mathcal{J}$ in Eq 3. By directly optimizing for R_{total} within this group-relative formulation, the model is incentivized to internalize the underlying temporal logic, effectively aligning its generation probabilities with temporally coherent video reasoning pathways without the computational overhead of a value network.

4 Experiments

4.1 Experimental Settings

4.1.1 Benchmarks

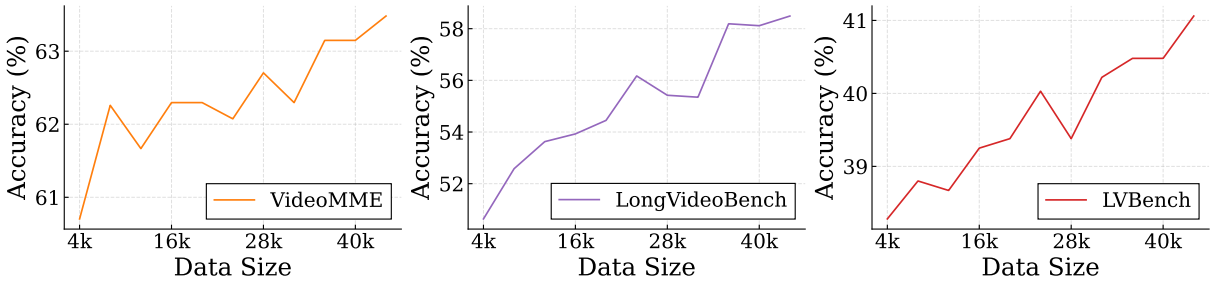
We conduct a comprehensive evaluation of MVP across six video benchmarks: VideoMME (Fu et al., 2025a), LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2025c), MLVU (Zhou et al., 2024), Video-Holmes (Cheng et al., 2025) and TempCompass (Liu et al., 2024). These benchmarks were selected to cover a broad spectrum of video reasoning capabilities. VideoMME and MLVU primarily assess general VideoQA performance. VideoMME comprises videos spanning diverse themes and categorized into short, medium, and long durations. Each video is paired with three questions, enabling a comprehensive assessment of the model's video question-answering capabilities. MLVU consists of nine subtasks, designed to evaluate model performance from multiple distinct perspectives. LVBench and LongVideoBench focus on long-video understanding, inlcuding extremly long videos ranging up to two hours. Video-Holmes serves as a reasoning benchmark, specifically targeting complex video reasoning abilities, while TempCompass evaluates the model's proficiency in temporal understanding.

4.1.2 Implementation Details.

We conduct GRPO training for MVP on three base models: Qwen2.5-VL-7B-Instruct (Bai et al.,

Model	Frames	VideoMME	LongVideoBench	LVBench	MLVU	Video-Holmes	TempCompass
Qwen2.5-VL-72B	768	73.3	60.7	47.3	74.6	–	74.8
Qwen3-VL-235B-A22B	768	79.0	–	63.6	83.8	–	–
InternVL3.5-241B-A28B	-	72.9	67.1	–	78.2	–	–
Jigsaw-7B*	64	61.2	53.2	36.1	61.4	36.9	69.0
	128	62.4	56.2	38.0	64.6	37.6	–
Qwen2.5-VL-7B-Instruct*	64	60.3	48.0	32.9	57.1	33.4	66.0
	128	60.9	51.4	37.5	63.0	35.0	–
+MVP (Ours)	64	$61.2^{\uparrow 0.9}$	$55.9^{\uparrow 7.9}$	$38.2^{\uparrow 5.3}$	$62.2^{\uparrow 5.1}$	$35.9^{\uparrow 2.5}$	$68.5^{\uparrow 2.5}$
	128	$63.5^{\uparrow 2.6}$	$58.6^{\uparrow 7.2}$	$41.1^{\uparrow 3.6}$	$66.0^{\uparrow 3.0}$	$36.7^{\uparrow 1.7}$	–
InternVL3.5-8B*	64	60.4	53.0	37.3	63.5	39.3	68.8
	128	62.2	55.6	42.4	65.1	37.7	–
+MVP (Ours)	64	**63.6**$^{\uparrow 3.2}$	$61.0^{\uparrow 8.0}$	**42.4**$^{\uparrow 5.1}$	**67.0**$^{\uparrow 3.5}$	$39.4^{\uparrow 0.1}$	$72.3^{\uparrow 3.5}$
	128	$64.0^{\uparrow 1.8}$	$59.5^{\uparrow 3.9}$	**44.2**$^{\uparrow 1.8}$	**69.0**$^{\uparrow 3.9}$	$38.8^{\uparrow 1.1}$	–
Qwen3-VL-8B-Thinking*	64	62.6	59.6	38.1	59.7	38.5	74.2
	128	67.4	62.9	41.3	65.4	41.8	–
+ MVP (Ours)	64	$62.7^{\uparrow 0.1}$	**62.0**$^{\uparrow 2.4}$	$39.8^{\uparrow 1.7}$	$63.6^{\uparrow 3.9}$	**40.1**$^{\uparrow 1.6}$	**74.7**$^{\uparrow 0.5}$
	128	**67.6**$^{\uparrow 0.2}$	**63.8**$^{\uparrow 0.9}$	$43.0^{\uparrow 1.7}$	$67.8^{\uparrow 2.4}$	**42.6**$^{\uparrow 0.8}$	–

Table 1: **Comparisons with state-of-the-art methods on various benchmarks.** We report results with 64 and 128 frames for models trained with MVP. The * denotes results evaluated in our experimental settings and the bold text means the best performance under this frame setting.



(a) VideoMME Results (b) LongVideoBench Results (c) LVBench Results

Figure 4: **Data scaling analysis on different benchmarks.** We train for 1 epoch on different data sizes and evaluate the results on three different benchmarks.

2025b), InternVL3.5-8B (Wang et al., 2025d), and Qwen3-VL-8B-Thinking (Bai et al., 2025a). The training utilizes 50k data samples synthesized from the LLaVA-Video-178K dataset (Zhang et al., 2025b). Each MVP training sample is constructed with a sequence length of 15 frames. Preliminary tests(included in the appendix E) indicate that the task presents a significant challenge to the models, therefore, we avoid masking an excessive number of frames. The final configuration consists of 10k samples with 2 masked positions, 25k samples with 3 masked positions and 15k samples with 4 masked positions. All models are trained under the same MVP formulation and reward design to ensure a fair comparison across architectures. For evaluation, we test each dataset using both 64-frame and 128-frame settings (except of TempCompass which include mostly short videos, so we just test it under 64-frame settings), with the token count per frame limited to 256. Detailed hyperparameters for the data synthesis process and training, along with prompts and other settings are provided in the appendix C.

4.2 Experiment Results

4.2.1 Results on Standard Benchmarks.

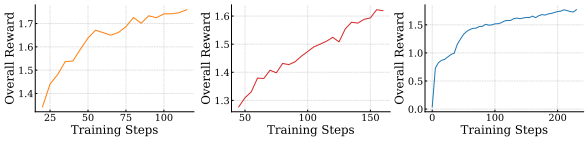
As presented in Table 1, MVP yields comprehensive and consistent improvements across a broad spectrum of video understanding domains, ranging from general perception and long-video understanding to complex reasoning and temporal logic. Crucially, these gains are observed uniformly across diverse base models, regardless of their specific architecture or training paradigm—spanning standard instruction-tuned models like Qwen2.5-VL-7B-Instruct, as well as reasoning-oriented models like Qwen3-VL-8B-Thinking and InternVL3.5-8B. For instance, we observe significant enhancements in long-context tasks (e.g., substantial improvements on LongVideoBench and LVBench across all backbones), alongside robust boosts in general QA (VideoMME, MLVU) and reasoning-intensive benchmarks (Video-Holmes, TempCompass). This universality underscores that MVP provides a fundamental visual supervision signal that effectively generalizes across different model types. To bench-

mark our approach against existing pre-training objectives, we implemented Jigsaw-7B (Wu et al., 2025) as a baseline, which is trained to reorder shuffled video segments using 100k samples synthesized from the same LLaVA-Video-178K dataset. Even with using only half the size of training data of Jigsaw-7B, our MVP method consistently delivers better results across most benchmarks, leading by margins such as 2.7 points on LongVideoBench. This demonstrates that while segment reordering provides only coarse-grained supervision, the MVP task compels the model to master fine-grained inter-frame relationships and intrinsic temporal properties, resulting in a more robust video representation. Notably, these enhancements persist across different input frame settings (64 and 128). This consistency demonstrates that the efficacy of MVP is independent of specific sampling densities, suggesting that the model acquires fundamental temporal reasoning skills that remain effective across varying temporal resolutions.

4.2.2 Ablation Study and Analysis.

Ablation on Reward Components. To validate the effectiveness of our fine-grained reward mechanism, we dissect the impact of each component in Table 2. We train the Qwen2.5-VL-7B-Instruct backbone under various reward settings. To ensure a rigorous comparison, all models are evaluated using checkpoints obtained at an identical number of training steps. We establish a baseline using a strict Exact Match policy (Row 1), where the model receives rewards only when both the frame content and its temporal index are perfectly predicted, with no partial credit for misplaced frames. Introducing the Content-Aware component (Row 2), which grants partial rewards for correctly identifying target frames regardless of their order, leads to immediate performance gains (e.g., +2.0 on LongVideoBench). This indicates that encouraging the model to distinguish relevant visual information from distractors is a crucial first step. Finally, incorporating the Sequence-Level Continuity Bonus (Row 3) yields the best performance across all benchmarks, which demonstrates that explicitly incentivizing the preservation of temporal substructures is essential for the model to master the intrinsic sequential logic and causal dynamics of videos.

Data Scaling Analysis. To validate the scalability of our approach, we investigate the impact of data scaling by training the Qwen2.5-VL-7B model on MVP subsets of varying sizes. We adhered to a rigorous experimental protocol: all models are trained for a single epoch with a fixed mask ratio (2:3:4 frames = 1:2:1) to isolate the effect of data volume. As shown in Figure 4, the results on VideoMME, LongVideoBench and LVBench demonstrate a monotonic increase in performance as the dataset size grows. Crucially, the performance curve shows no signs of saturation within our tested range, indicating that the MVP objective effectively leverages additional data to refine the model's temporal reasoning capabilities. This trend confirms the scalable nature of the MVP task. Moreover, given the model's responsiveness to data quantity, we posit that scaling up the quality and diversity of the training corpus would likely unlock further improvements in video reasoning.



(a) Qwen2.5-VL-7B-Instruct (b) Qwen3-VL-8B-Thinking (c) InternVL-3.5-8B

Figure 5: Training curves on different backbones.

Training Curves. To better illustrate the optimization dynamics, we visualize the training process by plotting the overall validation reward curves for all three backbone models. As shown in Figure 5, the reward values consistently increase over training iterations across all architectures, indicating smooth and stable learning behavior. This steady improvement provides strong empirical evidence for the effectiveness of our training strategy, demonstrating that the proposed fine-grained reward design reliably guides the models toward better frame selection and temporal ordering capabilities. Notably, the absence of oscillation or collapse suggests that the training process avoids common reinforcement learning instabilities. Moreover, the similar convergence trends observed across different backbones indicate that the proposed framework generalizes well across architectures, underscoring the robustness and model-agnostic nature of the MVP training paradigm.

Qualitative Analysis. To qualitatively illustrate MVP's impact on fine-grained perception and inter-frame reasoning, Figure 6 compares the reasoning trajectories of the base and MVP-aligned models. As shown, the base model is prone to temporal hallucinations driven by spurious correlations; it often incorrectly propagates past context (e.g., environ-

Content	Sequence	VideoMME	LongVideoBench	LVBench	MLVU	Video-Holmes	TempCompass
✗	✗	60.4	53.5	37.0	60.5	34.0	67.1
✓	✗	61.0	55.5	37.4	61.3	35.1	68.1
✓	✓	**61.2**	**55.9**	**38.2**	**62.2**	**35.9**	**68.5**

Table 2: **Ablation study on different reward components.** Content means content rewards, without which we only give a reward when the model predicts an exact match (position and content are all correct). Sequence means sequence-level continuity bonus. All experiments are conducted using 64 frames.

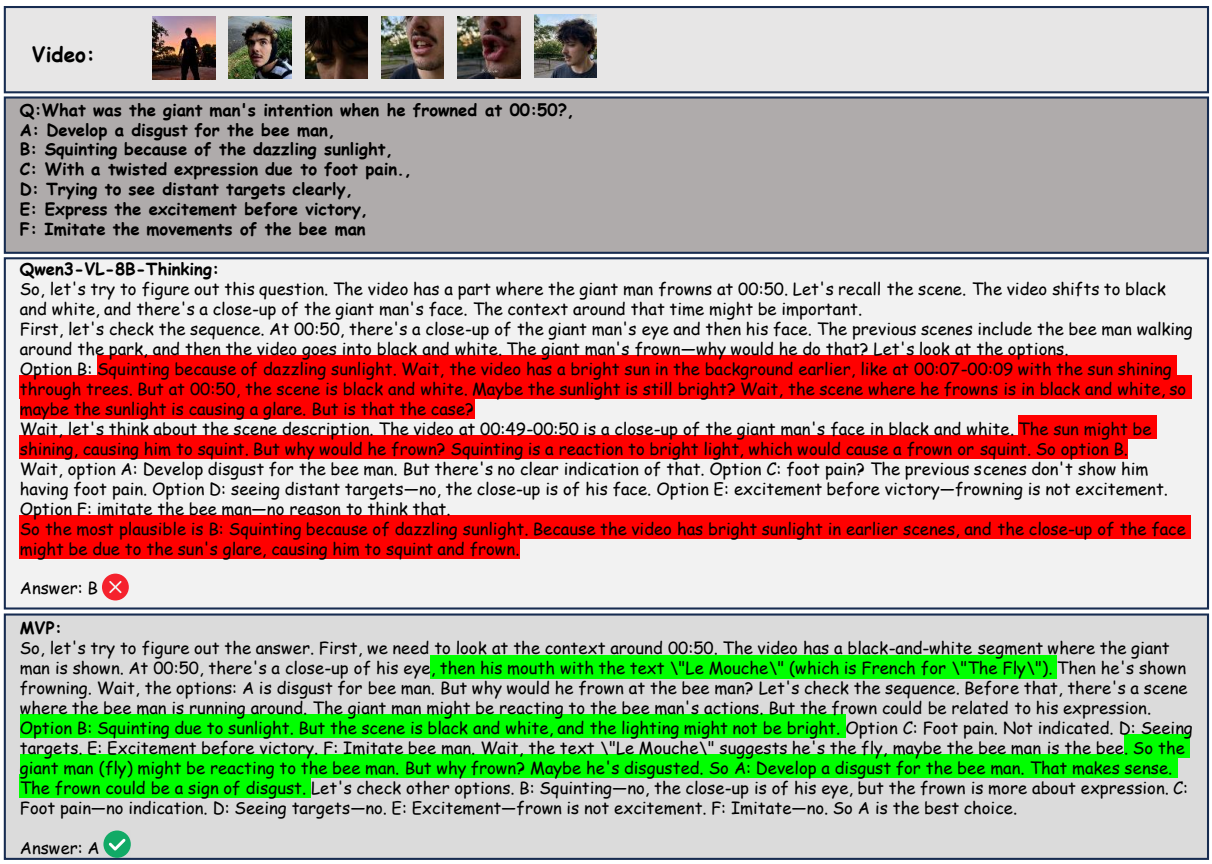


Figure 6: **An example of the reasoning process after training with MVP.** MVP training enhances the model's grasp of video temporality and focuses it more on visual cues to reduce hallucinations.

mental conditions) to explain subsequent events, resulting in plausible but factually incorrect physical interpretations. In contrast, the MVP model demonstrates superior grounding capabilities. By learning to verify temporal continuity through masked prediction, it effectively overrides such hallucinatory priors, instead anchoring its reasoning in subtle but critical visual details present in the relevant frames. The visualization confirms that MVP enables the model to correctly interpret complex causal dynamics, such as distinguishing emotional reactions from physical reflexes, by strictly adhering to the verified visual evidence rather than relying on static semantic associations. This capability reveals a robust verification mechanism, effectively transforming video reasoning from passive pattern matching into active, evidence-based inquiry.

5 Conclusion

In this work, we address the lack of explicit temporal supervision in VideoLLMs by introducing MVP, a novel post-training objective that encourages models to learn intrinsic temporal dynamics and fine-grained inter-frame relationships. Leveraging a scalable data synthesis pipeline and a hierarchical reward design within the GRPO framework, MVP effectively converts arbitrary video data into high-quality supervision for temporal reasoning. Extensive evaluations across multiple benchmarks show that MVP consistently improves performance on diverse video reasoning tasks. Moreover, we observe a monotonic performance gain with increased data scale, underscoring the scalability of the approach and its potential as a strong foundation for future video reasoning models.

Limitations

Although MVP demonstrates significant efficacy in enhancing video reasoning, we acknowledge certain limitations in the current scope of our work. First, while the MVP task necessitates implicit temporal logic, it lacks direct supervision on the explicit reasoning process. The model is trained to optimize the final selection and ordering outcomes, but is not explicitly guided to verbalize the underlying causal relationships or logical deductions behind its decisions. Future research will explore integrating process-level supervision to the model's reasoning trajectory, fostering a more robust and transparent understanding of video content and temporal relationships.

Second, due to time and computational resource constraints, we have not fully explored the asymptotic limits of the MVP task. Our current experiments rely on a relatively small amount of training data and limited training duration, meaning the model's performance has not yet truly converged. As indicated by our scaling analysis, performance continues to improve with increased data, suggesting that the task's full potential is yet to be realized. Investigating the performance upper bounds through large-scale training remains an important direction for future work.

Ethical Considerations

Data Privacy and Consent. Our study exclusively uses publicly available datasets. These datasets were released under open-source licenses by their original curators, ensuring that the data collection process adhered to ethical standards. We have not attempted to re-identify any individuals or extract personally identifiable information from the video data.

Ethical Use of Models. The proposed Masked Video Prediction (MVP) framework is designed for research purposes to enhance Video LLMs. We have conducted manual spot-checks on model outputs to ensure they do not generate offensive or biased content.

Environmental Impact. To reduce the carbon footprint, we prioritize energy-efficient self-supervised learning objectives and utilize pre-trained models whenever possible to minimize redundant computational costs.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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. 2025a. 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.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. 2025b. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, and 178 others. 2025. Seed1.5-vl technical report. *Preprint*, arXiv:2505.07062.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. 2025. Videossr: Video self-supervised reinforcement learning. *Preprint*, arXiv:2511.06281.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025a. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *Preprint*, arXiv:2501.13826.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025b. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *Preprint*, arXiv:2107.09609.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. 2025. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3318–3327.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. 2025. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv.org/abs/2402.03300*, 2(3):5.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. 2025. Tspo: Temporal sampling policy optimization for long-form video language understanding. *arXiv preprint arXiv:2508.04369*.

Zhou Tao, Shida Wang, Yongxiang Hua, Haoyu Cao, and Linli Xu. 2025. Dig: Differential grounding for enhancing fine-grained perception in multimodal large language model. *Preprint*, arXiv:2512.12633.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. 2025a. Fostering video reasoning via next-event prediction. *arXiv preprint arXiv:2505.22457*.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025b. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *Preprint*, arXiv:2505.15966.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025c. 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, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025d. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

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.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. 2025. Visual jigsaw post-training improves mllms. *arXiv preprint arXiv:2509.25190*.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. 2025. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*.

Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. 2025. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding. *arXiv preprint arXiv:2503.18943*.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. 2025. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, and 1 others. 2025. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*.

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. 2025a. Vcapsbench: A large-scale fine-grained benchmark for video caption quality evaluation. *arXiv preprint arXiv:2505.23484*.

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

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

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.

A Use of AI Assistants

We acknowledge the use of AI assistants in the preparation of this work. Their involvement was strictly confined to the following aspects:

- **Language Refinement:** The AI was used to improve the grammatical flow, clarity, and conciseness of the manuscript.
- **Engineering Support:** The AI assisted in writing boilerplate code and debugging scripts for the experimental infrastructure.

All fundamental scientific contributions, including the conceptualization of the research, the design of the methodology, the analysis of results and the primary drafting of the manuscript, were performed entirely by the human authors. The authors retain full responsibility for the accuracy and integrity of the final paper.

B Reproducibility and Stability

To ensure the reliability of our results, all experiments were repeated across three to five independent runs. We observed that the outcomes remained identical due to the use of fixed random seeds and a deterministic greedy decoding strategy (temperature = 0) during inference. Consequently, we report the consistent scores obtained across all trials.

C Further Implementation Details

C.1 Training Details

For the rollout process, the `rollout_batch_size` was configured to 256 for both Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Thinking, while a smaller size of 128 was utilized for InternVL-3.5-8B. Detailed configurations for the remaining rollout parameters are provided in Table 3. Both the `max_prompt_length` and `max_response_length` were fixed at 8192 tokens. During the training phase, an additional 2,000 samples were reserved exclusively for validation. All models were evaluated after a single training epoch to maintain consistency. For the reward function parameters, we set α for rewarding exact matches to 3.0, β for balancing format and correct rewards to 0.1 and γ for rewarding correct content selection to 0.9.

C.2 Data Construction Details

To identify and manage temporal redundancy within the video sequences, we set the redundancy threshold κ to 0.95. For each selected sample, we supplemented the m masked frames with $6-m$ distractor frames sampled from the same source video, thereby ensuring a consistent set of six candidates for the model to choose from.

To further guarantee the quality of temporal reasoning and content representation, we implemented a rigorous filtering pipeline. Specifically, each candidate sample was evaluated by Qwen2.5-VL-72B-Instruct through 10 independent rollouts. Samples

Parameter	Value
n (rollout number)	5
temperature	1.0
top_p	1.0
limit_images	0
gpu_memory_utilization	0.6
enforce_eager	false
enable_chunked_prefill	false
tensor_parallel_size	2
Validation Configs	
– temperature	0.6
– top_p	0.95
– n	1

Table 3: Rollout Configuration.

that failed to accrue any points across all rollouts—based on the reward function defined in Section 3—were deemed to lack meaningful temporal information and were subsequently excluded from the dataset.

D Prompts

We provide detailed prompts for the MVP training and for evaluating on diiferent benchmarks in this section, which are shown in Figure 7, 8.

E Performance of Base Models on MVP

Table 4 presents a comprehensive baseline evaluation of the vision-language models employed in our study, including Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-8B-Instruct-Thinking, and InternVL2-8B, specifically focusing on their zero-shot performance on the MVP task. To ensure statistical significance, the evaluation was conducted across 2,000 diverse samples, utilizing a controlled masking configuration where the number of masked frames followed a 2:3:4 distribution at a fixed 1:2:1 ratio. Our empirical results reveal

Model	Avg. Accuracy (%)	Avg. Format Rate
Qwen2.5-VL-7B	22.37	1.0000
Qwen3-VL-8B	21.54	0.7405
InternVL3.5-8B	5.00	0.3110

Table 4: **Performance Comparison of Different Base Models on MVP.**

that the MVP task poses a substantial challenge for current state-of-the-art models, as evidenced by the relatively low average accuracy across the board. The complexity of temporal reasoning and visual reconstruction inherent in this task suggests that setting an overly aggressive masking strategy could impede the convergence of the model. Consequently, these findings provide a critical heuristic for our training phase: to maintain a stable learning signal and avoid catastrophic forgetting or optimization difficulties, the mask length should be carefully calibrated and kept within a moderate range rather than being set to an excessively high value.

F Case Study

Further visualization and comparison with base models are provided in this section. We can see from most of the cases that after MVP training, the model focuses more on visual information, and reduces hallucinations that occur during the reasoning process.

MVP Training Prompt:

You will be shown a sequence of video frames with a gap. You must select 2 frames from the provided candidates and place them in the correct chronological order to fill this gap.

Provided Frames
Frames Before Gap:
<image>Time: 75
<image>Time: 76
<image>Time: 77
<image>Time: 78
Frames After Gap:
<image>Time: 81
<image>Time: 84
<image>Time: 85
<image>Time: 86
<image>Time: 87
<image>Time: 89
<image>Time: 92
<image>Time: 93
<image>Time: 95

Candidate Frames for Selection
<image>
Selection: a
<image>
Selection: b
<image>
Selection: c
<image>
Selection: d
<image>
Selection: e
<image>
Selection: f

Your Instructions
1. **Analyze the Gap:** Observe the action and object positions in frames before time*78** (before the gap) and frames after time**81** (after the gap).
2. **Identify the Sequence:** Determine which 2 candidate frames create a smooth and logical transition from time 78 to time 81.
3. **Determine Order:** Place these 2 selected frames in the correct temporal order to fill in the gap.
4. **Provide Your Answer:** Output **only** the selection letters of your 2 chosen frames in the correct order, using the exact format: [letter1, letter2]
(Where letter1 is your choice for Time 79,and so on ...).
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE put within <answer></answer> tags.

Figure 7: **Training prompt for MVP.**

Evaluation Prompt:

Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter (A, B, C, or D) of the correct option.
{question}
{options}
Only give the best option.
You FIRST think about the reasoning process as an internal monologue and then provide the final answer.
The reasoning process MUST BE enclosed within <think></think> tags.
The final answer MUST BE put within <answer></answer> tags.

Figure 8: **Evaluation prompt.**

Figure 9: **Example of an MVP sample.** The answer is [b,a,c].

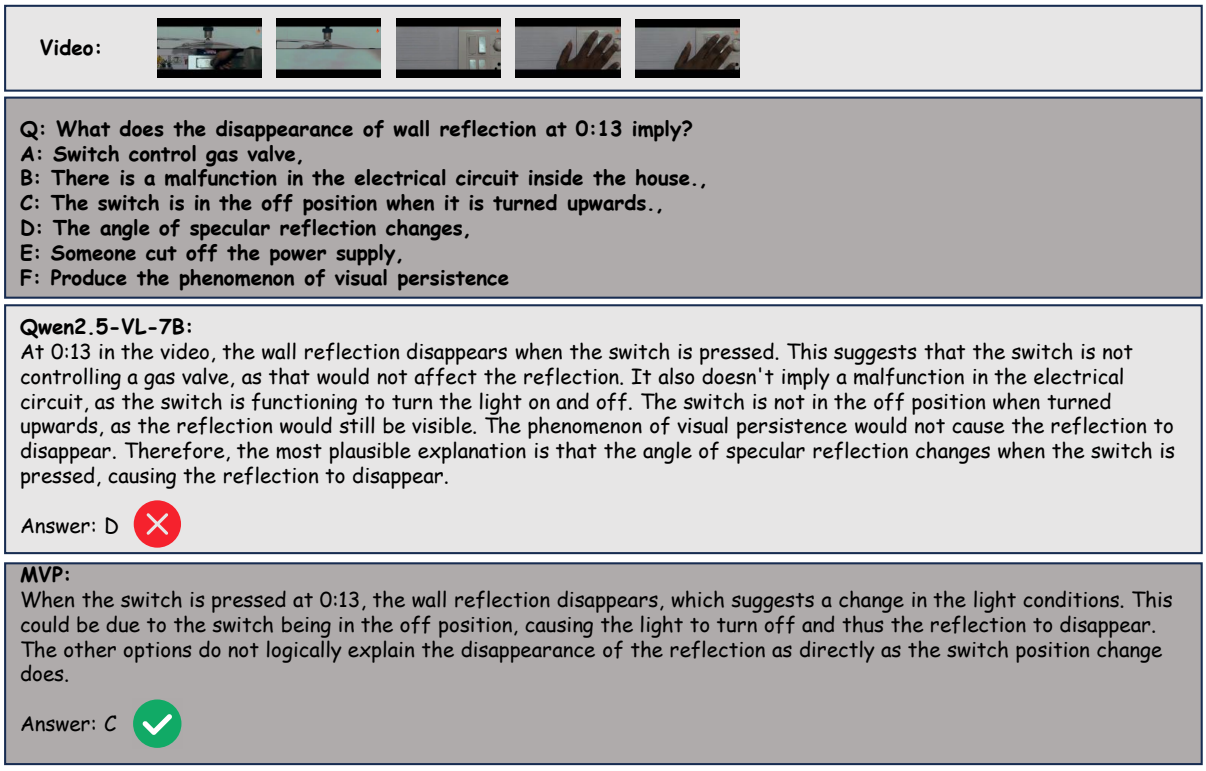
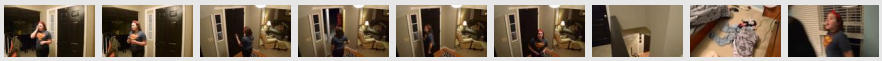


Figure 10: **An example of Qwen2.5-VL-Instruct before and after MVP training.**

Video: 

Q: What was the direct reason why the man in the jacket could finally enter the house?
A: Short-haired woman didn't close the window.
B: Unlock remotely by phone.
C: Use the spare key.
D: Break into the door by damaging the lock.
E: The woman with short hair forgot to lock the door.
F: The woman with short hair takes the initiative to invite.

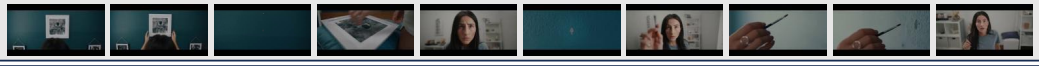
Qwen2.5-VL-7B:
"Let's analyze the sequence of events that led to the man in the jacket being able to enter the house:\n\n1. The woman with short hair was outside the house, talking on her phone.\n2. She was instructed to unlock the door remotely by the man in the jacket.\n3. She successfully unlocked the door remotely.\n4. The man in the jacket then entered the house.\n\nGiven these steps, the direct reason why the man in the jacket could finally enter the house was that the woman with short hair unlocked the door remotely by phone.\n\nTherefore, the correct answer is B: Unlock remotely by phone."

Answer: B ✗

MVP:
"Let's analyze the sequence of events that led to the man in the jacket being able to enter the house:\n\n1. The woman with short hair was initially unable to open the door.\n2. She then called someone on her phone, presumably to unlock the door remotely.\n3. After waiting for a while, she decided to go to the kitchen.\n4. She took off her jacket and placed it on the counter.\n5. She then went to the living room and started talking on her phone.\n6. The man in the jacket was able to enter the house.\n\nGiven these steps, the most logical reason why the man in the jacket could finally enter the house is that the woman with short hair forgot to lock the door. This is because she was distracted by the phone call and the situation in the kitchen, and she didn't take the time to secure the door properly.\n\nTherefore, the correct answer is E: The woman with short hair forgot to lock the door."

Answer: E ✓

Figure 11: **An example of Qwen2.5-VL-Instruct before and after MVP training.**

Video: 

Q: What was the intention of the girl in blue to try to color again after being pushed away at 1:00?
A: Test the quality of the paintbrush.
B: Cover wall surface abnormalities
C: Prove yourself to your sister.
D: Repair the photo frame position
E: Curiosity about supernatural phenomena
F: Escape from the pressure of reality

InternVL-3.5-8B:
"The video shows a girl in blue painting a wall and then being startled by something supernatural, causing her to stop. After being pushed away, she attempts to continue painting, which suggests she is trying to cover or deal with the abnormalities on the wall caused by the supernatural event. This indicates her intention is related to the wall surface rather than personal or emotional reasons."

Answer: B ✗

MVP:
"The video shows a girl in blue painting a wall and then being startled by something supernatural, causing her to stop. Her intention to try to paint again could be driven by curiosity about what she experienced, indicating an interest in understanding or investigating the supernatural occurrence."

Answer: E ✓

Figure 12: **An example of InternVL-3.5 before and after MVP training.**