

# SLTR-Video: Efficient Video Reinforcement Learning with Short- and Long-Term Rewards

Xiaodong Wang<sup>1,2</sup> and Peixi Peng<sup>1,2</sup>

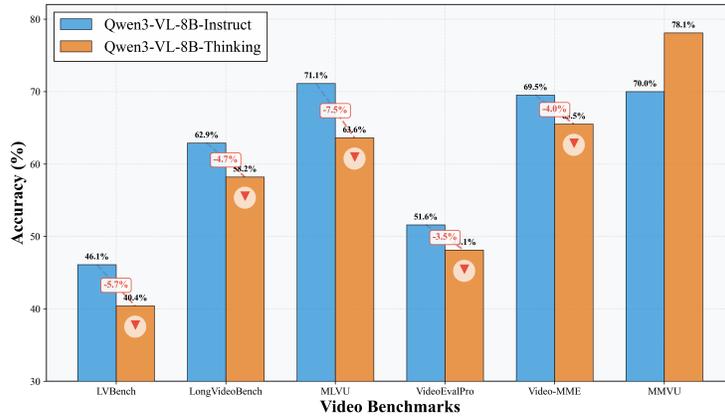
<sup>1</sup> School of Electronic and Computer Engineering, Peking University, Shenzhen, CHINA

<sup>2</sup> Pengcheng Laboratory, Shenzhen, CHINA  
wangxiaodong21s@stu.pku.edu.cn

**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 demonstrate that, through our post-training framework, the Qwen3-VL-Instruct-8B model is transformed into a more accurate and efficient video reasoning model, significantly outperforming the open-source Qwen3-VL-Thinking-8B baseline. Our framework highlights the importance of dynamic reasoning and efficient reward design, paving a new path for advancing multimodal reasoning in videos.

## 1 Introduction

Recently, multimodal reasoning models based on multimodal large language models (MLLMs) have made great progress in image reasoning and video reasoning tasks [5, 6, 11, 43, 44]. Existing methods mainly first perceive images or videos, and then use reinforcement learning (RL) to incentivize linguistic reasoning ability [11, 39]. This kind of paradigm has shown promising results for image reasoning [51, 61], but for more complex video reasoning, it is unable to effectively reason about the spatio-temporal information in videos. To delve into the impact



**Fig. 1:** Comparison of the instruct and thinking models of Qwen3-VL [3] on 128 video frames with a maximum of 2048 output tokens.

of reasoning ability in video understanding, we conducted a detailed comparison of the thinking version and the instruction version of the state-of-the-art model Qwen3-VL [3] on six mainstream video understanding benchmarks, as shown in Figure 1. The results show that the thinking model performs worse on five benchmarks, with improvements observed only on MMVU [65]. While MMVU demands greater expert knowledge and internal reasoning capability, the significant and widespread performance decline across other benchmarks is unsatisfactory, indicating that the thinking model still has major limitations in general video understanding.

Regarding the underperformance of the Qwen3-VL-Thinking model on video tasks, it is important to note that its training methodology follows the paradigm we previously identified: “First perceive images or videos, and then use RL for linguistic reasoning.” Specifically, while the model undergoes supervised fine-tuning (SFT) and long chain-of-thought (CoT) SFT on video data, the subsequent RL stage is trained exclusively on image data. This highlights a key insight: incorporating dynamic video data is crucial during both the long-CoT SFT and RL phases. Current approaches to video CoT construction also exhibit limitations. Some methods, such as Video-R1 [11], rely on non-thinking MLLMs (e.g., Qwen2.5-VL-72B) for CoT annotation. However, forcing non-thinking models to generate reasoning chains often compromises accuracy, as prior work shows that such practices lead to significant performance degradation in video understanding [11, 22, 44]. Other methods, like VideoRFT [39], use linguistic thinking models (e.g., DeepSeek-R1) for annotation in constructing textual CoTs. Using the linguistic thinking model, which depend solely on MLLM outputs or clip captions, fail to adequately capture the spatiotemporal information essential for video reasoning. To address these issues, we propose to utilize a state-of-the-art multimodal thinking model, Seed1.8 [6], which dynamically adjusts its reasoning chain based on video content and queries. Unlike prior methods that generate

uniformly long CoTs regardless of task complexity, this dynamic approach enables more efficient and context-aware reasoning.

Regarding the dynamics of video data used in the RL stage, the key factor lies in data difficulty. Although prior works [11,39] collect large amounts of video, they ultimately use only a small subset for training. This is because continuously introducing more data tends to cause instability during reinforcement learning. To mitigate this degradation, some studies [16,52] adjust the reward design based on the difficulty of training samples. Inspired by [3], we adopt a different strategy: for each query, we perform multiple pre-sampling passes to generate several reasoning trajectories and compute their accuracy, filtering out samples where the baseline model already achieves extremely high accuracy while retaining those that are more beneficial for reinforcement learning exploration.

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 second key insight is that incorporating reward evaluation on the reasoning process itself is crucial for video RL training. Recent works [3,11,12] only use verifiable multiple-choice questions or compute result similarity for open-ended questions to obtain outcome rewards. A recent approach [39] 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 [14] 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 highly accurate video reasoning models.

## 2 Related Work

### 2.1 Multimodal Large Language Models

MLLMs [2, 3, 5, 6, 9, 17, 68] have achieved significant advancements in multimodal tasks, such as image and video understanding. For image understanding, the early method [21] 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 [20, 24, 25]. 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 [25] to a zero-shot video understanding model, or further fine-tune the model on video data to achieve better video understanding capabilities [63]. Also, there are some works [34, 46, 59] 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 [26, 29].

To enhance temporal awareness in video understanding, some works propose explicit temporal textual prompts [34], temporal modules [58], and MRoPE techniques [4]. As for video training, many work adopt a hybrid data training strategy. For example, LLaVA-OneVision [19], InternVL2.5 [8], and Qwen2.5-VL [4] 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 [3, 4, 15, 68], such as continue training, strong-to-weak distillation [1], preference learning [33], and reinforcement learning [14]. In parallel, video benchmarks have been introduced to assess the various MLLMs, such as general video understanding tasks [13, 40, 50, 67] and video reasoning tasks [35, 55, 65]. More recently, the research frontier has shifted toward reinforcement learning-based reasoning enhancement [11, 12, 39, 44], tool-integrated video agents [60, 69], and evolving into video agents [57, 62] 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 [48], test-time scaling [18], RL [36], and agentic RL, etc. Inspired by DeepSeek-R1 [14, 36], 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 [27, 32, 41, 53, 56] 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 [10], CoT training [45], and RL training [44]. For video reasoning models’ RL training, several work [11, 39, 44] 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 [15], for example, integrates video data into pretraining and designs a post-training phase that combines SFT with RL. Keye-VL-1.5 [54] uses a slow-fast video encoder with iterative SFT and RL training, while InternVL3.5 [42] proposes a cascade RL framework that consists of a mixed preference optimization and an online RL stage. Qwen3-VL [3] 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 [12], 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 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:

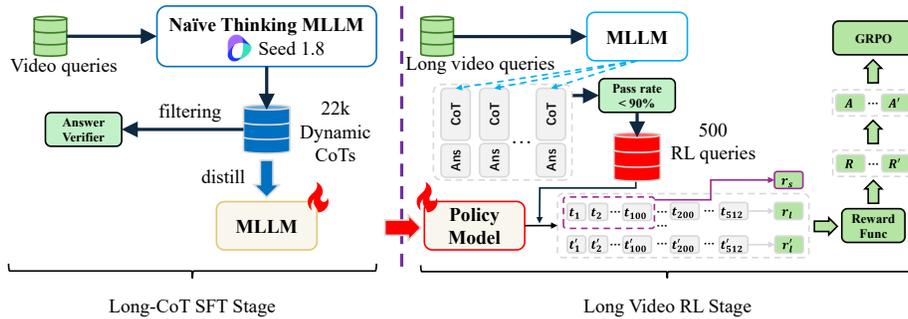
$$[T_{think}, T_{ans}] \sim P(T | V, Q), \quad (1)$$

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:

$$T_{ans} \sim P(\cdot | V, Q, T_{think}). \quad (2)$$

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 [3]), due to the complexity of the post-training process involved, it is difficult to further enhance



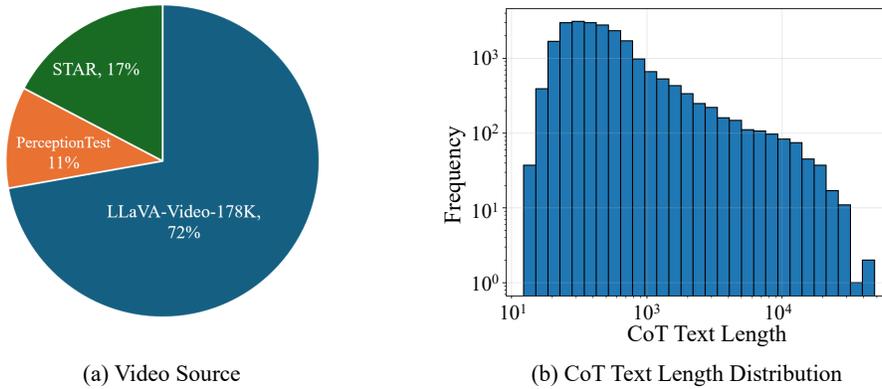
**Fig. 2:** The proposed framework to train a video reasoning model.

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. These methods do not take into account the dynamics of video data and tend to execute very long CoT 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 [6] 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 [31], STAR [49], LLaVA-Video-178k [64], 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



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

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  $P(T | V, Q)$  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 address this issue, we design a novel and effective way to measure the reward of reasoning CoT. When video model reason the  $t_{th}$  token  $T_{\text{think}}^t$ , to evaluate the importance of the current reasoning CoT tokens, we regard output tokens  $[T_{\text{think}}^0, \dots, T_{\text{think}}^t]$  as the context, and we re-input the video  $V$  and query  $Q$  along with context to model  $M$  to obtain a intermediate answer use the following template:

```

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} <STOP>
Assistant:
 $[T_{think}^0, \dots, T_{think}^t]$ 
The most correct option is:

```

**Table 1:** The input template used to prompt video reasoning models for intermediate answer.

The intermediate answer is the next predicted token  $\hat{T}_{ans}^0$ . The intermediate reward is compared with the ground truth answer  $A$ :

$$r = \begin{cases} 1, & \text{if } \hat{T}_{ans}^0 = A \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

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 final reward  $R$  to the trajectory. We define the final reward as follows:

$$R = 2(r_l - r_s) + 2(r_l - 0.5)(r_l - r_s + 2) \quad (4)$$

This final reward definition incorporates prior knowledge about the reasoning process: as reasoning unfolds, some steps contribute positively to arriving at the correct answer, while others detract from it. Based on this idea, we define the final reward values as shown in Table 2. Specifically, when the intermediate reward transitions from 0 to 1 as the reasoning progresses, we interpret this as the model reflecting on and correcting its reasoning—and thus assign a higher reward to such trajectories. Conversely, a transition from 1 to 0 indicates errors emerging in the reasoning content, which warrants a negative reward.

**Table 2:** Reward values for different combinations of  $r_s$  and  $r_l$ .

$(r_s, r_l)$	(0,0)	(0,1)	(1,0)	(1,1)
$R$	-2.0	5.0	-3.0	2.0

In this way, we first sampling multi trajectories  $\{T_1, T_2, \dots, T_G\}$  from a single video query, and we use the final reward function to score the trajectories, yielding  $G$  rewards  $\{R_1, R_2, \dots, R_G\}$ . Inspired by GRPO, we use the group normalized

reward as the advantages for all tokens, i.e.,  $\hat{A}_{i,t} = \frac{(R_i - \text{mean}(R))}{\text{std}(R)}$ . 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 [7] 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.

## 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 [66] 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 [13], LongVideoBench (LongVB) [50], MLVU [67], LVBench [40], VideoEval-Pro(VEP) [28], and on 4 video reasoning benchmarks including VSI-Bench [55], MMVU [65], LongVideoReason-eval [7] and Tomato [35]. More details can be found in the Appendix.

### 4.2 Main Results

*Video Generation Understanding.* We provide the quantitative comparison with state-of-the-art models in Table 3. 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 [3] is an open-source multimodal thinking model, but it needs denser frames and more reasoning tokens to achieve comparable performance. OneThinker [12]

**Table 3:** Results on video understanding benchmarks compared with state-of-the-art models. **Bold** indicates the best performance in the Thinking Models.

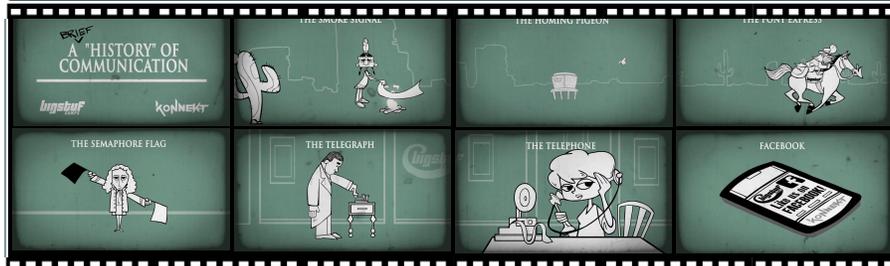
Model \ Benchmark	LVBench	LongVB	MLVU	VEP	Video-MME			
					val	M-Avg	mcq	Short
<b>Non-Thinking Models</b>								
Qwen2-VL-Base [38]	-	53.7	-	-	70.3	59.5	49.6	59.8
Qwen2-VL-Ins [38]	40.9	57.1	65.1	44.7	72.7	61.4	51.7	61.9
Qwen2.5-VL-Ins [4]	43.3	59.3	67.7	44.4	76.4	65.2	54.0	63.2
Qwen3-VL-Ins [3]	46.1	62.9	71.1	51.6	80.4	69.0	59.1	69.5
<b>Thinking Models</b>								
Video-R1-7B [11]	38.6	49.7	63.0	42.2	72.2	59.4	47.0	59.5
VideoChat-R1-7B [22]	-	-	-	-	-	-	-	60.0
LongVILA-R1-7B [7]	-	58.0	-	-	71.9	62.0	53.3	62.4
Open-o3 Video [30]	-	-	-	-	-	-	54.9	63.6
VideoRFT-7B [39]	41.4	56.5	65.6	44.8	76.0	64.2	52.7	64.3
Qwen3-VL-Thinking [3]	40.4	55.1	63.6	48.1	77.2	65.3	53.9	65.5
OneThinker-8B [12]	-	61.7	-	-	-	-	-	66.5
AdaTooler-V-7B [37]	-	-	-	-	-	-	-	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 [14]	45.1	60.8	69.0	50.0	79.0	68.4	54.7	67.4
Ours	<b>47.3</b>	<b>63.1</b>	<b>70.3</b>	<b>52.5</b>	<b>81.3</b>	<b>71.2</b>	<b>57.2</b>	<b>69.9</b>

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 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.

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

Model \ Benchmark	VSI-Bench			
	MMVU	Long	Video-Reason	Tomato
	mc	eval		
<b>Non-Thinking Models</b>				
LLaVA-OV-7B [19]	32.4	49.2	-	25.5
ViLA-1.5-8B [23]	28.9	49.2	-	-
VideoTree [47]	-	54.2	-	-
LLaVA-Video-7B [64]	36.2	60.2	-	24.9
Qwen2-VL-7B-Ins [38]	33.4	63.5	-	30.1
Qwen2.5-VL-7B-Ins [4]	38.1	68.8	-	25.6
Qwen3-VL-8B-Ins [3]	59.4	70.0	72.2	32.9
<b>Thinking Models</b>				
Video-R1-7B [11]	37.8	64.3	68.1	26.9
VideoRFT [39]	36.8	68.5	75.3	24.9
Qwen3-VL-8B-Thinking [3]	<b>56.6</b>	<b>78.1</b>	71.5	31.2
Qwen3-VL-Ins (Force-think, base)	52.0	68.6	73.4	33.0
GRPO [14]	54.1	70.4	75.8	33.0
Ours	54.4	71.0	<b>76.1</b>	<b>33.6</b>



**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 "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.".\n6. So, the order of communication methods in the video is: Smoke Signal -> Homing Pigeon -> Pony Express -> Semaphore Flag -> Telegraph.\n7. Now, 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.\n8. Therefore, the correct answer should be the Pony Express.\n9. Let'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.\n10. The most accurate and immediate predecessor to Semaphore is the Pony Express.\n</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.\n1. Given this sequence, the method listed immediately after "The Homing Pigeon" is "The Pony Express." Thus, the correct answer is D. Pony Express.\n2. To 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.\n</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.\n</think>\n<answer>D</answer>

**Fig. 4:** Comparison of reasoning results on short video questions.

**Table 5:** Ablation study on proposed components. The baseline without any training uses Qwen3-VL-8B-Instruct to force thinking. CoT-SFT denotes the long-CoT SFT stage. We also conducted ablation experiments on three components in RL: data filtering (filter), using the default outcome reward (outcome), and the combination of short-term and long-term rewards (short-long).

CoT-SFT	RL			LVBench	LongVB	MLVU	VEP	Video-MME
	filter	outcome	short-long	val	M-Avg	mcq	overall	
				42.2	58.2	66.5	47.4	66.6
		✓		44.6	58.1	67.4	49.1	67.1
✓				44.6	59.9	67.0	50.0	67.1
✓	✓	✓		45.1	60.8	69.0	50.0	67.4
✓	✓	✓	✓	<b>47.3</b>	<b>63.1</b>	<b>70.3</b>	<b>52.5</b>	<b>69.9</b>

**Table 6:** Ablation study on different short-term and long-term rewards.

Short term	Long term	LVBench	LongVB	MLVU	VEP	Video-MME
		val	M-Avg	mcq	overall	
400	512	45.2	60.4	69.1	51.0	69.3
300	512	45.6	60.9	69.0	51.8	69.7
200	512	46.0	61.5	69.7	50.5	68.8
100	512	<b>47.3</b>	<b>63.1</b>	<b>70.3</b>	<b>52.5</b>	<b>69.9</b>

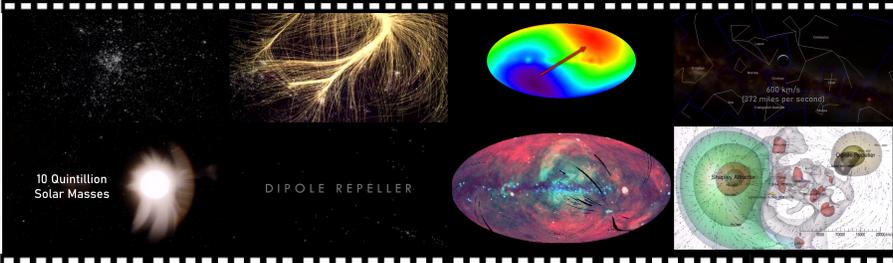
*Qualitative comparison.* Figure 4 shows the results of video reasoning on short videos. In this scenario, our model does not output lengthy reasoning processes like other models, but instead performs concise and accurate reasoning. Figure 5 shows the reasoning results on long videos. For more complex videos and questions, 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 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

**Table 7:** Generalization results on Qwen2.5-VL-7B-Instruct model. **Bold** indicates the best performance in the Thinking Models.

Model \ Benchmark	LVBench	LongVB	MLVU	VEP	Video-MME			
					val	M-Avg	mcq	Short
<b>Non-Thinking Models</b>								
Qwen2.5-VL-Ins [4]	43.3	59.3	67.7	44.4	76.4	65.2	54.0	63.2
<b>Thinking Models</b>								
VideoR1-7B [11]	38.6	49.7	63.0	42.2	72.2	59.4	47.0	59.5
VideoRFT-7B [39]	41.4	56.5	<b>65.6</b>	<b>44.8</b>	<b>76.0</b>	<b>64.2</b>	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 [14]	40.9	49.2	61.5	41.9	73.6	63.5	51.2	62.8
Ours	<b>42.5</b>	<b>57.2</b>	64.8	44.7	75.8	64.1	<b>54.1</b>	<b>64.7</b>



**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.</think>A. 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.</think>B. 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.</think>C. 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.</think>D. 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.</think>After 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.</think>Therefore, the aspect NOT described in the video is the challenges posed by the Zone of Avoidance.</think><answer>C</answer>

**VideoRFT:** <think>To determine which aspect is not described in the video, let's review the content step-by-step.</think>The 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.</think>Next, 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.</think>Now, let's evaluate each option:</think>Option 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.</think>Option 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.</think>Option 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.</think>Option 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.</think>Based on this analysis, Option A is the aspect that is not described in the video.</think><answer>A</answer>

**Ours:** <think>Got it, let's analyze each option.</think>First, 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.</think>Option B: The video shows the Great Attractor's influence on Laniakea (the supercluster the Local Group is in), so that's covered.</think>Option 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.</think>Option D: The video covers observational history (like the CIZA project, ROSAT, Anglo-Australian Telescope, etc.), so that's covered.</think>So the one not described is A.</think><answer>A</answer>

**Fig. 5:** Comparison of reasoning results on long video questions.

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 Figure 6, we evaluate the short-term reward at positions 100, 200, 300, and 400, while keeping the position at 512 fixed as it corresponds to the outcome reward—which we consider as the long-term reward. Experimental 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 an early short-term reward and the final outcome reward is most effective. When the token positions for calculating the two rewards are closer together, the benefits brought by the rewards gradually diminish. On short and medium video benchmarks, as well as on challenging benchmarks like VideoEvalPro, specific reward combinations also yield strong reasoning accuracy. For instance, the combination of positions 300 and 512 achieves competitive performance on Video-MME and VideoEvalPro, approaching the results of our default configuration (100 and 512).

#### 4.4 Generalization Experiments.

In Table 7, we validate the generalization ability of our method on a widely used foundation model, Qwen2.5-VL-7B-Instruct. First, comparing our method with GRPO on the baseline model shows that our approach brings significant and consistent performance improvements across all benchmarks. All thinking models in the table are based on the same base model, ensuring a fair comparison. Our model achieves the best performance on LVBench, LongVideoBench, and Video-MME. On MLVU, our model performs slightly below VideoRFT, while on VideoEvalPro and the short and medium subtasks of Video-MME, the results are closely matched. Notably, our method uses significantly fewer training samples than VideoRFT.

## 5 Conclusion

In this work, we identify that existing multimodal reasoning models inherit a language-centric bias, leading to inefficient and ineffective reasoning in long video understanding tasks. To address this, we propose a simple yet effective post-training framework that transforms Instruction multimodal models into efficient video reasoning models. Our framework consists of two key components: Long-CoT SFT to equip the model with dynamic video reasoning capabilities, followed by RL with a hybrid short-term and long-term reward mechanism to further optimize reasoning accuracy and efficiency. Extensive experiments on diverse video benchmarks show that our framework efficiently trains the Qwen3-VL-Instruct-8B model into a high-performing video reasoning model, significantly outperforming the Qwen3-VL-Thinking-8B. This highlights the effectiveness of our framework in building accurate and efficient video reasoning models.

## References

1. Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Garea, S.R., Geist, M., Bachem, O.: On-policy distillation of language models: Learning from self-generated mistakes. In: The twelfth international conference on learning representations (2024)
2. Anthropic: Introducing the next generation of claude, 2024 (2024), <https://www.anthropic.com/news/claude-3-5-sonnet>
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. ByteDance: Introduction to techniques used in Seed1.6 (2025), [https://seed.bytedance.com/en/seed1\\_6](https://seed.bytedance.com/en/seed1_6)
6. ByteDance: Introduction to techniques used in Seed1.8 (2025), [https://seed.bytedance.com/zh/seed1\\_8](https://seed.bytedance.com/zh/seed1_8)
7. Chen, Y., Huang, W., Shi, B., Hu, Q., Ye, H., Zhu, L., Liu, Z., Molchanov, P., Kautz, J., Qi, X., et al.: Scaling rl to long videos. arXiv preprint arXiv:2507.07966 (2025)
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M.L., Hsu, W.: Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230 (2024)
11. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
12. Feng, K., Zhang, M., Li, H., Fan, K., Chen, S., Jiang, Y., Zheng, D., Sun, P., Zhang, Y., Sun, H., et al.: Onethinker: All-in-one reasoning model for image and video. arXiv preprint arXiv:2512.03043 (2025)
13. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 (2024)
14. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
15. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)

16. Guo, Z., Hong, M., Jin, T.: Observe-rl: Unlocking reasoning abilities of mllms with dynamic progressive reinforcement learning. arXiv preprint arXiv:2505.12432 (2025)
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
18. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
19. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
20. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
21. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900 (2022)
22. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-rl: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
23. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: CVPR. pp. 26689–26699 (2024)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR. pp. 26296–26306 (2024)
25. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), <https://llava-v1.github.io/blog/2024-01-30-llava-next/>
26. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4122–4134 (2025)
27. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)
28. Ma, W., Ren, W., Jia, Y., Li, Z., Nie, P., Zhang, G., Chen, W.: Videoeval-pro: Robust and realistic long video understanding evaluation. arXiv preprint arXiv:2505.14640 (2025)
29. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
30. Meng, J., Li, X., Wang, H., Tan, Y., Zhang, T., Kong, L., Tong, Y., Wang, A., Teng, Z., Wang, Y., et al.: Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. arXiv preprint arXiv:2510.20579 (2025)
31. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems* **36**, 42748–42761 (2023)
32. Peng, Y., Zhang, G., Zhang, M., You, Z., Liu, J., Zhu, Q., Yang, K., Xu, X., Geng, X., Yang, X.: Lmm-rl: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536 (2025)

33. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. *NIPS* **36** (2024)
34. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multi-modal large language model for long video understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14313–14323 (2024)
35. Shangguan, Z., Li, C., Ding, Y., Zheng, Y., Zhao, Y., Fitzgerald, T., Cohan, A.: Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models (2024), <https://arxiv.org/abs/2410.23266>
36. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024)
37. Wang, C., Feng, K., Chen, D., Wang, Z., Li, Z., Gao, S., Meng, M., Zhou, X., Zhang, M., Shang, Y., et al.: Adatooler-v: Adaptive tool-use for images and videos. *arXiv preprint arXiv:2512.16918* (2025)
38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024)
39. Wang, Q., Yu, Y., Yuan, Y., Mao, R., Zhou, T.: Videorf: Incentivizing video reasoning capability in mllms via reinforced fine-tuning (2025)
40. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., et al.: Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035* (2024)
41. Wang, W., Gao, Z., Chen, L., Chen, Z., Zhu, J., Zhao, X., Liu, Y., Cao, Y., Ye, S., Zhu, X., et al.: Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291* (2025)
42. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265* (2025)
43. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265* (2025)
44. Wang, X., Peng, P.: Open-r1-video. <https://github.com/Wang-Xiaodong1899/Open-R1-Video> (2025)
45. Wang, Y., Zeng, Y., Zheng, J., Xing, X., Xu, J., Xu, X.: Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355* (2024)
46. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: *European Conference on Computer Vision*. pp. 396–416. Springer (2024)
47. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 3272–3283 (2025)
48. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)

49. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: STAR: A benchmark for situated reasoning in real-world videos. In: Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) (2021)
50. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding (2024), <https://arxiv.org/abs/2407.15754>
51. Xia, J., Zang, Y., Gao, P., Li, Y., Zhou, K.: Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. CoRR **abs/2505.14677** (2025). <https://doi.org/10.48550/ARXIV.2505.14677>, <https://doi.org/10.48550/arXiv.2505.14677>
52. Xiao, W., Gan, L., Dai, W., He, W., Huang, Z., Li, H., Shu, F., Yu, Z., Zhang, P., Jiang, H., et al.: Fast-slow thinking for large vision-language model reasoning. arXiv e-prints pp. arXiv-2504 (2025)
53. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440 (2024)
54. Yang, B., Wen, B., Ding, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., et al.: Kwai keye-vl 1.5 technical report. arXiv preprint arXiv:2509.01563 (2025)
55. Yang, J., Yang, S., Gupta, A., Han, R., Fei-Fei, L., Xie, S.: Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. arXiv preprint arXiv:2412.14171 (2024)
56. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., Zhang, B., Chen, W.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)
57. Yeo, W., Kim, K., Yoon, J., Hwang, S.J.: Worldmm: Dynamic multimodal memory agent for long video reasoning. arXiv preprint arXiv:2512.02425 (2025)
58. Zeng, X., Li, K., Wang, C., Li, X., Jiang, T., Yan, Z., Li, S., Shi, Y., Yue, Z., Wang, Y., Wang, Y., Qiao, Y., Wang, L.: Timesuite: Improving MLLMs for long video understanding via grounded tuning. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=nAVejJURqZ>
59. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
60. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. arXiv preprint arXiv:2508.04416 (2025)
61. Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937 (2025)
62. Zhang, X., Jia, Z., Guo, Z., Li, J., Li, B., Li, H., Lu, Y.: Deep video discovery: Agentic search with tool use for long-form video understanding. arXiv preprint arXiv:2505.18079 (2025)
63. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>
64. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
65. Zhao, Y., Xie, L., Zhang, H., Gan, G., Long, Y., Hu, Z., Hu, T., Chen, W., Li, C., Song, J., Xu, Z., Wang, C., Pan, W., Shangguan, Z., Tang, X., Liang, Z., Liu, Y., Zhao, C., Cohan, A.: Mmvu: Measuring expert-level multi-discipline video understanding (2025), <https://arxiv.org/abs/2501.12380>

66. Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., Zhou, W., Chen, Y.: Swift:a scalable lightweight infrastructure for fine-tuning (2024), <https://arxiv.org/abs/2408.05517>
67. Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv preprint arXiv:2406.04264 (2024)
68. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
69. Zou, J., Huang, Z., Zhang, S., Zhang, L., Shen, W.: Videobrain: Learning adaptive frame sampling for long video understanding. arXiv preprint arXiv:2602.04094 (2026)