LEARNING TO REASON EFFICIENTLY IN VIDEOS: REINFORCEMENT LEARNING WITH SHORT- AND LONG-TERM REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in multimodal reasoning models have shown promising results, but these models often inherit a language-centric reasoning bias and fall short in true multimodal understanding. In long video understanding, a core challenge in multimodal reasoning, such models tend to generate lengthy and reflective chain-of-thought (CoT) outputs that do not meaningfully contribute to video comprehension. Moreover, the excessive token generation during reasoning leads to inefficiencies in further training. To address these issues, we propose a simple post-training framework for building efficient video reasoning models. Specifically, to mitigate the text-biased reasoning patterns inherited from language models, we first distill an Instruction multimodal model using the video reasoning trajectories of a native multimodal reasoning model, allowing it to learn dynamic video reasoning. During the reinforcement learning stage, to improve data efficiency, we leverage the instruction-tuned model to filter long-form video question-answering data, retaining only a compact set of high-quality training samples. For reward design, we introduce a hybrid mechanism that incorporates both short-term and long-term rewards, optimized via group relative policy optimization (GRPO). Extensive experiments on diverse video benchmarks show that our framework transforms Qwen3-VL-Instruct into a more accurate and efficient video reasoning model, substantially outperforming the open-source Qwen3-VL-Thinking baseline. Consistent gains on Qwen2.5-VL and Qwen3.5 further demonstrate its generalizability. These results highlight the importance of dynamic reasoning and efficient reward design for advancing multimodal video reasoning.

1 INTRODUCTION

Recently, multimodal reasoning models based on multimodal large language models (MLLMs) have made great progress in image reasoning and video reasoning tasks (ByteDance, 2025a;b; Wang et al., 2025d; Wang & Peng, 2025; Feng et al., 2025a). Existing methods mainly first perceive images or videos, and then use reinforcement learning (RL) to incentivize linguistic reasoning ability (Feng et al., 2025a; Wang et al., 2025b). This kind of paradigm has shown promising results for image reasoning (Zhang et al., 2025b; Xia et al., 2025), but for more complex video reasoning, it is unable to effectively reason about the spatio-temporal information.

To investigate whether stronger reasoning capabilities necessarily improve video understanding, we compare Qwen3-VL-Thinking and Qwen3-VL-Instruct (Bai et al., 2025a) on six mainstream video benchmarks, as shown in the upper part of Fig. 1. Surprisingly, the thinking model underperforms the instruction model on five benchmarks and achieves an improvement only on MMVU (Zhao et al., 2025), which requires more specialized knowledge and complex reasoning. These results suggest that stronger linguistic reasoning does not automatically translate into better video understanding. Instead, effective video reasoning must remain grounded in spatiotemporal evidence and allocate an appropriate amount of computation according to question difficulty.

One possible reason lies in the construction and training of current video reasoning models. Although Qwen3-VL-Thinking uses video data during supervised and long-CoT fine-tuning,

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.

its reinforcement learning stage relies only on image data, leaving the final reasoning policy insufficiently optimized for dynamic visual information. Existing video CoT construction methods also have limitations. Video-R1 (Feng et al., 2025a) prompts a non-thinking MLLM to generate long reasoning chains, which may introduce inaccurate or weakly grounded rationales (Feng et al., 2025a; Li et al., 2025). VideoRFT (Wang et al., 2025b), by contrast, uses a language reasoning model that depends on MLLM outputs or clip captions rather than directly observing the video, making it difficult to preserve fine-grained temporal and cross-frame information. Moreover, these methods often produce similarly long responses for questions of different complexity. The lower part of Fig. 1 further compares our model with Video-R1 and VideoRFT across different difficulty levels. Our model achieves the highest accuracy of 69.9% while producing substantially more concise responses. Its reasoning length increases with question difficulty, while incorrect answers—typically associated with harder or more uncertain questions—tend to involve longer reasoning chains than correct ones. This indicates that the model learns to allocate reasoning effort according to the demands of each question rather than generating uniformly verbose CoTs. These observations motivate the construction of higher-quality reasoning trajectories that are directly grounded in dynamic video content.

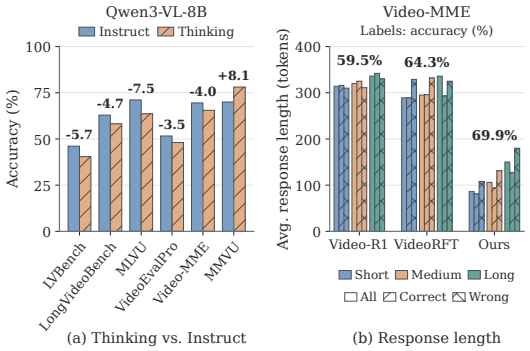


Figure 1: Comparison of reasoning effectiveness in video understanding. Top: Qwen3-VL-Thinking underperforms Qwen3-VL-Instruct on most benchmarks. Bottom: our model achieves higher accuracy on Video-MME with difficulty-adaptive and more concise reasoning than others.

Beyond data quality, most existing video RL training methods rely solely on final outcomes for reward evaluation, failing to leverage process-level reasoning that captures video dynamics. Our key insight is that incorporating reward evaluation on the reasoning process itself is crucial for video RL training. Recent works (Bai et al., 2025a; Feng et al., 2025b;a) only use verifiable multiple-choice questions or compute result similarity for open-ended questions to obtain outcome rewards. A recent approach (Wang et al., 2025b) introduces a semantic-consistency reward to measure alignment between the reasoning trace and the input video. However, it extracts only video description sentences from the trace, uses separate text and image encoders to compute cosine similarity between text features and averaged video features. This method cannot evaluate rewards for reasoning steps beyond video descriptions, nor can it assess the contribution of these steps toward arriving at the correct outcome.

To address this issue, we propose a novel method for measuring process rewards, enabling reward computation at any token after the reasoning process begins in the thinking model. Specifically, at any point in the reasoning token sequence, we take the existing reasoning trace as context, append a simple prompt, and feed it into the current policy model to obtain an immediate answer. This answer can be directly compared with the ground truth, allowing us to compute an accurate reward that reflects the importance of the current reasoning trace in reaching the correct outcome. During RL training, we select a reward from an early reasoning stage as the short-term reward and a reward from a later stage as the long-term reward. We then design a reward function with prior knowledge to compute the final reward for each sampled trajectory, and apply group relative policy optimization (Guo et al., 2025a) to calculate advantages and optimize the model.

Overall, our proposed video reasoning training framework consists of two stages. In the long-CoT SFT stage, we curate 22k video question-answer samples to enhance foundational reasoning capabilities, and leverage dynamic CoT data generated by the multimodal thinking model Seed1.8 to distill dynamic reasoning from a stronger model to the base model. In the RL stage, we introduce a multiple pre-sampling method to filter low-quality samples, yielding 500 high-quality video QA instances for RL training. We then propose a novel approach to accurately measure process rewards and design a prior-knowledge-based reward function, followed by group relative policy optimization for model update. Comprehensive experiments on video understanding and reasoning benchmarks demonstrate that our framework can effectively build stable and accurate video reasoning models.

2 Related Work

2.1 Multimodal Large Language Models

MLLMs (Hurst et al., 2024; Anthropic, 2024; ByteDance, 2025a;b; Zhu et al., 2025; Bai et al., 2025a) have achieved significant advancements in multimodal tasks, such as image and video understanding. For image understanding, the early method (Li et al., 2022) uses Q-Former to aggregate visual tokens to inject visual information into large language models. Then, the mainstream method to build MLLMs is to train a simple MLP projector with a pretraining stage and an instruction tuning stage (Li et al., 2024b; Liu et al., 2024a;b). For video understanding, a simple way is to treat video input as a sequence of images, and leverage the existing MLLMs trained on image data (Liu et al., 2024b) to a zero-shot video understanding model, or further fine-tune the model on video data to achieve better video understanding capabilities (Zhang et al., 2024a). Also, there are some works (Zhang et al., 2023; Ren et al., 2024; Wang et al., 2024d) that introduce Q-Former to aggregate temporal information and reduce the video tokens, or use simple MLP projectors for effectively injecting video information into large language models (Liu et al., 2025a).

To enhance temporal awareness in video understanding, some works propose explicit temporal textual prompts (Ren et al., 2024), temporal modules (Zeng et al., 2025), and MRoPE techniques (Bai et al., 2025b). As for video training, many work adopt a hybrid data training strategy. For example, LLaVA-OneVision (Li et al., 2024a), InternVL2.5 (Chen et al., 2024), and Qwen2.5-VL (Bai et al., 2025b) are trained on a combination of single images, multi-frame image sequences, and videos. Additionally, post-training techniques are widely used to improve video reasoning performance (Bai et al., 2025b; Zhu et al., 2025; Guo et al., 2025b; Bai et al., 2025a), such as continue training, strong-to-weak distillation (Agarwal et al., 2024), preference learning (Rafailov et al., 2024), and reinforcement learning (Guo et al., 2025a). In parallel, video benchmarks have been introduced to assess the various MLLMs, such as general video understanding tasks (Fu et al., 2024; Wang et al., 2024b; Wu et al., 2024; Zhou et al., 2024) and video reasoning tasks (Yang et al., 2024; Zhao et al., 2025; Shangguan et al., 2024). More recently, the research frontier has shifted toward reinforcement learning-based reasoning enhancement (Wang & Peng, 2025; Feng et al., 2025a; Wang et al., 2025b; Feng et al., 2025b), tool-integrated video agents (Zhang et al., 2025a), and evolving into video agents are the cutting-edge directions for the development of MLLMs towards more powerful and intelligent video understanding.

2.2 Multimodal Reasoning Models

Large language reasoning models have achieved great success in reasoning and instruction-following tasks, such as mathematics, coding, and agentic tasks. This success is driven by techniques including Chain-of-Thought (CoT) reasoning (Wei et al., 2022), test-time scaling (Jaech et al., 2024), RL (Shao et al., 2024), and agentic RL, etc. Inspired by DeepSeek-R1 (Shao et al., 2024; Guo et al., 2025a), which show that large-scale RL with verifiable rewards unlocks emerging reasoning abilities in LLMs, there has been growing interest in incorporating reasoning into MLLMs. This direction has shown encouraging progress, particularly through a line of research (Xu et al., 2024; Yang et al., 2025b; Liu et al., 2025b; Peng et al., 2025; Wang et al., 2025c) investigating RL for image-based reasoning tasks. To explore the multimodal reasoning effect for complex video understanding tasks, some works focus on step-by-step reasoning (Fei et al., 2024), CoT training (Wang et al., 2024c), and RL training (Wang & Peng, 2025). For video reasoning models' RL training, several work (Wang & Peng, 2025; Feng et al., 2025a; Wang et al., 2025b) introduce GRPO training for MLLM to reasoning for fully understanding the video-language relationship before the final answer. More generally, recent RL training frameworks typically incorporate both offline and online training stages. Seed1.5-VL (Guo et al., 2025b), for example, integrates video data into pretraining and designs a post-training phase that combines SFT with RL. Keye-VL-1.5 (Yang et al., 2025a) uses a slow-fast video encoder with iterative SFT and RL training, while InternVL3.5 (Wang et al., 2025e) proposes a cascade RL framework that consists of a mixed preference optimization and an online RL stage. Qwen3-VL (Bai et al., 2025a) designs long-CoT cold start, strong-to-weak distillation, general and reasoning RL training strategies to build multimodal reasoning models. Recent works, such as OneThinker (Feng et al., 2025b), have made progress by collecting image and video data across diverse visual tasks to build unified reasoning models. However, these approaches largely transfer image-based reasoning paradigms to video, overlooking the unique challenges of dynamic

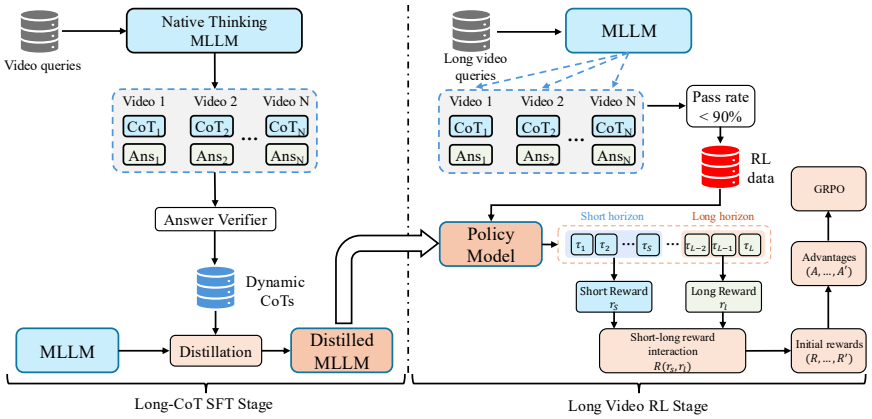


Figure 2: Overview of the proposed framework. Left: long-CoT SFT with dynamically generated and verified reasoning data. Right: long-video RL with interaction between short-horizon and long-horizon rewards.

video reasoning. In contrast, this work specifically targets dynamic video reasoning by designing a more efficient reinforcement learning method tailored for long-form video content.

3 METHOD

Currently, video reasoning models are that the models can first reason (textual reasoning or multimodal reasoning) and then finalize the answer. This is different from Instruct models, which directly output the answer and are proficient in doing so, but are not good at reasoning. Given a video V and a query Q, the model M can generate the reasoning content T_{think} and answer T_{ans} as follows:

$$[T_{\text{think}}, T_{\text{ans}}] \sim P(T \mid V, Q), \tag{1}$$

where P denotes the conditional probability distribution implicitly modeled by model M, and the reasoning content is enclosed within special tags (such as ¡think¿ and ¡/think¿), while the answer is enclosed within special tags (such as ¡answer¿ and ¡/answer¿). Meanwhile, the result tokens are the output results of the model conditioned on the original input and the existing reasoning content:

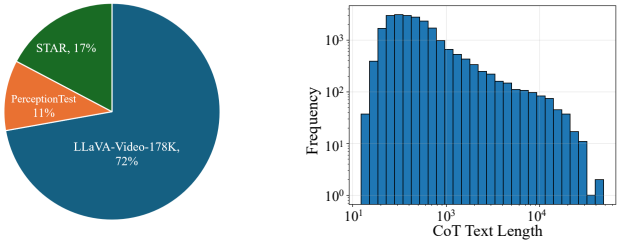
$$T_{\text{ans}} \sim P(\cdot \mid V, Q, T_{\text{think}}). \tag{2}$$

Regarding the benefits of the reasoning content, one is to provide the interpretability of the model for perception, and the other is to enhance the multimodal understanding ability through reasoning, which can solve more complex tasks.

Although there are thinking-version video reasoning models available in open-source models at present (such as Qwen3-VL-Thinking (Bai et al., 2025a)), due to the complexity of the post-training process involved, it is difficult to further enhance the capabilities based on this version. Therefore, the industry consensus is that the construction of reasoning capabilities in the Instruct model is more common and effective. Hence, the purpose of the framework proposed in this paper is to start from the Instruct model and build a more flexible and effective video reasoning model. Our proposed framework has a long-CoT SFT stage with dynamic video data, and an RL stage with short and long-term rewards as shown in Figure 2.

3.1 LONG-COT SFT STAGE WITH DYNAMIC VIDEO DATA

Long-CoT SFT, as a cold start stage, is a fundamental stage that establishes the basic reasoning abilities for multimodal models. Existing methods use relatively fixed video annotation data, such as using non-thinking models or using linguistic thinking models for annotation. Video-R1 uses a non-thinking model Qwen2.5-VL-72B to generate CoT rationales for every image and video. Video-RFT first prompts GPT-4o-mini to generate textual descriptions for every video, then uses a linguistic thinking model, DeepSeek-R1, to answer the question and reason step by step, and finally collects the reasoning CoTs.



(a) Video Source (b) CoT Text Length Distribution

Figure 3: Statistical data of Video CoT data. (a) Sources of video data. (b) Distribution of CoT text length.

These methods do not take into account the dynamics of video data and tend to execute very long CoTs for each video question. The dynamic nature of video data lies in the fact that if the video content is simple or the question is simple, then the CoT will be shorter. However, if the video content or the question is complex, a longer CoT is required. To introduce the dynamic capability of reasoning, we chose to use a native multimodal thinking model, Seed1.8 (ByteDance, 2025b) as the strong model, and used its reasoning trajectory to distill the base model. The video data is collected from several public sources, including PerceptionTest (Patraucean et al., 2023), STAR (Wu et al., 2021), LLaVA-Video-178k (Zhang et al., 2024b), and we only use multi-choice video QA data.

We leverage a native multimodal thinking model, Seed1.8, which can adjust the reasoning effort as needed. There are four modes (minimal, low, medium, and high). We set the highest mode as "high" so that it can conduct an in-depth analysis of complex videos. After the annotation process, we extract the reasoning content and the answers, and filter out the samples that did not match the ground truth answers. Ultimately, we obtain 22k CoT annotation data. The video source of our collected videos is shown in the left figure in Figure 3. The distribution of the text length of the reasoning content is shown in the right figure in Figure 3. Based on these video CoT data with dynamic thinking content, we fine-tune the model by maximizing the log-likelihood of the annotated reasoning CoT and answer $(T_{\text{think}}, T_{\text{ans}})$ given the input (V, Q), where the conditional distribution $P(T \mid V, Q)$ is defined in Eq. (1). The model used here is Qwen3-VL-Instruct-8B. At this stage, not only does the model learn how to perform video reasoning, but it also enables it to think dynamically based on different levels of difficulty.

3.2 RL Stage with Short and Long-term Rewards

The RL stage is an effective way to enhance the reasoning ability of thinking models. For example, the RL algorithms Group Relative Policy Optimization (GRPO) and Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) are widely used to incentivize linguistic reasoning and multimodal reasoning abilities. Regarding video reasoning reinforcement learning training, the existing methods do not adequately take into account the characteristics of videos when designing the algorithms or rewards. Qwen3-VL only used image data in the RL stage and did not utilize video data. During the training of its reasoning ability, it was prone to neglect the spatiotemporal information of videos. Recent work in RL training has incorporated image and video data, but only used outcome rewards, or considered semantic-consistency rewards for reasoning CoT. The reward for the reasoning process has not been fully explored.

To evaluate the contribution of a reasoning prefix, we prompt model M with the video V, query Q, and the first $t+1$ reasoning tokens $[T^0_{\text{think}}, \ldots, T^t_{\text{think}}]$ to elicit an intermediate answer, using the template in Table 1.

```
System-message <STOP>
User:  The assistant first thinks about the reasoning process in the
mind and then provides the user with the answer.
Question:  {V} {Q} \n <STOP>
Assistant:  [T^0_think,...,T^t_think]
The most correct option is:
```

Table 1: Prompt template for eliciting intermediate answers.

Table 2: Results on video understanding benchmarks compared with state-of-the-art models. **Bold** indicates the best performance in Thinking Models.

Model \Benchmark	LVBench	LongVB	MLVU	VEP	Video-MME			
		val	M-Avg	mcq	Short	Medium	Long	Overall
Non-Thinking Models								
Qwen2-VL-Base (Wang et al., 2024a)	-	53.7	-	-	70.3	59.5	49.6	59.8
Qwen2-VL-Ins (Wang et al., 2024a)	40.9	57.1	65.1	44.7	72.7	61.4	51.7	61.9
Qwen2.5-VL-Ins (Bai et al., 2025b)	43.3	59.3	67.7	44.4	76.4	65.2	54.0	63.2
Qwen3-VL-Ins (Bai et al., 2025a)	46.1	62.9	71.1	51.6	80.4	69.0	59.1	69.5
Thinking Models								
Video-R1-7B (Feng et al., 2025a)	38.6	49.7	63.0	42.2	72.2	59.4	47.0	59.5
VideoChat-R1-7B (Li et al., 2025)	-	-	-	-	-	-	-	60.0
LongVILA-R1-7B (Chen et al., 2025)	-	58.0	-	-	71.9	62.0	53.3	62.4
Open-o3 Video (Meng et al., 2025)	-	-	-	-	-	-	54.9	63.6
VideoRFT-7B (Wang et al., 2025b)	41.4	56.5	65.6	44.8	76.0	64.2	52.7	64.3
Qwen3-VL-Thinking (Bai et al., 2025a)	40.4	55.1	63.6	48.1	77.2	65.3	53.9	65.5
OneThinker-8B (Feng et al., 2025b)	-	61.7	-	-	-	-	-	66.5
AdaTooler-V-7B (Wang et al., 2025a)	-	-	-	-	-	-	-	66.7
Qwen3-VL-Ins (Force-think, base)	42.2	58.2	66.5	47.4	78.2	66.9	54.8	66.6
+ GRPO (Guo et al., 2025a)	45.1	60.8	69.0	50.0	79.0	68.4	54.7	67.4
+ Ours	**47.3**	**63.1**	**70.3**	**52.5**	**81.3**	**71.2**	**57.2**	**69.9**

In this way, we can effectively measure the reward of existing reasoning content at any time, and this reward is closely related to the expected answers, reflecting the importance of the reasoning content to the final outcome. Further, we consider the short-term reward r_s and long-term reward r_l to comprehensively evaluate the reward R for the trajectory. We define the interaction reward:

$$R = \underbrace{(2r_l - 1)}_{\text{outcome reward}} + \underbrace{r_l(1 - r_s)}_{\text{correction bonus}} - \underbrace{r_s(1 - r_l)}_{\text{degradation penalty}} = 3r_l - r_s - 1, \qquad r_s, r_l \in \{0, 1\}. \tag{3}$$

This reward incorporates prior knowledge about reasoning dynamics by separating final correctness from changes in reasoning quality. The outcome term $2r_l - 1$ assigns a positive or negative base reward according to the final result. The correction term $r_l(1-r_s)$ adds a bonus only for a transition from 0 to 1, indicating successful reflection and correction. Conversely, the degradation term $r_s(1 - r_l)$ penalizes a transition from 1 to 0, indicating that subsequent reasoning introduces errors. The reward are shown in Table 3. Additionally, overly short responses receive ± 1 based on correctness, whereas overlong responses without a final answer receive -3. See Appendix C.3 for details.

Table 3: Reward values for different combinations of r_s and r_l, where $r_s, r_l \in \{0, 1\}$.

(r_s, r_l)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
R	-1.0	2.0	-2.0	1.0

In this way, we first sampling multi trajectories $\{T_1, T_2, \cdots, T_G\}$ from a single video query, and we use the interaction reward function to score the trajectories, yielding G rewards $\{R_1, R_2, \cdots, R_G\}$. Inspired by GRPO, we use the group normalized reward as the advantages for all tokens, i.e., $\hat{A}_{i,t} = \frac{(R_i - \texttt{mean}(R))}{\texttt{std}(R)}$. These advantages are then incorporated into the model update following the same optimization approach as GRPO. During the training process, we observed that the accuracy on many samples was excessively high—meaning that the trajectories obtained from multiple samplings were mostly correct. This led to the issue of gradient vanishing. While this phenomenon does not theoretically hinder the model's learning capability, it significantly affects the stability of the training process and results in considerable fluctuations in model performance. We propose to collect high-quality video data for RL training. We choose LongVideo-Reason (Chen et al., 2025) training set as the data source, which has long video samples. We use the Qwen3-VL-8B-Instruct model to generate 16 responses per video query and filtered out any samples with an average accuracy exceeding 90%. We used the remaining 0.5k samples for training, which encouraged the model to engage in more exploration. The presence of both accurate and inaccurate sampled trajectories provided valuable feedback for model training. Additionally, as we focused on RL training over long videos, the spatiotemporal information in long videos is progressively reasoned through step-by-step reasoning tokens. This aligns well with our approach of coordinating short-term and long-term rewards, ultimately making the model training more effective.

Table 4: Performance on public video reasoning benchmarks compared to previous models. **Bold** indicates the best performance in Thinking Models.

Model \Benchmark	VSI-Bench	MMVU mc	LongVideo-Reason eval	Tomato
Non-Thinking Models				
LLaVA-OV-7B (Li et al., 2024a)	32.4	49.2	-	25.5
ViLA-1.5-8B (Lin et al., 2024)	28.9	49.2	-	-
VideoTree (Wang et al., 2025f)	-	54.2	-	-
LLaVA-Video-7B (Zhang et al., 2024b)	36.2	60.2	-	24.9
Qwen2-VL-7B-Ins (Wang et al., 2024a)	33.4	63.5	-	30.1
Qwen2.5-VL-7B-Ins (Bai et al., 2025b)	38.1	68.8	-	25.6
Qwen3-VL-8B-Ins (Bai et al., 2025a)	59.4	70.0	72.2	32.9
Thinking Models				
Video-R1-7B (Feng et al., 2025a)	37.8	64.3	68.1	26.9
VideoRFT (Wang et al., 2025b)	36.8	68.5	75.3	24.9
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	**56.6**	**78.1**	71.5	31.2
Qwen3-VL-Ins (Force-think, base)	52.0	68.6	73.4	33.0
+ GRPO (Guo et al., 2025a)	54.1	70.4	75.8	33.0
+ Ours	54.4	71.0	**76.1**	**33.6**

4 EXPERIMENT

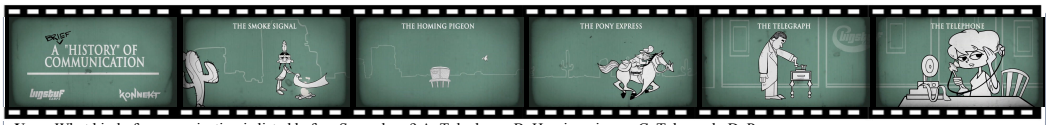
4.1 SETUP

We use Qwen3-VL-8B-Instruct as the base model, with all parameters of the MLLM fully fine-tuned. We use SWIFT as the codebase to perform multimodal SFT and RL training. All experiments are conducted on H20 GPUs. We sample 128 frames per video for both the training and inference stages. For thinking models, we set the maximum number of output tokens to 2048. In the Long-CoT SFT stage, the model was fine-tuned on CoT samples for 3 epochs, which takes 48 hours. In the RL stage, it was fine-tuned on RL samples for 1 epoch, taking 10 hours. We evaluate our models and state-of-the-art models on 5 video general understanding benchmarks, including Video-MME (Fu et al., 2024), LongVideoBench (LongVB) (Wu et al., 2024), MLVU (Zhou et al., 2024), LVBench (Wang et al., 2024b), VideoEval-Pro(VEP) (Ma et al., 2025), and on 4 video reasoning benchmarks including VSI-Bench (Yang et al., 2024), MMVU (Zhao et al., 2025), LongVideo-Reason-eval (Chen et al., 2025) and Tomato (Shangguan et al., 2024). More details can be found in the Appendix A.

4.2 MAIN RESULTS

General Video Understanding. We provide the quantitative comparison with state-of-the-art models in Table 2. These benchmarks evaluate general comprehension capabilities for long videos, among which Video-MME also assesses understanding abilities for short and medium-length videos. Qwen3-VL-Thinking (Bai et al., 2025a) is an open-source multimodal thinking model, but it needs denser frames and more reasoning tokens to achieve comparable performance. OneThinker (Feng et al., 2025b) also uses Qwen3-VL-8B-Instruct as the base model, and it trains a reasoning model on multiple image and video understanding tasks. From the same base model, our framework builds a thinking model that is consistently better than OneThinker and GRPO on all benchmarks. In thinking models, our model shows promising results and achieves the best results on all benchmarks. Compared to non-thinking models, our model outperforms Qwen3-VL-8B-Ins on four out of five benchmarks. Compared to the open-source Qwen3-VL-Thinking, our model demonstrates stronger general video understanding capabilities under the same maximum output token limit.

Video Reasoning. In Table 4, we show a comprehensive comparison of 4 mainstream video reasoning benchmarks. These benchmarks assess capabilities in video reasoning, including specialized knowledge, complex reasoning, spatial reasoning, and temporal reasoning. Based on the Qwen3-VL-8B-Instruct model, our proposed method consistently outperforms GRPO on 4 video reasoning benchmarks. On VSI-Bench and MMVU, our method still does not surpass the Qwen3-VL-8B-Thinking model. The reason may be that longer chains of thought bring significant performance improvements on such benchmarks, as Qwen3-VL-8B-Thinking can output up to 2048 tokens in some cases, whereas our model typically outputs a maximum of only 512 tokens. On LongVideo-Reason

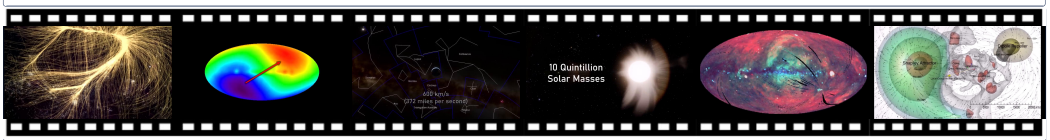


User: What kind of communication is listed before Semaphore? A. Telephone. B. Homing pigeon. C. Telegraph. D. Pony express.

VideoR1: <think>Let's break down the sequence of communication methods shown in the video:\n1. The video starts with *XXX*.\n2. Next, it shows *XXX*.\n3. Then, we see *XXX*.\n4. After the pony express, *XXX*.\n5. Following the *XXX*.\n\nTherefore, *XXX*:\n A. *XXX*.\n B. *XXX*.\n C. *XXX*.</think>\n<answer>B</answer> **(176 words)**

VideoRFT: <think>To determine the correct answer, I need *XXX*. Next, it shows *XXX*. Thus, the *XXX*. This confirms *XXX*.</think>\n<answer>D</answer> **(110 words)**

Ours: <think>Got it, let's check the video sequence. First, the order is: Smoke Signal, Homing Pigeon, Pony Express, Semaphore Flag. So before Semaphore (Semaphore Flag, which is at 28-31 seconds), the one before is Pony Express (21-27 seconds). Let's confirm the options: D is Pony express.</think><answer>D</answer> **(46 words)**



User: What aspect is NOT described in the video? A. The role of dark matter in the formation of the Great Attractor. B. Its gravitational influence on the local galaxy supercluster, Laniakea. C. The challenges posed by the Zone of Avoidance in studying the Great Attractor. D. Observational history and methods used to investigate the Great Attractor region.

VideoR1: <think>Let's break down the options one by one:\n\nA. *XXX*.\n\nB. *XXX*.\n\nC. *XXX*.\n\nD. *XXX*.\n\nAfter analyzing all the options, *XXX*. The video focuses more on *XXX*.\n\nTherefore, the aspect NOT described in the video is the challenges posed by the Zone of Avoidance.</think>\n<answer>C</answer> **(229 words)**

VideoRFT: <think>To determine which aspect is not described in the video, let's review the content step-by-step.\n\nThe video begins with *XXX*.\n\nNext, the video covers *XXX*.\n\nNow, let's evaluate each option:\n\nOption A: *XXX*.\n\nOption B: *XXX*.\n\nOption C: *XXX*.\n\nOption D: *XXX*.\n\nBased on this analysis, Option A is the aspect that is not described in the video.</think>\n<answer>A</answer> **(340 words)**

Ours: <think>Got it, let's analyze each option. \n\nFirst, option A: The video doesn't mention *XXX*.\n\nOption B: The video shows *XXX*.\n\nOption C: The Zone of Avoidance is *XXX*.\n\nOption D: The video covers *XXX*, so that's covered.\n\nSo the one not described is A.</think><answer>A</answer> **(128 words)**

Figure 4: Comparison on short and long video questions, "*XXX*" denotes compressed content.

Table 5: Ablation studies on training components and reward settings. The untrained baseline uses Qwen3-VL-8B-Instruct with forced thinking. Filter denotes RL data filtering; outcome denotes the default outcome reward. Short and Long specify the short- and long-term reward settings; – indicates that these rewards are disabled. The full model (h) serves as the shared reference.

ID	CoT-SFT	RL				LVBench	LongVB	MLVU	VEP	Video-MME
		Filter	Outcome	Short-long						
				Short	Long		val	M-Avg	mcq	overall
(a)				–	–	42.2	58.2	66.5	47.4	66.6
(b)			✓	–	–	44.6	58.1	67.4	49.1	67.1
(c)	✓			–	–	44.6	59.9	67.0	50.0	67.1
(d)	✓	✓	✓	–	–	45.1	60.8	69.0	50.0	67.4
(e)	✓	✓	✓	400	512	45.2	60.4	69.1	51.0	69.3
(f)	✓	✓	✓	300	512	45.6	60.9	69.0	51.8	69.7
(g)	✓	✓	✓	200	512	46.0	61.5	69.7	50.5	68.8
(h)	✓	✓	✓	100	512	**47.3**	**63.1**	**70.3**	**52.5**	**69.9**

and Tomato, our model achieves the best performance compared with existing thinking models and non-thinking models. This demonstrates the superiority of our model in long-video reasoning and temporal reasoning capabilities.

Qualitative comparison. Figure 4 presents two qualitative examples of video reasoning. In the upper case, our model produces a concise and accurate reasoning processes, whereas the compared methods generate unnecessarily lengthy reasoning traces. In the lower, more complex case, our model also performs detailed reasoning to arrive at the correct answer.

4.3 Ablation Study

The effect of proposed components. In Table 5, we evaluate the effectiveness of both the long-CoT SFT stage and RL stage. The first row is the base model, Qwen3-VL-8B-Instruct, which is forced to think without any training. The second row is that without the long-CoT SFT and directly fine-tune the base model by RL training with outcome reward on randomly sampled video data. The third row is that only fine-tunes the base model through the long-CoT SFT stage. Comparing these

Table 6: Generalization results on Qwen2.5-VL-7B and Qwen3.5-9B models. **Bold** indicates the best performance in Thinking Models.

Model \Benchmark	LVBench	LongVB	MLVU	VEP	Video-MME			
		val	M-Avg	mcq	Short	Medium	Long	Overall
Non-Thinking Models								
Qwen2.5-VL-7B-Ins (Bai et al., 2025b)	43.3	59.3	67.7	44.4	76.4	65.2	54.0	63.2
Qwen3.5-Base (Qwen Team, 2026)	50.2	68.8	78.5	60.2	81.1	72.6	60.8	71.5
Thinking Models								
VideoR1-7B (Feng et al., 2025a)	38.6	49.7	63.0	42.2	72.2	59.4	47.0	59.5
VideoRFT-7B (Wang et al., 2025b)	41.4	56.5	65.6	44.8	76.0	64.2	52.7	64.3
Qwen2.5-VL-Ins (Force-think, base)	39.0	44.7	57.0	42.2	72.7	62.5	49.7	61.6
+ GRPO (Guo et al., 2025a)	40.9	49.2	61.5	41.9	73.6	63.5	51.2	62.8
+ Ours	42.5	57.2	64.8	44.7	75.8	64.1	54.1	64.7
Qwen3.5-9B-Base	33.7	55.7	70.3	47.6	78.7	62.7	42.3	61.2
+ GRPO (Guo et al., 2025a)	56.1	65.5	77.2	55.8	82.1	73.8	63.0	72.9
+ Ours	**58.1**	**66.0**	**78.7**	**62.8**	**82.7**	**76.1**	**65.7**	**74.8**

two standalone training approaches, the results across five benchmarks consistently demonstrate that long-CoT SFT yields better performance gains than RL, which is also consistent with the current consensus in the field. The fourth row represents first going through the CoT SFT stage, followed by RL training with outcome reward using filtered RL data. Attributed to our data filtering approach, using the basic GRPO reinforcement learning algorithm can further enhance the performance of the CoT model. Further employing our proposed combination of short-term and long-term rewards leads to significant performance improvements on all benchmarks compared to using only outcome rewards. These ablation studies validate the effective contributions of our used long-CoT SFT, data filtering, and reward design methods to the overall approach.

The impact of different short- and long-term rewards. As shown in Table 5, we evaluate the short-term reward at positions 100, 200, 300, and 400, while fixing the long-term outcome reward at position 512. Results on LVBench, LongVideoBench, and MLVU show that as the short-term reward position increases, the model's reasoning accuracy consistently decreases. This suggests that for long video understanding tasks, leveraging both early short-term rewards and final outcome rewards is most effective. As the distance between the token positions used to compute the two rewards decreases, the resulting performance gains diminish. Specific reward combinations also achieve strong reasoning accuracy on short- and medium-video benchmarks and challenging benchmarks like VideoEvalPro. For instance, the 300/512 combination approaches our default 100/512 configuration on Video-MME and VideoEvalPro.

4.4 Generalization Experiments.

Table 6 evaluates our method on Qwen2.5-VL-7B-Instruct and Qwen3.5-9B. On both backbones, our method consistently outperforms GRPO across all benchmarks, demonstrating strong generalization across model families and scales. With Qwen2.5-VL-7B, it achieves the best results among thinking models on LVBench, LongVideoBench, and the Long and Overall settings of Video-MME, while remaining competitive on the other benchmarks. With Qwen3.5-9B, it achieves the best performance among thinking models across all evaluation settings. Notably, these improvements are obtained using substantially fewer training samples than other models.

5 Conclusion

In this work, we identify that existing multimodal reasoning models inherit a language-centric bias, resulting in ineffective and inefficient reasoning for long-video understanding. To address this issue, we propose a simple yet effective post-training framework that transforms instruction-tuned multimodal models into efficient video reasoning models. Our framework consists of a Long-CoT SFT stage, which equips the model with dynamic video reasoning capabilities, and an RL stage with interaction between short-term and long-term rewards to further improve reasoning accuracy and efficiency. Experiments on diverse video benchmarks show that our framework transforms Qwen3-VL-Instruct into a high-performing model that significantly outperforms Qwen3-VL-Thinking. Consistent gains on Qwen2.5-VL and Qwen3.5 further demonstrate its generalizability across backbone models and effectiveness in building accurate and efficient video reasoning models.

AI USE STATEMENT

Generative AI tools were used solely for grammar correction and language polishing and were not used for any other tasks. All AI-assisted revisions were reviewed and approved by two authors to ensure accuracy, appropriate wording, and consistency with the intended meaning. The authors take full responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth ICLR*, 2024.

Anthropic. Introducing the next generation of claude, 2024, 2024. https://www.anthropic.com/news/claude-3-5-sonnet.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, and et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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*, 2025b.

ByteDance. Introduction to techniques used in Seed1.6, 2025a. https://seed.bytedance.com/en/seed1_6.

ByteDance. Introduction to techniques used in Seed1.8, 2025b. https://seed.bytedance.com/zh/seed1_8.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

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. *arXiv preprint arXiv:2501.03230*, 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*, 2025a.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*, 2025b.

Chaoyou Fu, Yuhan Dai, Yondong 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. *arXiv preprint arXiv:2405.21075*, 2024.

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

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025b.

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.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900, 2022.

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 *CVPR*, pp. 26689–26699, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the CVPR*, pp. 4122–4134, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025b.

Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. Videoeval-pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*, 2025.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *NIPS*, 36:42748–42761, 2023.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NIPS*, 36, 2024.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the CVPR*, pp. 14313–14323, 2024.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models, 2024. URL https://arxiv.org/abs/2410.23266.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Chaoyang Wang, Kaituo Feng, Dongyang Chen, Zhongyu Wang, Zhixun Li, Sicheng Gao, Meng Meng, Xu Zhou, Manyuan Zhang, Yuzhang Shang, et al. Adatooler-v: Adaptive tool-use for images and videos. *arXiv preprint arXiv:2512.16918*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

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

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025c.

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

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

Xiaodong Wang and Peixi Peng. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.

Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*, 2024c.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pp. 396–416. Springer, 2024d.

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

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. *NIPS*, 35: 24824–24837, 2022.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth NIPS*, 2021.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *CoRR*, abs/2505.14677, 2025. doi: 10.48550/ARXIV.2505.14677. URL https://doi.org/10.48550/arXiv.2505.14677.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

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

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8579–8591, 2025.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving MLLMs for long video understanding via grounded tuning. In *The Thirteenth ICLR*, 2025. URL https://openreview.net/forum?id=nAVejJURqZ.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025a.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025b.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024a. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding, 2025. URL https://arxiv.org/abs/2501.12380.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 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

In addition to the experiments on Qwen3-VL, Qwen2.5-VL, and Qwen3.5 reported in the main paper, we further investigate the dependence on teacher-generated supervision, the frame-sampling configuration, long-context information retrieval, and cross-architecture generalization. In particular, we show that our RL objective remains effective without teacher-generated distillation data and extend the architecture coverage with Qwen2-VL.

A.1 Effectiveness without Teacher Distillation

Our method does not require teacher-generated distillation data to improve the base model. To isolate the contribution of our RL objective, we apply it directly to Qwen3-VL-8B-Instruct without the preceding teacher-SFT stage. As shown in Table A1, our RL-only variant improves the overall Video-MME score from 66.6 to 68.2 and outperforms GRPO-only by 0.9 points. It also exceeds GRPO-only by 3.2 points on the long-video subset and by 0.5 points on LongVideoBench. Although combining teacher SFT with our RL objective yields the strongest overall performance, the consistent gains of the RL-only variant demonstrate that the effectiveness of our RL method is not contingent on teacher supervision.

Table A1: Effectiveness of our RL objective with and without teacher-generated supervision.

Model	Video-MME				LongVB
	short	medium	long	overall	
Qwen3-VL-8B-Ins	78.2	66.9	54.8	66.6	58.2
+ Teacher-SFT only	78.5	68.0	55.0	67.1	<u>59.9</u>
+ GRPO only	79.0	<u>68.4</u>	54.7	67.3	58.1
+ Our RL only	<u>79.2</u>	67.6	**57.9**	<u>68.2</u>	58.6
+ Our SFT+RL	**81.3**	**71.2**	<u>57.2</u>	**69.9**	**63.1**

A.2 Frame Sampling

For a controlled and fair comparison, our method and all baselines use uniform temporal sampling during both training and evaluation. Although adaptive frame selection is an important research direction, this work focuses on improving video reasoning under the same standard uniform-sampling protocol. Table A2 compares matched training and evaluation budgets of 32, 64, and 128 frames. Our method consistently improves the Qwen3-VL-8B-Instruct baseline at every frame budget, yielding overall Video-MME gains of 3.2, 1.5, and 3.3 points with 32, 64, and 128 frames, respectively. These results show that the improvements are robust to the number of sampled frames rather than being tied to a particular frame budget.

Table A2: Impact of the number of uniformly sampled frames used during training and evaluation.

Model	Frames	Video-MME			
	train/eval	short	medium	long	overall
Qwen3-VL-8B-Ins	128/128	78.2	66.9	54.8	66.6
+ Ours	128/128	**81.3**	**71.2**	**57.2**	**69.9**
Qwen3-VL-8B-Ins	64/64	77.2	61.8	52.8	63.9
+ Ours	64/64	**77.6**	**65.4**	**53.3**	**65.4**
Qwen3-VL-8B-Ins	32/32	72.6	57.2	52.1	60.6
+ Ours	32/32	**76.7**	**61.7**	**53.0**	**63.8**

A.3 Video-NIAH Evaluation

We further evaluate long-context information retrieval on the Video-NIAH task from Haystack-LVBench (Ye et al., 2025). As shown in Table A3, our method outperforms both the base models and their GRPO-trained counterparts across two architectures. Specifically, it improves over GRPO

by 3.2 points on Qwen3-VL-8B-Instruct and by 3.5 points on Qwen2.5-VL-7B-Instruct, indicating a stronger ability to locate and retain relevant information in long videos.

Table A3: Video-NIAH evaluation on Haystack-LVBench.

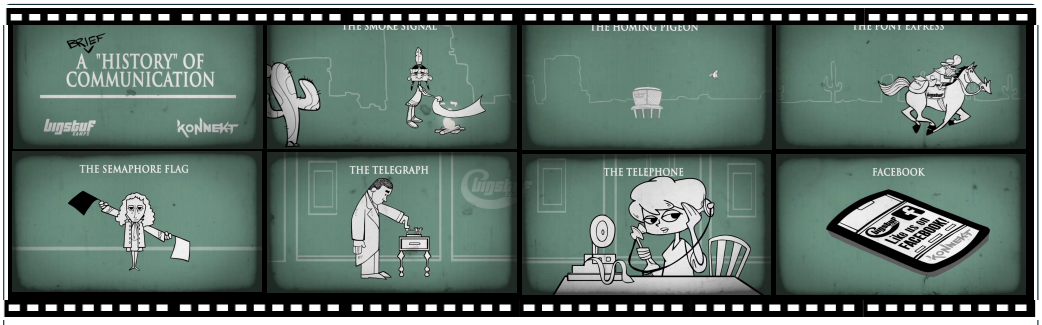
Model	Base	+ GRPO	+ Ours
Qwen3-VL-8B-Ins	51.7	51.5	**54.7**
Qwen2.5-VL-7B-Ins	43.6	47.4	**50.9**

A.4 Generalization across Model Architectures

The main paper evaluates our method on Qwen3-VL, Qwen2.5-VL, and Qwen3.5. To further assess whether its effectiveness transfers to an earlier and structurally distinct architecture, we additionally conduct experiments on Qwen2-VL. Table A4 reports the new Qwen2-VL results alongside Qwen2.5-VL for reference. On Qwen2-VL-7B-Instruct, our method improves the overall Video-MME score from 43.9 to 57.0 and the LongVideoBench score from 21.1 to 41.1. The consistent gains across both model families and all video-duration categories demonstrate that our method generalizes beyond the architectures evaluated in the main paper.

Table A4: Generalization results on Qwen2.5-VL and Qwen2-VL.

Model	Video-MME				LongVB
	short	medium	long	overall	
Qwen2.5-VL-7B-Ins	72.7	62.5	49.7	61.6	44.7
+ Ours	**76.5**	**65.4**	**54.4**	**65.4**	**57.2**
Qwen2-VL-7B-Ins	48.7	42.3	40.6	43.9	21.1
+ Ours	**68.5**	**56.1**	**46.4**	**57.0**	**41.1**



User: What kind of communication is listed before Semaphore? A. Telephone. B. Homing pigeon. C. Telegraph. D. Pony express.

VideoR1: <think>Let's break down the sequence of communication methods shown in the video:\n\n1. The video starts with \"The Smoke Signal,\" which is an early form of communication.\n2. Next, it shows \"The Homing Pigeon,\" indicating that this method comes after the smoke signal.\n3. Then, we see \"The Pony Express,\" which follows the homing pigeon.\n4. After the pony express, there is \"The Semaphore Flag.\"\n5. Following the semaphore flag, the video shows \"The Telegraph.\"\n\nSo, the order of communication methods in the video is: Smoke Signal -> Homing Pigeon -> Pony Express -> Semaphore Flag -> Telegraph.\n\nNow, the question asks what kind of communication is listed before Semaphore. From the above sequence, the communication listed immediately before Semaphore is the Pony Express.\n\nTherefore, the correct answer should be the Pony Express.\n\nLet's verify if any other options fit:\n- A. Telephone: This is not mentioned before Semaphore.\n- B. Homing pigeon: This is mentioned before Semaphore but not immediately before.\n- C. Telegraph: This is mentioned after Semaphore.\n\nThe most accurate and immediate predecessor to Semaphore is the Pony Express.</think>\n<answer>B</answer>

VideoRFT: <think>To determine the correct answer, I need to recall the sequence of communication methods presented in the video. The video begins with \"The Smoke Signal,\" followed by \"The Homing Pigeon.\" Next, it shows \"The Pony Express,\" then \"The Semaphore Flag,\" and finally \"The Telegraph.\" Therefore, the order is: Smoke Signal → Homing Pigeon → Pony Express → Semaphore Flag → Telegraph.\n\nGiven this sequence, the method listed immediately after \"The Homing Pigeon\" is \"The Pony Express.\" Thus, the correct answer is D. Pony Express.\n\nTo confirm, the video's sequence is: Smoke Signal → Homing Pigeon → Pony Express → Semaphore Flag → Telegraph. This confirms that the method listed right after Homing Pigeon is indeed Pony Express.</think>\n<answer>D</answer>

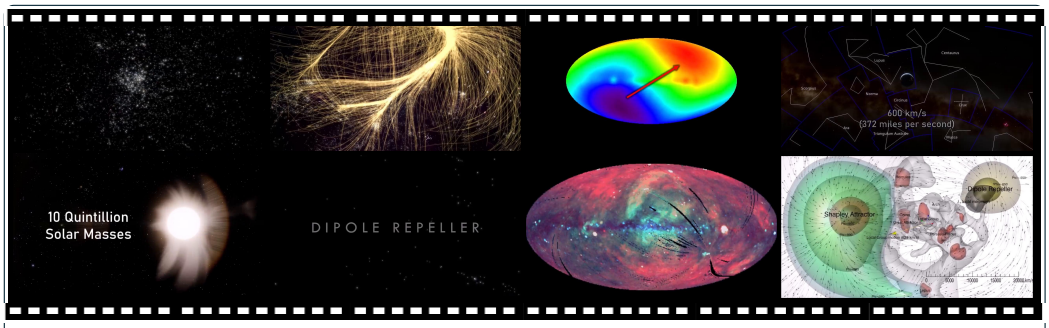
Ours: <think>Got it, let's check the video sequence. First, the order is: Smoke Signal, Homing Pigeon, Pony Express, Semaphore Flag. So before Semaphore (Semaphore Flag, which is at 28-31 seconds), the one before is Pony Express (21-27 seconds). Let's confirm the options: D is Pony express.</think><answer>D</answer>

Figure A1: Details of the comparison results on short video questions.

A.5 DETAILED QUALITATIVE COMPARISONS.

Due to space limitations, we combine the two qualitative examples and reorganize their content into a compact layout in the main text. Figures A1 and A2 provide the corresponding expanded results, including the complete reasoning traces generated by each method.

Figure A1 presents the complete outputs for the short-video example. The question requires identifying the communication method immediately preceding the Semaphore Flag in the video. Although VideoR1 reconstructs the correct sequence and even identifies the Pony Express as the immediate predecessor, its final answer incorrectly selects the Homing Pigeon, revealing an inconsistency between its reasoning and prediction. VideoRFT produces the correct answer but repeatedly restates the event sequence and performs redundant verification. In contrast, our model directly grounds the relevant events using their temporal positions, identifies the Pony Express at 21–27 seconds as preceding the Semaphore Flag at 28–31 seconds, and concisely returns the correct answer.



User: What aspect is NOT described in the video? A. The role of dark matter in the formation of the Great Attractor. B. Its gravitational influence on the local galaxy supercluster, Laniakea. C. The challenges posed by the Zone of Avoidance in studying the Great Attractor. D. Observational history and methods used to investigate the Great Attractor region.

VideoR1: <think>Let's break down the options one by one:\n\nA. The role of dark matter in the formation of the Great Attractor. - This seems relevant because the video likely discusses the gravitational effects of dark matter, which is crucial for understanding the formation of structures like the Great Attractor.\n\nB. Its gravitational influence on the local galaxy supercluster, Laniakea. - This also appears relevant as the video probably talks about the gravitational interactions between galaxies and the formation of larger structures like the Laniakea supercluster.\n\nC. The challenges posed by the Zone of Avoidance in studying the Great Attractor. - This is less likely to be discussed in the video. While the Zone of Avoidance is a real phenomenon, it is not mentioned in the context provided in the video summary.\n\nD. Observational history and methods used to investigate the Great Attractor region. - This is also relevant as the video likely covers the historical observations and techniques used to study this region.\n\nAfter analyzing all the options, it is clear that the challenge posed by the Zone of Avoidance in studying the Great Attractor is not explicitly described in the video. The video focuses more on the gravitational influences and observational methods rather than the specific challenges faced due to the Zone of Avoidance.\n\nTherefore, the aspect NOT described in the video is the challenges posed by the Zone of Avoidance.</think>\n**<answer>C</answer>**

VideoRFT: <think>To determine which aspect is not described in the video, let's review the content step-by-step.\n\nThe video begins with a view of space, featuring a galaxy and stars, and then zooms into Earth from space. It transitions to a black screen with text about the Great Attractor, followed by a depiction of the Milky Way and its surroundings. The video then shows a diagram of the Local Group and other nearby galaxies, including the Virgo Cluster and M87. It also includes a map of the cosmic web and a discussion of the Great Attractor's gravitational influence on the local supercluster, Laniakea.\n\nNext, the video covers the Zone of Avoidance, explaining how it affects our ability to study galaxies behind the Milky Way. It mentions the use of X-ray telescopes like ROSAT and the Anglo-Australian Telescope to identify hidden galaxies. The video also discusses the Shapley Supercluster and the Hercules-Corona Borealis Great Wall, as well as the Vela Supercluster.\n\nNow, let's evaluate each option:\n\nOption A: The role of dark matter in the formation of the Great Attractor. The video does not explicitly mention dark matter's role in the formation of the Great Attractor. While it discusses the Great Attractor's gravitational influence, there is no detailed explanation of dark matter's contribution to its formation.\n\nOption B: Its gravitational influence on the local galaxy supercluster, Laniakea. The video clearly shows the Great Attractor's gravitational influence on the local supercluster, Laniakea, using diagrams and text.\n\nOption C: The challenges posed by the Zone of Avoidance in studying the Great Attractor. The video explicitly discusses the Zone of Avoidance and its impact on studying galaxies behind the Milky Way, mentioning the use of X-ray telescopes to identify hidden galaxies.\n\nOption D: Observational history and methods used to investigate the Great Attractor region. The video covers observational history, including the use of X-ray telescopes and the Anglo-Australian Telescope, and discusses the methods used to identify hidden galaxies.\n\nBased on this analysis, Option A is the aspect that is not described in the video.</think>\n**<answer>A</answer>**

Ours: <think>Got it, let's analyze each option. \n\nFirst, option A: The video doesn't mention dark matter's role in forming the Great Attractor. The video talks about the Great Attractor's gravitational pull, the Zone of Avoidance, observational methods (like X-ray surveys, telescopes), and its influence on Laniakea (the supercluster). But dark matter's specific role in its formation isn't described.\n\nOption B: The video shows the Great Attractor's influence on Laniakea (the supercluster the Local Group is in), so that's covered.\n\nOption C: The Zone of Avoidance is mentioned as a challenge, since it's obscured by the Milky Way, making observation hard. That's in the video.\n\nOption D: The video covers observational history (like the CIZA project, ROSAT, Anglo-Australian Telescope, etc.), so that's covered.\n\nSo the one not described is A.</think>**<answer>A</answer>**

Figure A2: Details of the comparison results on long video questions.

Figure A2 shows the complete outputs for the more challenging long-video example. Answering this question requires integrating evidence distributed across the video and determining which candidate topic is not discussed. VideoR1 incorrectly selects the Zone of Avoidance, overlooking its role in obstructing observations of the Great Attractor. VideoRFT reaches the correct answer through an exhaustive option-by-option analysis, but produces a substantially longer reasoning trace. Our model verifies that the video covers the Great Attractor's influence on Laniakea, the observational challenges caused by the Zone of Avoidance, and the corresponding investigation methods. It then identifies the role of dark matter in the formation of the Great Attractor as the only undescribed aspect and returns the correct answer. Together, these examples illustrate that our model can adapt

its reasoning to videos of different lengths while maintaining both evidential grounding and response conciseness.

B Dataset Details

For a comprehensive evaluation, we conduct experiments on nine widely used video benchmarks. These include five general video-understanding benchmarks—Video-MME (Fu et al., 2024), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2024), VideoEvalPro (Ma et al., 2025), and LVBench (Wang et al., 2024b)—and four video-reasoning benchmarks—VSI-Bench (Yang et al., 2024), MMVU (Zhao et al., 2025), Tomato (Shangguan et al., 2024), and LongVideo-Reason (Chen et al., 2025).

Video-MME is a large-scale benchmark for evaluating multimodal understanding in video-based LLMs, covering tasks such as question answering, captioning, and spatiotemporal reasoning. It comprises a diverse collection of real-world videos with comprehensive annotations to support objective evaluation.

LongVideoBench (LongVB) focuses on assessing models' ability to process extended video content, requiring both fine-grained moment analysis and high-level summarization. Its annotated videos often span several minutes to hours, challenging existing context-length limitations.

Multi-Level Video Understanding (MLVU) provides a dataset encompassing tasks from basic object-action recognition to high-level story comprehension. It evaluates a model's capacity for multi-semantic reasoning across a wide range of real-world video clips.

VideoEval-Pro targets long-form video understanding with high-quality videos and queries aligned with real-world scenarios. It requires models to demonstrate robust performance in both multiple-choice and open-ended question answering.

LVBench features a diverse set of tasks aimed at long video comprehension and information extraction. It demands long-term memory and extended comprehension capabilities, with videos spanning several hours.

VSI-Bench is a video-based benchmark comprising nearly 5,000 queries across approximately 290 real-world indoor-scene videos. By leveraging continuous temporal input, it mirrors natural human observation and supports richer spatial understanding compared to static images.

MMVU is designed to evaluate models on expert-level, multi-discipline video understanding and reasoning. It includes 1,500 specialized-domain videos across 27 subjects in four key disciplines: Science, Healthcare, Humanities & Social Sciences, and Engineering, requiring domain knowledge and advanced reasoning skills.

Tomato is a video benchmark that rigorously assesses temporal reasoning by requiring models to interpret frames as a continuous sequence. It features 1,484 human-annotated questions across six dynamic tasks, drawing on diverse video sources that include both real-world and simulated content.

LongVideo-Reason is a high-quality long-video benchmark specialized for reasoning evaluation. Its annotations are derived from segmenting videos into short clips and generating detailed captions, with a focus on temporal, goal-oriented, spatial, and narrative reasoning.

C Experiment Details

C.1 Long-CoT supervised fine-tuning.

During the long-CoT SFT stage, we collect 22K video–question samples from VideoRFT (Wang et al., 2025b) and use Seed1.8 (ByteDance, 2025b) to generate their responses. We fine-tune Qwen3-VL-8B-Instruct with MS-Swift (Zhao et al., 2024) using the script below. Unless otherwise specified, we cap the number of sampled frames at 128 during training to match the default 128-frame evaluation setting.

```
CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 \
NPROC_PER_NODE=8 \
```

```
VIDEO_MAX_PIXELS=25176 \
FPS_MAX_FRAMES=128 \
MAX_PIXELS=503520 \
VIDEO_MAX_TOKEN_NUM=192 \
USE_AUDIO_IN_VIDEO=False \
swift sft \
    --model Qwen3-VL-8B-Instruct \
    --train_type full \
    --torch_dtype bfloat16 \
    --dataset {code start data} \
    --load_from_cache_file true \
    --max_length 20480 \
    --num_train_epochs 3 \
    --per_device_train_batch_size 1 \
    --per_device_eval_batch_size 1 \
    --learning_rate 1e-6 \
    --gradient_accumulation_steps 1 \
    --save_strategy steps \
    --eval_strategy steps \
    --eval_steps 100000 \
    --save_steps 500 \
    --save_total_limit 5 \
    --logging_steps 1 \
    --output_dir checkpoint \
    --warmup_ratio 0 \
    --dataloader_num_workers 0 \
    --deepspeed zero3_offload \
    --report_to wandb
```

C.2 RL TRAINING.

During the RL stage, we sample 16 candidate responses from the base model for each video–question pair in the LongVideo-Reason training set (Chen et al., 2025). We use the default "pt" sampling engine for efficient generation, as shown below:

```
FPS_MAX_FRAMES=128 CUDA_VISIBLE_DEVICES=${i} swift sample \
        --model Qwen3-VL-8B-Instruct \
        --sampler_engine pt \
        --num_return_sequences 8 \
        --max_new_tokens 1024 \
        --dataset {dataset} \
        --output_dir {output} \
        --system {system prompt}
```

We then remove samples whose average accuracy exceeds 90%, retaining 0.5K challenging, high-quality long-video examples for RL training. The corresponding training script is shown below:

```
CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 \
NPROC_PER_NODE=8 \
VIDEO_MAX_PIXELS=25176 \
FPS_MAX_FRAMES=128 \
MAX_PIXELS=503520 \
VIDEO_MAX_TOKEN_NUM=192 \
USE_AUDIO_IN_VIDEO=False \
swift rlhf \
    --rlhf_type grpo \
    --model {SFT_model} \
    --external_plugins ms-swift/examples/train/grpo/plugin/plugin.py \
    --reward_funcs external_acc_mc \
    --use_vllm false \
    --train_type full \
```

```
    --torch_dtype bfloat16 \
    --dataset {RL_dataset} \
    --load_from_cache_file true \
    --max_completion_length 2048 \
    --num_train_epochs 1 \
    --per_device_train_batch_size 1 \
    --per_device_eval_batch_size 1 \
    --learning_rate 1e-6 \
    --gradient_accumulation_steps 1 \
    --save_strategy steps \
    --eval_strategy steps \
    --eval_steps 10000 \
    --save_steps 100 \
    --save_total_limit 6 \
    --logging_steps 1 \
    --output_dir checkpoint \
    --warmup_ratio 0.01 \
    --dataloader_num_workers 4 \
    --num_generations 8 \
    --temperature 1.0 \
    --system ms-swift/examples/train/grpo/prompt.txt \
    --deepspeed zero3 \
    --log_completions true \
    --report_to wandb \
    --num_iterations 1 \
    --async_generate false \
    --beta 0.001 \
    --budget 512 \
```

During the RL stage, we extend MS-Swift to generate immediate answers within GRPO training and optimize the model using our short-term and long-term rewards. We train the GRPO baseline with the default MS-Swift implementation. For a fair comparison, all RL runs use the same maximum generation budget of 512 tokens.

C.3 Reward assignment.

Let r_s and r_l denote the binary correctness indicators of the intermediate answer obtained at the 100-token checkpoint and the final answer, respectively. For responses that reach this checkpoint and produce a valid final answer within the 512-token budget, Table A5 summarizes the reward R for each combination.

Table A5: Reward values for different combinations of r_s and r_l.

(r_s, r_l)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
R	-1.0	2.0	-2.0	1.0

We handle responses outside these four cases according to their termination conditions. If a response terminates before the short token budget, r_s is unavailable, and the response receives a reward of 1.0 for a correct final answer and -1.0 otherwise. If a response terminates between the short and long token budget with a valid final answer, its reward is still determined by (r_s, r_l) according to Table A5, regardless of its exact length. Finally, a response that exhausts the long token budget without producing a valid final answer receives a reward of -3.0, irrespective of its intermediate correctness.