LEARNING TO LOOK BEYOND A SINGLE FRAME FOR VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large multimodal models (LMMs) have achieved incredible success in video understanding. However, they exhibit a fundamental limitation: single-frame bias. Instead of exploiting the full temporal sequence, they base their understanding on static cues from a few frames, typically the first or the last frame. This shortcut undermines the model's ability to develop a holistic understanding of video content, often resulting in systematic errors. To mitigate this problem, we propose MF-GRPO, a novel framework that employs reinforcement learning (RL) to encourage LMMs to engage with the full temporal context. At its core is the *multi-frame focus advantage enhancement*, which is triggered using a contrastive objective. Specifically, the model receives the advantage enhancement when its understanding of the entire video sequence surpasses its understanding of a partially masked version of the same sequence. This enhancement mechanism explicitly encourages the model to integrate information across the entire temporal span, beyond the excessive reliance on superficial cues of a single frame. Furthermore, this RL training is designed to be progressive. We start with shorter video sequences and gradually extend their length, enabling the model to develop the multi-frame focus ability smoothly from easy to hard. Quantitative experiments show MF-GRPO effectively improves the model performance on various video understanding benchmarks. Critically, qualitative visualizations confirm that our model learns to allocate attention more broadly across frames, successfully mitigating temporal blindness.

1 INTRODUCTION

Video understanding is among the hardest problems in computer vision. It requires a thorough comprehension of spatial-temporal relationships, dynamics, and causality. Recent Large Multimodal Models (LMMs) (Hurst et al., 2024; Bai et al., 2025; Cheng et al., 2024; Maaz et al., 2024; Feng et al., 2025; Tang et al., 2025) have made significant strides in this domain. They encode frames as sequences of language tokens, use powerful decoder-based language models to jointly interpret the video and the question, and then generate answers autoregressively.

Compared with static image understanding, the critical difference of video understanding lies in that it requires a comprehensive understanding of multiple frames, and extracting and modeling spatiotemporal, dynamic, and causal information from them. When humans understand videos, they never focus solely on a single frame (e.g., the first or last frame). Instead, they aggregate information from all frames to make judgments. However, do LMMs really understand the whole content of the input video like humans when they answer the question? Previous studies (Liu et al., 2024b; Feng et al., 2025; Yu et al., 2025) reveal the **single-frame bias** of LMMs in video understanding. LMMs tend to overly focus on a single frame (usually the first or the last frame) rather than comprehend all frames. Due to the underexploitation of video frames, the generated response can be misled by a single frame and cannot correctly answer the question. Figure 1 illustrates an example of this phenomenon. The baseline model overly focuses on the first frame where the woman is in bed, but overlooks the key information in the subsequent frames that show her getting up. So, it thinks the woman has been lying in bed the entire time and has never got up throughout the video, leading to the incorrect answer. Therefore, for robust video understanding, it is important to allocate the attention of LMMs more broadly in the temporal dimension. Note that "broadly" does not mean "uniform attention". Similar to the difference between nearsighted and normal people, the goal is to prevent

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

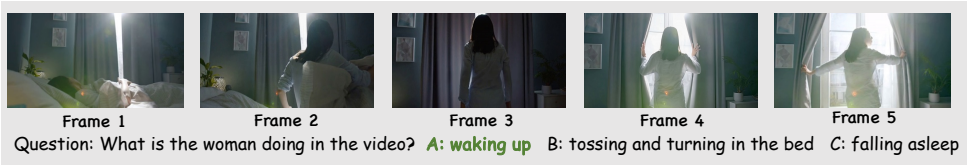


Figure 1: **The baseline model overly focuses on the first frame** where the woman is in bed, while ignoring subsequent frames indicating she is getting up and walking to the window. Consequently, its response is misled by the first frame, thinking that the woman has been lying in bed the entire time and has never got up throughout the video. In contrast, Ours comprehends all frames, such as sitting up in bed (Frame 2), getting out of bed and walking to the window (Frames 2–5), and is therefore able to derive the correct answer.

the model from only seeing first/last frames, and instead find and attend to the really useful critical frames.

To mitigate the single-frame bias and encourage LMMs to comprehend the whole content of the video, we propose **Multi-frame Focus GRPO (MF-GRPO)**, a novel Reinforcement Learning (RL) method for video based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Specifically, MF-GRPO introduces an innovative Multi-frame Focus (MF) advantage enhancement to encourage the model to exploit more frames beyond the first and last frames. Inspired by (Feng et al., 2025; Xue et al., 2025), we employ a contrastive paradigm to guide the learning. A contrast video is designed to contain half of the original input frames, among which the first and last frames are guaranteed to be included. The model generates a group of responses based on the original video for GRPO, and a single deterministic contrast response based on the contrast video. The MF advantage enhancement is granted to correct responses in the group if the average accuracy reward of the group is greater than the contrast response. This enhancement encourages the model to exploit more frames in the full temporal context, beyond the first and the last ones. The core rationale is that, if a response is correct but the contrast response is wrong, the correct response must have utilized information from the other half of the frames not included in the contrast video. Additionally, if the total number of frames is large, it will be more challenging to enhance focus for each frame position. Thus, we propose to progressively develop the ability of multi-frame focus from fewer input frames to more frames, ensuring a smooth learning from easy to hard. The ability to achieve multi-frame focus on a small number of frames serves as the foundation for the success on a larger number of frames. Quantitative experiments on a spectrum of video understanding benchmarks demonstrate the efficacy of MF-GRPO. The qualitative visualization of attention maps shows that MF-GRPO facilitates the model to focus on more frames. These results underscore the significance of multi-frame focus in video understanding. In summary, our main contributions are:

1. To mitigate the single-frame bias and encourage LMMs to see more frames across the entire temporal span, we propose **MF-GRPO** that leverages a novel Multi-frame Focus (MF) advantage enhancement to facilitate the model to exploit information beyond a single frame.
2. We propose a progressive multi-frame focus learning strategy to develop the multi-frame focus ability smoothly from fewer to more frames.
3. Quantitative experiments on multiple video understanding benchmarks demonstrate the effectiveness of our approach. Meanwhile, the qualitative visualization shows the trained model exhibits better multi-frame focusing capability.

2 Related Works

2.1 LMMs for Video Understanding

The rapid advancement of Large Multimodal Models (LMMs) has revolutionized video understanding by bridging visual perception and language reasoning. They encode visual input into a sequence of tokens in the language space, and thus can leverage the strong pretrained large language model to understand both the visual and textual inputs. Early LMMs for images (e.g., GPT-4V (OpenAI, 2023) and LLaVA (Liu et al., 2023)) demonstrated strong static visual comprehension. As the higher-dimensional data than images, videos contain rich spatiotemporal, dynamic, and causal information. Recent research on Video LMMs is committed to addressing video-specific issues, such as temporal-order perception (Xue et al., 2025; Feng et al., 2025), spatial-temporal modeling (Cheng et al., 2024), long video understanding (Song et al., 2024), temporal reasoning failure caused by language and image priors (Qi et al., 2025), etc. Besides, a line of works incorporate the chain-of-thought (Wei et al., 2022) into Video LMMs to enhance video reasoning ability, using prompting (Fei et al., 2024) or reinforcement finetuning (Feng et al., 2025). In this paper, we investigate a fundamental problem for Video LMMs, i.e., do they really understand the whole content of the input video when they answer the question?

2.2 Single-Frame Bias in Video Understanding

We distinguish two types of single-frame bias, i.e., data-induced bias and model-induced bias. Early works reveal the data-induced single-frame bias in video understanding benchmarks, which refers to a phenomenon that the static appearance of a single frame is sufficient to answer the question in the video benchmarks (Buch et al., 2022; Lei et al., 2023). For example, Lei et al. (Lei et al., 2023) find that a single-frame trained model without considering temporal information can outperform video-based methods. Buch et. al (Buch et al., 2022) gives a similar observation. They train a model to select only the most appropriate frame for a video understanding task, and find that it can outperform video-based models. The underlying reason for data-induced single-frame bias primarily stems from the static appearance bias in the test benchmarks. Later, some studies address this issue by enhancing the temporal and dynamic properties of videos in the benchmark (Liu et al., 2024b; Cores et al., 2024; Li et al., 2024b). They ensure that these video-related problems cannot be effectively solved using a single frame. Recently, after the LMMs have predominated in video understanding, Yu et al. (Yu et al., 2025) and Feng et al. (Feng et al., 2025) point out the single-frame bias inherent in the inference of LMMs. With a video input, LMMs tend to overly focus on a single frame (usually the first or the last frame), taking shortcuts to answer the question without truly understanding the full content of the video. Yu et al. (Yu et al., 2025) propose to mitigate this issue by explicitly modeling spatiotemporal attributes (e.g., bounding boxes) alongside a specially designed query strategy, relying on a suite of external expert models, but they do not internalize the multi-frame focusing capability. In this paper, the single bias issue denotes this sort of model-induced bias inherent in LMMs, and we avoid the data-induced bias by selecting highly temporal and dynamic benchmarks. We are the first to attempt to fundamentally address the single-bias issue of LMMs.

2.3 Reinforcement Learning for LMMs

Reinforcement learning has emerged as a popular post-training paradigm to enhance the video understanding and reasoning ability of LMMs. Video-R1 (Feng et al., 2025) and VideoChat-R1 (Li et al., 2025) endow the video LLMs with the chain-of-thought reasoning ability by adopting DeepSeek-R1-like rule-based RL, specifically leveraging GRPO that uses verifiable outcomes (e.g., answer correctness, format correctness) for optimization. Besides, Video-R1 propose to enhance the understanding of temporal order by assigning an extra reward when the ordered sequence outperforms the shuffled one. ArrowRL (Xue et al., 2025) also proposes to enhance temporal order by encouraging the divergence between the ordered and the shuffled sequence. VideoChat-R1 proposes to enhance spatio-temporal perception by integrating accuracy, IoU, and caption rewards. TW-GRPO (Dang et al., 2025) proposes to prioritize tokens with high informational density to promote concise and task-aware video reasoning. Orthogonal to previous studies, our work is devoted to enhancing multi-frame focus for LMMs via GPRO-based reinforcement learning.

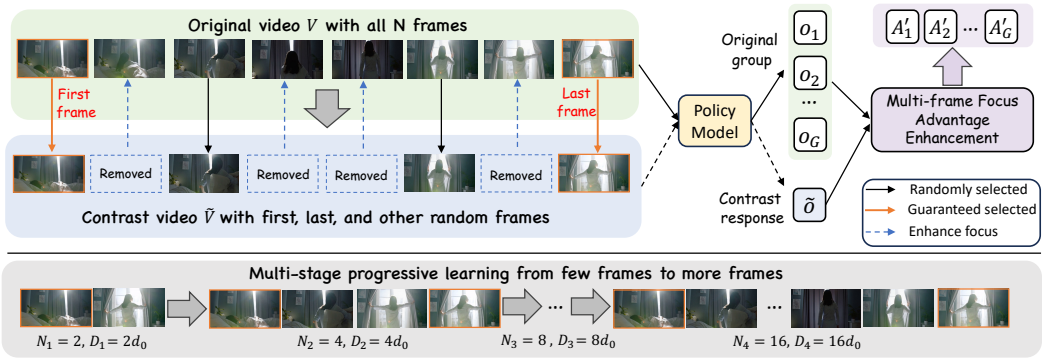


Figure 2: **The illustration of our MF-GRPO framework.** (a) The upper panel demonstrates the Multi-frame Focus advantage enhancement. The contrast video $\tilde{V}$ is constructed by selecting half of the frames of V, including the first, last, and other random frames. The MF advantage enhancement is defined as Equation (4), which encourages the LMM to enhance the focus on these not-selected (removed) frames. (b) The lower panel shows the progressive learning process. The model gradually learns multi-frame focus capacity from fewer input frames to more frames.

3 METHOD

In this section, we elaborate on the proposed MF-GRPO framework. There are two main components of our method. **Firstly,** we propose the Multi-frame Focus (MF) advantage enhancement mechanism, which facilitates LMMs to break free from single-frame bias by providing additional advantage enhancement for multi-frame focus behaviours. **Secondly,** considering that it is more challenging for the model to enhance focus for each frame when the total number of frames is large, we propose to progressively foster this multi-frame focus ability from fewer to more frames. The overall framework is illustrated in Figure 2.

3.1 MULTI-FRAME FOCUS ADVANTAGE ENHANCEMENT

We propose the Multi-frame Focus (MF) advantage enhancement to encourage the multi-frame focus behaviours, to facilitate LMMs to comprehend the whole content of the video rather than overly focus on a single frame. We will elaborate on the two critical points of the approach, i.e, how to determine multi-frame focus behaviours and how to encourage such behaviours.

Determine multi-frame focus behaviours by a contrastive mechanism. The key of MF advantage enhancement lies in providing advantage enhancement when the model exhibits multi-frame focus behaviours, to consolidate such behaviours. Therefore, a fundamental question is how to determine when the model exhibits multi-frame focus behaviours. To start with, we hold the following natural assumptions:

1. *A video V possess more information than its arbitrary sub-video $\tilde{V}$, where $\tilde{V} \subsetneq V$.*
2. *For video understanding with LMMs, understanding more information always yields better answers.*
3. *If a question cannot be solved with a set of frames, but can be solved by adding extra frames, the model must leverage the critical information in the additional frames.*

Therefore, inspired by (Feng et al., 2025; Xue et al., 2025), we employ a contrastive mechanism to produce multi-frame focus behaviours. Let $V = \{v_t\}_{t=1}^{N}$ denote the video with N frames. We construct a contrast video $\tilde{V}$ by selecting half of the frames of V. Note that we enforce the inclusion of the first and last frames, and the remaining frames are selected randomly. The underlying principle is that the first and last frames are always overly focused (Feng et al., 2025; Yu et al., 2025), so we expect the model to exploit the other frames beyond the first and last frames. It is formulated as:

$$\tilde{V} = \{v_1, v_N\} \cup \{v_{t_i}\}_{i=1}^{N'-2},$$
$$\{t_i\}_{i=1}^{N'-2} \subseteq \{2, \ldots, N-1\},\ N' = N//2. \tag{1}$$

Let q denote the input question. We generate a group of responses $O = \{o_i\}_{i=1}^{G}$ with the question q and the original video V, i.e.,

$$\{o_i\}_{i=1}^{G} \sim \pi_\theta(O|q, V), \tag{2}$$

where π_θ is the policy model (i.e., the LMM) and G is the number of generations in the GRPO group. We randomly sample the original response group to encourage exploration. In contrast, given q and the contrast video $\tilde{V}$, we generate only a single contrast response $\tilde{o}$ using greedy decoding:

$$\tilde{o} \sim \pi_\theta(o|q, \tilde{V}). \tag{3}$$

This design considers both effectiveness and efficiency. Greedy decoding typically yields stronger performance on deterministic MLLM tasks, providing a stricter contrast for triggering the MF enhancement. Meanwhile, generating only one contrast response avoids the additional sampling cost of a full contrast group. We then compute the rewards $\{R_i\}_{i=1}^{G}$ for the original response group, where R_i may include other rewards besides the accuracy reward, such as the format reward.

Encourage multi-frame focus behaviours via direct advantage enhancement. Different from Video-R1 (Feng et al., 2025) and ArrowRL (Xue et al., 2025), we propose to directly enhance the advantages of the responses that we expect to encourage, rather than add an extra reward, to avoid the reward vanishing problem after group standardization. According to GRPO, the advantages of the group are calculated by $A_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}$. The reward vanishing issue occurs when there are only two different levels of reward value. For example, let us consider a group of two correct and two incorrect responses. Their rewards are $[r_1, r_1, r_2, r_2]$, where $r_1 > r_2$. They obtain the advantages of $[a_1, a_1, a_2, a_2]$ after group standardization, where $a_1 = \frac{r_1 - mean(r_1, r_1, r_2, r_2)}{std(r_1, r_1, r_2, r_2)} = \frac{\sqrt{3}}{2} \cdot \frac{r_1 - r_2}{|r_1 - r_2|} = \frac{\sqrt{3}}{2}$ and $a_2 = -\frac{\sqrt{3}}{2}$. We can see that the advantage is a constant, independent of the absolute magnitudes of r_1 and r_2. So, even though the rewards will change to $[r_1 + \beta, r_1 + \beta, r_2, r_2]$ after adding an extra reward of β to the correct responses, the advantages will remain unchanged. Therefore, to avoid the reward vanishing problem and truly encourage the target responses, we propose to directly enhance the advantages of the responses that are expected to encourage. **Our MF advantage enhancement is defined as:**

$$A_i^{'} = \begin{cases} (1+\alpha) * A_i, & \text{if } r_g > \tilde{r} \text{ and } A_i > 0 \\ A_i, & \text{otherwise,} \end{cases}$$
$$\text{where } r_g = \text{mean}\left(\text{Acc}(\{o_i\}_{i=1}^{G})\right),\ \tilde{r} = \text{Acc}(\tilde{o}), \tag{4}$$

where α is the hyper-parameters controlling the strength of advantage enhancement, and ACC(.) is the accuracy reward function, e.g., correctness for multiple choice and ROUGE scores for free-form questions. r_g is the average accuracy reward of the group, while $\tilde{r}$ is the accuracy reward of the contrast response.

The MF advantage enhancement encourages the LMM to exploit the information contained in the frames which are in the original video V but not in the contrast video $\tilde{V}$. When $r_g > \tilde{r}$, the average performance of the group surpassed the contrast response, indicating the occurrence of multi-frame focus behaviour. The rationale is that, if the question q cannot be effectively solved with $\tilde{V}$ but can be solved by V, the model must have leveraged the critical information in the additional frames, i.e., $V' = V - \tilde{V}$. Recall that $\tilde{V}$ is constructed by selecting the first, last frame, and other randomly selected frames. Thus, V' can contain frames of arbitrary positions, except the first and the last ones. By encouraging the multi-frame focus behaviors, LMMs are learned to extend their focus to multiple frames. We use Figure 4 to support this statement. Figure 4 shows the attention maps of the baseline and the model trained with MF-GRPO. Whether the model exhibits excessive focus on the first frame or the last frame, our method enhances the attention of the model to other frames and reduces excessive reliance on a single frame.

We use Figure 3 to further illustrate the principle of the MF advantages enhancement. We present a correct response in the group with more input frames, and an incorrect contrast response with fewer

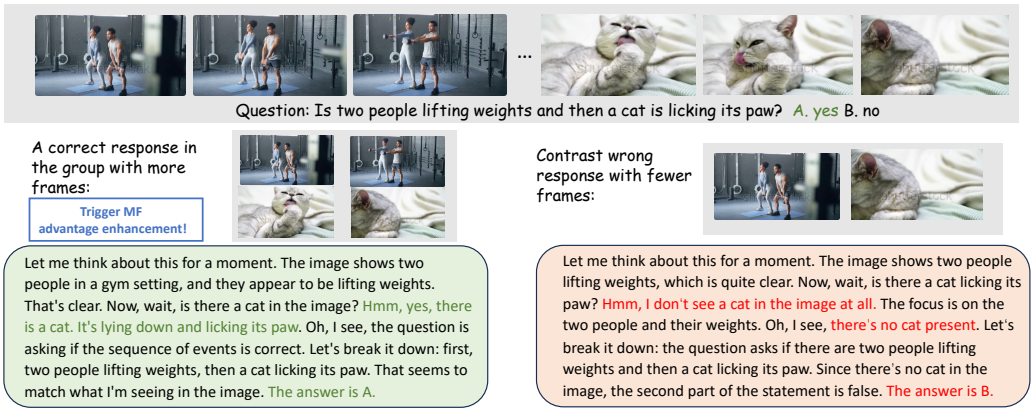


Figure 3: **An example when MF advantage enhancement is triggered.** With only 2 frames, the contrast response is wrong, because it cannot recognize the cat according to the selected non-typical cat frame. On the contrary, by understanding the additional frames which provide the exact information of the cat, the model can produce a correct response. It means a multi-frame focus behaviour emerges, so the MF advantage enhancement is triggered to encourage this behaviour.

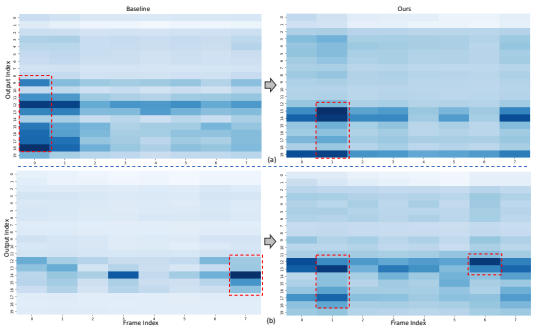


Figure 4: The visualization of the attention maps of baseline and ours. (a) and (b) show the situations where the model overly focuses on the first and the last frame, respectively. Ours enhances the attention to intermediate frames.

frames. The contrast response cannot recognize the cat because the input only contains a side-view cat frame that is confusing for the model. On the contrary, by understanding the additional frames that provide the exact appearance of the cat, the model is confident to identify the cat. It indicates that a multi-frame focus behaviour occurs. If the model cannot understand the additional frames, it will not answer correctly. As it is a multiple-choice question, the correct and the wrong responses get an accuracy reward of 0 and 1, respectively. So, we must have $r_g > \tilde{r}$ and the MF advantage enhancement will be triggered, encouraging this behaviour of understanding more frames.

Finally, the model is trained to maximize the GRPO objective with the enhanced advantage $A_i^{'}$:

$$\mathcal{J}_{\text{MF-GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}\Big[\frac{1}{G}\sum_{i=1}^{G}\big(\min\big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i^{'},\ \text{clip}\big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}\big)A_i^{'}\big) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\big)\Big] \quad (5)$$

3.2 Progressive Multi-frame Focus Learning

We have introduced MF advantage enhancement to encourage multi-frame focus behaviours. The advantage is enhanced when the model looks at V' beyond $\tilde{V}$. The number of the frames that the model is encouraged to focus on is $|V^{'}| = N//2$. However, if the number of frames $(N//2)$ is large, it will be more challenging to care for each frame position. To tackle this problem, we propose a progressive multi-frame focus learning strategy, which gradually learns the ability of multi-frame

focus from fewer input frames to more frames. The ability to achieve multi-frame focus on a small number of frames serves as the foundation for learning on a larger number of frames.

Specifically, we divide the training process into multiple stages $\{S_i\}_{i=1}^{M}$ with corresponding numbers of frames $\{N_i\}_{i=1}^{M}$ and numbers of training steps $\{D_i\}_{i=1}^{M}$, where M is the total number of stages. It means that for each stage S_i, we only sample N_i frames to construct the original video V, and train the model for D_i steps. We set $N_i = 2^i$ and $D_i = d_0 * 2^i$ which exhibits exponential growth. The base value d_0 is determined by the total number of training steps. Larger N_i naturally needs more training steps to learn multi-frame focus. As a special case, we only select the first frame to construct the contrast video $\tilde{V}$ when V only contains 2 frames ($N_i = 2$). In our experiments, we set $M = 4$, which means we have four stages S_1, S_2, S_3, S_4 with corresponding frame numbers $2, 4, 8, 16$. **There are two benefits of the progressive strategy:**

(a) *Ensuring a smooth learning from easy to hard. The model learns to focus on 2^i frames before it advances to learn on 2^{i+1} frames.*

(b) *When the model is learning to focus on V with 2^{i+1} frames, it can already well understand $\tilde{V}$ with 2^i frames, making $\tilde{V}$ a robust contrast sample.*

Ablation study in Section 4.3 demonstrates the efficacy of the progressive learning strategy.

4 Experiments

4.1 Experimental Setup

Benchmarks. To effectively evaluate the model, we choose six video understanding benchmarks with rich dynamics and temporality, i.e., MMVU (Zhao et al., 2025), VSI-Bench (Yang et al., 2025), MVbench (Li et al., 2024b), TempCompass (Liu et al., 2024b), TVbench (Cores et al., 2024), and AoTBench (Xue et al., 2025). Among them, MMVU, VSI-Bench, and MVbench are general benchmarks for video understanding. In contrast, TempCompass, TVBench, and AoTBench are specifically designed for temporal perception and understanding. They are more sensitive to temporality and therefore require capturing information from multiple frames, as multiple frames constitute the foundation of temporal understanding. For example, some questions in TempCompass temporally concatenate two clips and ask the model to distinguish their event order. TVBench includes tasks of action counting and object shuffling, which are designed to be unsolvable from a single frame. Figure 3 also provides a visual example where the model must see intermediate frames of the people lifting weights and the cat's face. The reported results are averaged over three independent runs with different random seeds.

Training and inference details. We use the Qwen2.5-VL-7B-COT-SFT model (Feng et al., 2025) as our base model. The model is a thinking version of Qwen2.5-VL-7B by finetuning on chain-of-thought reasoning data, and is proved to provide a better initial state for RL (Feng et al., 2025). The final reward function contains two types of rewards, i.e, format reward (to maintain the thinking ability) and the accuracy reward. For multi-choice questions, the accuracy reward is the correctness of the answer, i.e., 0 or 1. For free-form questions, the accuracy reward is ROUGE scores. The format reward is granted when the response follows the format of "`<think>...</think>`". The training data is constructed from Video-R1-260k (Feng et al., 2025) by only selecting the video data. We set 4 stages for the progressive learning, which means the least number of frames of V is 2 while the most is 16. We train the model with a total 1K RL steps, so d_0 is 33.33 and each stage is trained with 67, 133, 267 and 533 steps, respectively. We set the enhancement strength α to 0.2. For the group generation, top p is set to 0.95 with temperature=1. For contrast response generation, greedy sampling is used to provide a strict contrast and improve sampling efficiency, which is consistent with common inference configurations. The size of the group G is 8. The model is trained on 5 H200 GPUs for about 10 hours. We use the Adam optimizer with a learning rate of 1e-6 and batch size of 4. The KL weighting of GRPO is 0.04. In the inference, we use greedy sampling with 16 input frames with fps=2 on all benchmarks.

4.2 Main Results

Table 1 summarizes the quantitative results of different models across six diverse video understanding benchmarks. Our MF-GRPO consistently outperforms both the base model (Qwen2.5-VL-7B-COT-SFT) and its reinforcement-learning baseline (Vanilla GRPO) on all datasets, demonstrating the effectiveness of introducing the proposed multi-frame focus enhancement mechanism. Concretely, MF-GRPO achieves an improvement of 0.6–3.5 points over Vanilla GRPO, and 0.1–5.5 points over the base model. The most notable gains appear on these temporality-sensitive benchmarks, i.e, TempCompass (+2.5), TVbench (+2.0), and AoTBench (+3.5), which are designed to assess temporal perception and understanding. The reason is that temporality-sensitive tasks are more susceptible to single-frame biases, as temporality inherently relies on multiple frames. A single image can contain appearance and motion cues, but it definitely cannot possess temporality by itself. Overall, these results indicate that explicitly enhancing multi-frame focus facilitates the model to achieve a more holistic and effective understanding of videos.

Table 1: **The results of different models on six video understanding benchmarks.** We use Qwen2.5-VL-7B-COT-SFT as our base model for RL. We compare our MF-GRPO to the base model and the vanilla GRPO without our MF advantage enhancement. We also report other advanced models for reference.

Models	General Video Benchmarks			Temporality-sensitive Benchmarks		
	MMVU(mc)	VSI-Bench	MVbench	TempCompass	TVbench	AoTBench(mc)
GPT-4o (Hurst et al., 2024)	75.4	34.0	-	74.8	55.2	-
VideoLLaMA2 (Cheng et al., 2024)	44.8	-	54.6	-	-	-
Video-UTR-7B (Yu et al., 2025)	-	-	58.8	59.7	-	-
LLaVA-OneVision-7B (Li et al., 2024a)	49.2	32.4	56.7	69.6	48.9	46.2
VILA-1.5-8B (Zhang et al., 2024)	-	28.9	-	58.8	-	-
LongVA-7B (Lin et al., 2024)	-	29.2	61.1	56.9	-	-
Kangaroo-8B (Liu et al., 2024a)	-	-	61.1	62.5	-	-
Video-R1-7B (Feng et al., 2025)	64.2	34.6	62.7	72.6	-	-
Qwen2.5-VL-7B-COT-SFT (Feng et al., 2025)	60.2	31.8	60.4	68.8	46.1	48.1
+ Vanilla GRPO	62.5	33.7	64.6	71.8	47.2	44.7
+ MF-GRPO (Ours)	**64.2**$_{+1.7}$	**35.1**$_{+1.4}$	**65.2**$_{+0.6}$	**74.3**$_{+2.5}$	**49.2**$_{+2.0}$	**48.2**$_{+3.5}$

Table 2: **The results of ablation study.** Different α values represent different strength of the MF advantage enhancement. MF-GRPO (w/o prog) denotes removing the progressive learning strategy. MF-GRPO (full random) means fully randomly selecting frames for the contrast video, without guaranteeing that the first and last frames will definitely be selected. MF-GRPO (no first last) means definitely not selecting first and last frames. MF-GRPO (uniform) means uniformly sampling frames, where a fixed half of the frames are consistently excluded. MF-GRPO (w/o prog), MF-GRPO (full random), MF-GRPO (no first last), and MF-GRPO (uniform) all use $\alpha = 0.20$ to fairly compare with the final version MF-GRPO ($\alpha = 0.20$).

Models	General Video Benchmarks			Temporality-sensitive Benchmarks			Avg.
	MMVU(mc)	VSI-Bench	MVbench	TempCompass	TVbench	AoTBench(mc)	
Qwen2.5-VL-7B-COT-SFT (Feng et al., 2025)	60.2	31.8	60.4	68.8	46.1	48.1	52.5
Vanilla GRPO	62.5	33.7	64.6	71.8	47.2	44.7	54.0
MF-GRPO (w/o prog)	63.8	36.5	64.8	73.8	46.8	47.3	55.5
MF-GRPO (full random)	64.5	37.0	64.7	73.4	48.2	47.3	55.8
MF-GRPO (no first last)	63.5	33.9	65.1	73.6	47.2	47.2	55.0
MF-GRPO (uniform)	62.2	33.1	65.1	74.1	48.8	48.3	55.2
MF-GRPO ($\alpha = 0.10$)	63.5	36.5	65.0	73.1	50.0	48.2	56.0
MF-GRPO ($\alpha = 0.30$)	62.1	37.3	64.9	74.0	47.8	47.3	55.6
MF-GRPO ($\alpha = 0.20$)	64.2	35.1	65.2	74.3	49.2	48.2	**56.0**

4.3 Ablation Study and More Analysis

The method is not sensitive to α within an appropriate range. We further investigate the sensitivity of MF-GRPO to the hyperparameter α, which controls the enhancement strength. As shown in Table 2, all configurations with α values between 0.10 and 0.30 outperform both the base model and the vanilla GRPO baseline, demonstrating the robustness of our method to moderate variations in α.

The effectiveness of the definite selection of the first and last frames. As stated in Section 3.1, the first and the last frames are guaranteed to be included in the contrast video. The reason is that

they have already obtained enough attention and we expect the model to look at other positions. To validate the effectiveness of this selection strategy, we conduct an experiment where the frame selection is fully random. In this case, the first and last frames might not be included in the contrast sample. As shown in Table 2, MF-GRPO (full random) degrades the performance on four of the six benchmarks compared with MF-GRPO ($\alpha = 0.20$). In particular, degradations are observed across all three temporality-sensitive benchmarks, indicating a decline in the multi-frame focus capability of the model. Moreover, MF-GRPO (no first last) definitely excludes the first and last frames, which causes further degradation. Both MF-GRPO (full random) and MF-GRPO (no first last) give the chance to maintain the biased attention on the first and last frames, which is not consistent with our goal to alleviate single-frame bias. Additionally, MF-GRPO (uniform) samples frames uniformly (e.g., frames 0, 2, 4, 6, ...), which consistently excludes the other half of the frames (1, 3, 5, ...). This also undermines performance, as it encourages the model to focus on a fixed set of positions while neglecting the remaining positions. Despite this, all these settings still surpass the Vanilla GRPO, indicating the efficacy of looking at more frames by MF-GRPO.

The effectiveness of the progressive multi-frame focus learning. As shown in Table 2, MF-GRPO (w/o prog) shows the results when removing the progressive learning. It degrades the performance on five of the six benchmarks. Especially on temporality-sensitive benchmarks, it leads to a loss of 0.5, 2.4 and 0.9, respectively. These results show the efficacy of progressively learning the multi-focus focus ability. Moreover, Figure E.2 visualizes the attention maps with and without progressive learning when the number of input frames is 4, 8 and 16, respectively. For each N, the attention with progressive learning is overall distributed more broadly than the non-progressive one, enhancing the focus on multiple frames. In contrast, the model without progressive learning becomes more difficult to focus on multiple frames as the frame number increases. These results demonstrate the effectiveness of the progressive strategy. It facilitates the model to gradually develop multi-frame focus capacity from fewer to more frames. Besides, the progressive learning ensures that the contrast video composed of half of the frames can be well understood, since the model has already learned on this number of frames in the last stage, making the contrast video a robust contrast sample. For example, understanding 4 and 8 frames makes them good contrast samples for learning multi-frame focus on 8 and 16 frames.

Analysis of the training curve. Figure 5 shows the accuracy reward over 1K RL training steps. The reward exhibits a clear upward trend and gradually stabilizes as training proceeds, indicating that the model progressively improves its task performance through reinforcement learning with our MF-GRPO.

We provide additional experimental results on more models and benchmarks in the Appendix, together with further details on the training and inference settings, benchmark configurations, and qualitative results.

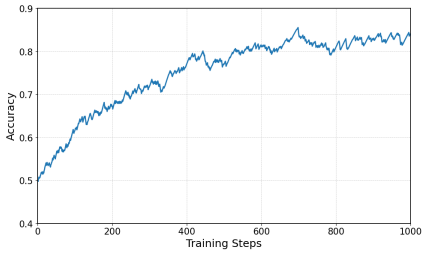


Figure 5: Accuracy reward during RL training.

5 Conclusion

In this work, we aim to mitigate a fundamental limitation of LMMs in video understanding, namely the single-frame bias, which prevents models from forming a holistic comprehension of video content. To address this issue, we introduce an RL-based framework termed MF-GRPO, designed to encourage LMMs to leverage the full temporal context rather than over-relying on the static cues present in a single frame. At the heart of our method lies the Multi-frame Focus advantage enhancement, a mechanism that activates when the model demonstrates desirable multi-frame focus behaviours through a contrastive mechanism, so as to reinforce such behaviours. In addition, we adopt a progressive learning paradigm to ensure a smooth learning from easy (fewer frames) to hard (more frames), allowing the model to gradually build the multi-frame focus capability. Extensive quantitative results across six diverse benchmarks demonstrate that our approach outperforms the base model and the vanilla GRPO, especially on temporal-sensitive benchmarks that rely more heavily on multiple-frame understanding. Additionally, qualitative visualization of attention maps shows that our method truly facilitates the model to allocate the attention more broadly across all frames. We hope that MF-GRPO offers a promising direction and new insights for developing stronger LMMs that can truly understand and integrate information distributed across the entire temporal span.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, or formulate mathematical claims. Dataset cleaning and reformatting were not applicable to this work. Additionally, we used generative AI tools to polish the writing. We reviewed all AI-assisted work. For example, we checked the correctness of the AI-polished text. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2917–2927, 2022.

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

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*, 2024.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *arXiv preprint arXiv:2505.24718*, 2025.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: step-by-step video reasoning from perception to cognition. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13109–13125, 2024.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, 2023.

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

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

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

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 34892–34916, 2023.

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

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8731–8772, 2024b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

OpenAI. Gpt-4v(ision) system card. 2023. URL https://api.semanticscholar.org/CorpusID:263218031.

Daiqing Qi, Dongliang Guo, Hanzhang Yuan, Handong Zhao, Mengxuan Hu, Lehan Yang, and Sheng Li. Improve temporal reasoning in multimodal large language models via video contrastive decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

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

Yunlong Tang, Jing Bi, Pinxin Liu, Zhenyu Pan, Zhangyun Tan, Qianxiang Shen, Jiani Liu, Hang Hua, Junjia Guo, Yunzhong Xiao, et al. Video-lmm post-training: A deep dive into video reasoning with large multimodal models. *arXiv preprint arXiv:2510.05034*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.

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

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

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Additional Experiments

A.1 Generalization to Different Models

To further evaluate the generalization of our method, we conduct experiments on InternVL3-2B (Zhu et al., 2025) in addition to Qwen2.5-VL. As shown in Table A.1, our method consistently improves performance across all evaluated benchmarks, increasing the average score from 58.3 to 60.8. These results demonstrate that our method generalizes well to different multimodal large language models.

Table A.1: Results on InternVL3-2B across different video understanding benchmarks.

Model	Method	TempCompass	MMVU	MVBench	TVBench	AoT QA	Avg.
InternVL3-2B	Vanilla GRPO	63.4	56.7	69.0	51.7	50.7	58.3
	Ours	**66.1**	**58.9**	**71.6**	**53.3**	**53.9**	**60.8**

A.2 Evaluation on Long-Video Understanding

We further evaluate our method on the long-video benchmark Video-MME-Long (Fu et al., 2025) to examine its effectiveness for long-temporal reasoning. Our method improves the performance from 47.33% to 49.22%, showing that mitigating single-frame bias also benefits long-video understanding.

A.3 Evaluation on Sparse Key-Frame Retrieval

We also evaluate whether MF-GRPO sacrifices the model's performance on tasks which only require attention to few key frames. Our MF-GRPO encourages broad attention rather than uniformly distributed attention. It mitigates the blindness of the model, rather than treating each frame equally. So, theoretically, it will not undermine the capacity to focus on few key frames required by certain tasks. To support this, we conduct experiments on VideoNIAH (Zhao et al., 2024), where the model needs to retrieve a few inserted "needle" frames from a long video. Vanilla GRPO and MF-GRPO both achieve an accuracy of 43.1%, indicating that our method preserves the model's ability to focus on sparse key frames.

B Training and Inference Details

Training and inference configurations. We use a batch size of 4 and a learning rate of 1e-6. Following Video-R1 , for efficiency, we limit the maximum number of video frames to 16, so the frame number for each stage is 2, 4, 8 and 16. Each frame is processed at a max resolution of 128x28x28 pixels. For inference, we use a fixed frame number of 16, and relax the max resolution to 256x28x28.

Post-training data. We use a subset of Video-R1-260k by filtering out all image data to avoid its interference in our multi-frame focus video training. The filtered video-only dataset consists of 116K video question-answering pairs, including multiple-choice and free-form problems. Our training only uses 4,000 of them, as we only train 1K steps for MF-GRPO.

Prompt template. In both training and inference, following Video-R1, we use a chain-of-thought style prompt to encourage thinking step by step. The prompt is shown in Figure B.1.

Prompt Template for Training and Inference

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> and </think> tags, and then give your final answer between the <answer> and </answer> tags.

Figure B.1: Prompt template for training and inference

C BENCHMARK DETAILS

We use three general video benchmarks (i.e., MMVU, VSI-Bench, MVbench) and three temporal-sensitive benchmarks (i.e., TempCompass, TVbench, AoTBench). MMVU targets expert-level multi-discipline video understanding, including questions from science, engineering, healthcare, etc. We use 625 multiple-choice problems of it. VSI-Bench is designed to evaluate the visual-spatial intelligence of LMMs, consisting of over 5,000 problems derived from 288 egocentric videos, including object count, route plan, relative direction, etc. MVbench contains 4,000 problems involving understanding actions, position, count, etc. Temporal-sensitive benchmarks are enhanced for temporal understanding, which rely more heavily on multi-frame understanding. Specifically, TempCompass constructs conflicting videos which share the same static cues but differ in a specific temporal aspect, explicitly requiring correct understanding across the temporal span rather than single-frame information. TVbench defines 10 temporally challenging tasks such as action count, temporal localization, and temporal sequential ordering. AoTBench is dedicated to assessing temporal direction sensitivity, which collects problems by filtering temporal-sensitive videos based on a temporal-related metric. These temporal-sensitive benchmarks are more effective for evaluating whether the model understands the whole video, as temporality is inherently built on multiple frames. A single frame has appearance and even motion cues, but it will never possess temporal information. Figure C.1 show some examples of temporal-sensitive benchmarks.

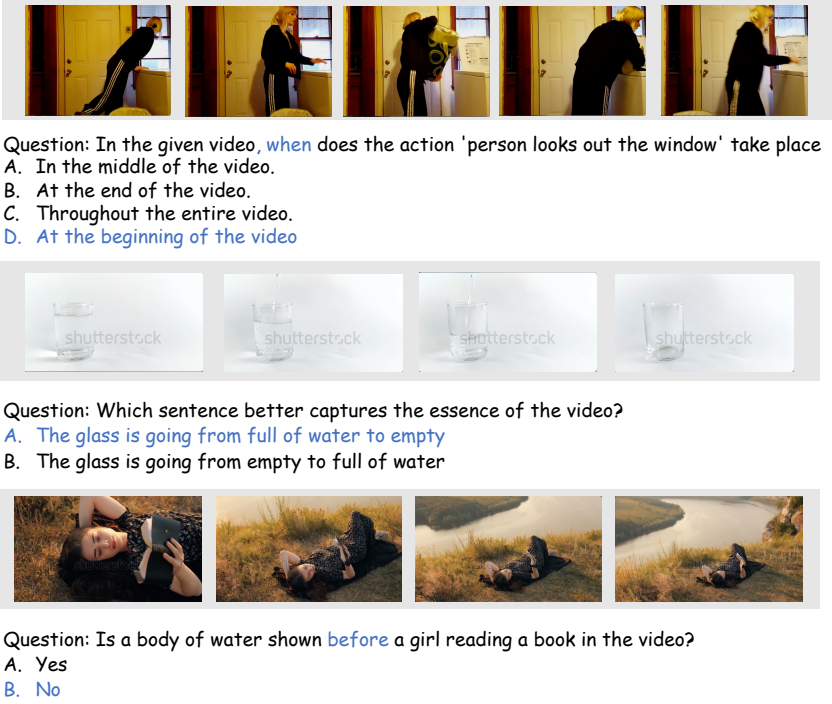


Figure C.1: Several examples of the temporal-sensitive benchmarks.

D ADDITIONAL ABLATIONS

We further compare MF-GRPO with two additional ablation baselines: curriculum-only training and random frame masking. As shown in Table D.1, curriculum-only training does not improve over Vanilla GRPO, indicating that the performance gain does not come from the progressive curriculum alone. Its benefit arises only when coupled with our MF advantage enhancement, where it enables the model to learn multi-frame focus progressively from fewer to more frames. Random frame masking improves performance, but remains inferior to our method, suggesting that retaining the first and last frames in the contrast video provides a more effective contrast for mitigating single-frame bias.

Table D.1: Comparison with additional control baselines.

Method	Avg. Acc.
Vanilla GRPO	54.0
Curriculum-only	53.8
Random frame masking	55.2
Ours	**56.0**

E Additional Qualitative Results

Figure E.1 presents two extra samples of the comparison between the baseline and our method. Only by understanding the entire video content can the model answer correctly, highlighting the significance of our multi-frame focus enhancement post-training.

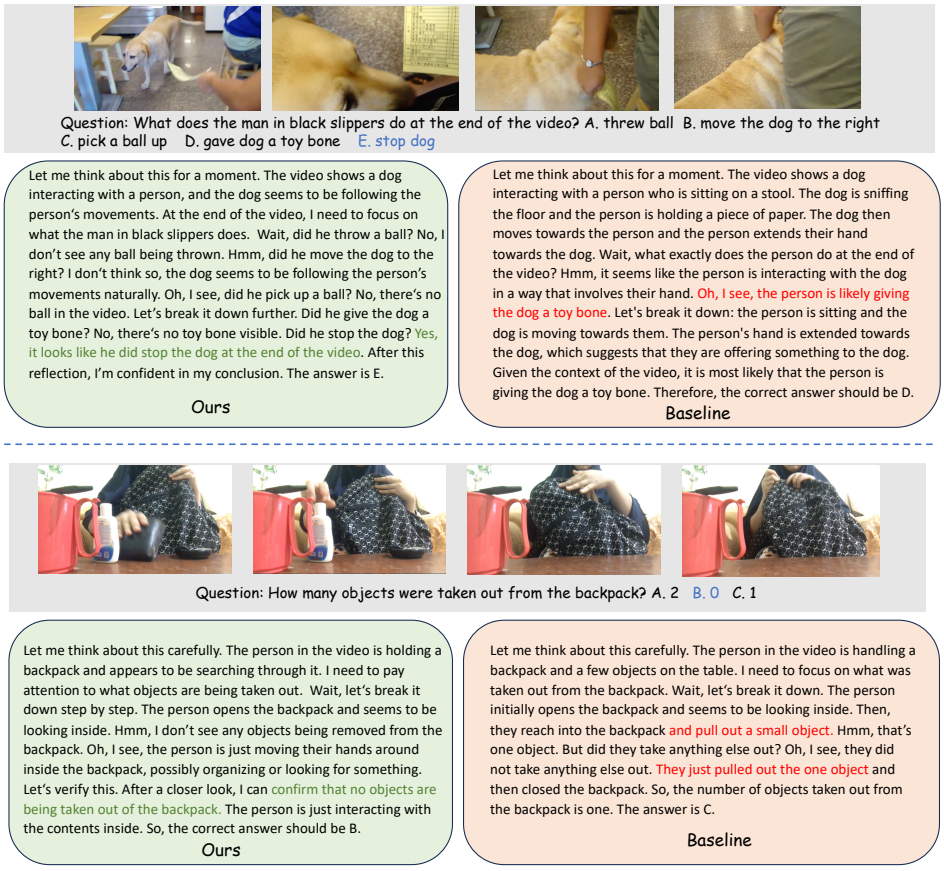


Figure E.1: Additional examples of comparison between baseline (vanilla GRPO) and our method. Only by understanding the entire video content can the model answer correctly.

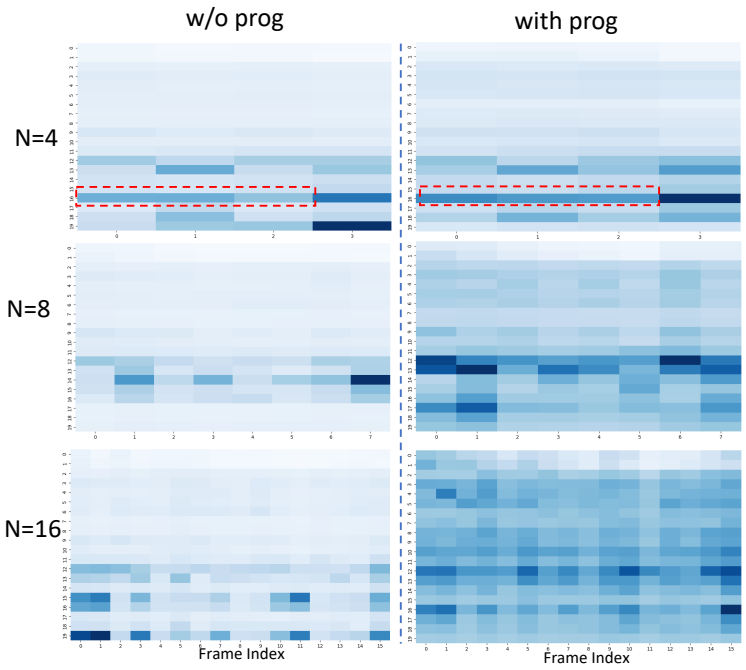


Figure E.2: **The visualization of the attention maps without and with our progressive learning.** For each N (the number of the input frames), the model with progressive learning allocates attention more broadly, mitigating the issue of overly relying on a single frame. Besides, the model without progressive learning becomes more difficult to focus on multiple frames as the frame number increases from 4 to 16.

Algorithm E.1 The proposed MF-GRPO algorithm.

for $i = 1$ **to** M **do**
  Stage $= S_i$, frame number $N = N_i$, training steps $D = D_i$
  for step $= 1$ **to** D **do**
    $V = \{v_t\}_{t=1}^{N}$ ▷ original video with N frames
    $\tilde{V} = \{v_1, v_N\} \cup \{v_{t_i}\}_{i=1}^{N'-2}$ ▷ contrast video
    $\{o_i\}_{i=1}^{G} \sim \pi_\theta(O \mid q, V)$ ▷ sample group
    $\tilde{o} \sim \pi_\theta(o \mid q, \tilde{V})$ ▷ sample contrast
    Calculate the reward R_i
    Calculate the advantages A_i
    Obtain the enhanced advantages $A_i^{'}$ (Eq. 4)
    Optimize π_θ (Eq. 5)
  end for
end for
Output: The optimized model π_θ