Incentivizing Versatile Video Reasoning in MLLMs via Data-Efficient Reinforcement Learning

Xiaodong Wang[1,2], Zhirong Wu[1], Langling Huang[1], Yuxi Zheng[1], Peixi Peng[1,2*]
[1] Peking University
[2] Pengcheng Laboratory
wangxiaodong21s@stu.pku.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have made great progress in video understanding tasks. However, when it comes to understanding complex or lengthy videos, MLLMs tend to overlook details or produce hallucinations. To alleviate these issues, recent work has attempted to leverage reinforcement learning (RL) to boost models' deep linguistic reasoning of complex videos. But these methods have two main problems: First, the RL framework they used has unstable training, high training costs, and is difficult to train satisfactory video reasoning models; Second, the linguistic reasoning process is difficult to guarantee the reliability of visual information. To alleviate these problems, we propose to use multimodal elements for reasoning, and we design a novel framework to build and enhance versatile video reasoning capabilities on MLLMs. We carefully design a multi-task cold start and multi-task reinforcement learning to improve the model's visual perception and proficiency in multiple capabilities. In the inference phase, we leverage multimodal reasoning and dynamic sampling to further improve the performance. We verified the efficiency of the framework on a base MLLM (Qwen2-VL-7B-Base). Through cold-start with 3k data and reinforcement learning training with 5k data, combined with inference design, our final model significantly outperforms the base model on seven public video benchmarks, even surpassing and approaching the state-of-the-art Instruct Models such as Qwen2.5-VL-7B-Instruct trained with large-scale data.

1. Introduction

The recent surge of large language reasoning models [16, 19, 31, 49] has marked great progress towards a new era of artificial intelligence, particularly in addressing challenging and realistic tasks such as mathematics, reasoning, etc. These advances have also promoted the rapid development of multimodal large language models (MLLMs) [15, 17, 28]. A notable trend is the extension of reinforcement learning (RL) methods from the linguistic to the multimodal domain. Recent studies focus on improving RL algorithms [27], designing more effective verifiers [36, 39], and expanding to diverse modalities [38, 59].

Despite the promising performance of MLLMs trained with RL algorithms on image understanding tasks [5, 55], RL training on video tasks has not achieved comparable improvements [12, 34], and establishing a stable and efficient RL framework for video understanding remains an open challenge. Existing RL training frameworks [7, 12] for videos typically involve high-cost long Chain-of-Thought (CoT) reasoning annotations, large-scale CoT cold start, and large-scale RL training. For example, Video-R1 [12] curates 165k data for cold start and 260k data for RL training. These models build on Instruct Models (e.g., Qwen2.5-VL-Instruct), which have already undergone large-scale SFT and post-training on image and video tasks using a direct-response paradigm, i.e., generating a brief answer immediately in response to a question. However, the inherent prior of such models favors direct responses over step-by-step reasoning, limiting their performance on tasks requiring logical reasoning [57]. Therefore, it requires large-scale data and extensive training to enable the model to develop robust reasoning capabilities and adapt to reasoning formats and tasks. Although recent methods [12, 34] have enhanced the model's ability to reason over video content and generate summaries, they still lag behind the original Instruct Models, revealing a performance gap between reasoning training and direct-response models.

Furthermore, recent studies [12, 15] only use RL training to incentivize the linguistic reasoning path, which may include reasoning contents such as problem analysis, video perception, information reasoning, and summary. However, using only a textual reasoning path makes it difficult to ensure the long-term accuracy of visual information, and long-term text reasoning is prone to lead to incorrect and hallucination [20]. For video tasks, more effective and accurate reasoning paradigms involving multimodal elements should be explored instead of relying merely on text elements, in order to better cope with reasoning-intensive video understanding tasks. Meanwhile, although giving the model a longer budget during reasoning (e.g., more than 1k text tokens) can stimulate a self-reflective mechanism during the reasoning process, it also leads to a longer inference time, which will become a crucial bottleneck in real-world video applications.

To address these limitations, we propose **VideoReasoner**, an efficient framework that builds and enhances versatile video reasoning capabilities for base MLLMs. To construct a simple and usable framework and verify the effectiveness of this framework. We set the task to conduct training based on Base MLLMs. The basic goal is to use this framework to enhance video understanding capabilities more efficiently; that is, the video understanding performance needs to exceed the corresponding Instruct Model, and explore whether this framework can bring about effects beyond expectations. As mentioned earlier, there is an inherent performance gap when using the Instruct Model. The usage of the Base Model is proposed here to avoid the performance gap and to verify the effectiveness of the basic framework. Furthermore, in the experimental section, we also presented the training results based on some Instruct Models. As for how to bridge this gap, more subsequent work is needed for exploration. Meanwhile, since this framework subsequently designs the reasoning process of multimodal elements, which involves multitask learning, the Base Model is more suitable as a baseline because it has only undergone multimodal pre-training and can adapt to multitask learning more efficiently.

To avoid error accumulation and hallucinations of visual content understanding caused by only using textual reasoning, we extend it to three perspectives: event reasoning, keyframe reasoning, and direct-response. As important contents in videos, events and keyframes can express clearer information than text. To this end, we design a two-stage training method to enable the model to learn and enhance its reasoning ability from these perspectives. In the first stage, a multi-task SFT is designed as a cold start. We design a unified instruction for multiple tasks as shown in Figure 1. These tasks have different task prefixes and subsequent specific contents. An example is shown in the left part of Figure 1. For keyframe reasoning, we do not directly output the indices of keyframes for reasoning. Instead, we adopt a simple approach that predicts the key elements. In the second stage, we propose a novel multi-task reinforcement learning method. For a given video-question pair, multiple sets of responses are generated using prompts with different task prefixes, and task-specific rewards are defined for Group Relative Policy Optimization (GRPO). After training, we design an inference pipeline leveraging the model's three video reasoning capabilities. Specifically, the model performs event reasoning and keyframe reasoning in parallel, conducts dense sampling of the outputs, and then feeds the sampled video frames back into the model to generate a direct response.

Through extensive experiments on various public video benchmarks, including general video understanding, video reasoning, and video temporal grounding benchmarks, we validate the effectiveness of the proposed VideoReasoner framework. Based on Qwen2-VL-7B-Base, through a cold start with 3k samples and RL training with 5k samples, combined with the proposed inference pipeline, the framework achieves substantial improvements across seven benchmarks, outperforming Qwen2-VL-7B-Instruct on five benchmarks and Qwen2.5-VL-7B-Instruct on three benchmarks. Compared with the data and training costs required for SFT or post-training used in training two Instruct Models, our framework requires only 8k data while achieving comparable results, demonstrating its efficiency and highlighting its potential for real-world applications.

2. Related Work

2.1. Multimodal Large Language Models for Videos

MLLMs [1, 2, 4, 10, 18, 62] have achieved significant advancements in video understanding tasks, and open-source models are gradually catching up with closed-source models in terms of multimodal capabilities. MLLMs treat video input as a sequence of images and bridge the visual tokens and language space through a modality alignment module, and these works use Q-Former [22] to aggregate temporal information or simple MLP projectors. The training paradigms of MLLMs for video understanding continue to evolve. Recently, Qwen2.5-VL [2] fuses adjacent frames and further compresses encoded multiple visual tokens into a single token, which is then connected to the language model via MLP. To enhance temporal awareness, some works propose explicit temporal textual prompts [30], temporal module [53], and MRoPE techniques [2]. As for video training, many works adopt a hybrid data training strategy. For example, InternVL2.5 [8] and Qwen2.5-VL 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 [2, 15, 62]. In parallel, video benchmarks have been introduced to assess the various MLLMs, such as general video understanding tasks [13, 35, 45] and video reasoning tasks [50, 60]. Recently, the use of reinforcement learning to enhance the reasoning ability of models [12, 38], the ability to use tools [54], and evolve into video agents [**?**] are the cutting-edge directions for the development of MLLMs towards more powerful and practical video understanding.

2.2. Multimodal Large Language Models for Reasoning

Large language reasoning models using CoT [44], test-time scaling [19], and RL [31] have achieved great success for the reasoning and instruction-following abilities, such as in mathematics, coding, and agentic tasks. Recently, DeepSeek-R1 [31] demonstrates that large-scale RL with verifiable rewards induces emerging reasoning capabilities in LLMs. Inspired by this, the introduction of reasoning and design reasoning on MLLMs is continuously evolving and has demonstrated some promising results [25, 29, 36, 47, 51]. To explore the multimodal reasoning effect for complex video understanding tasks, some works focus on step-by-step reasoning, CoT training, and RL training. VideoCoT [41] propose a high-quality video dataset with chain-of-thought reasoning annotations. Video-of-Thought [11] breaks down a complex video task into simpler sub-problems, such as tracking or action analysis, and it addresses them using step-by-step reasoning from a low-level pixel perception to high-level cognitive interpretation. For Video-MLLM RL training, [12, 38] introduce GRPO training for MLLM to reasoning for fully understanding the video-language relationship before the final answer. Recent RL training frameworks often involve offline and online training stages. Seed1.5-VL [15] incorporates video data into the pretraining phase and designs a post-training phase through a combination of supervised fine-tuning (SFT) and RL techniques. Keye-VL-1.5 leverages a slow-fast video encoding method, and the post-training stage is continuous iterative SFT and RL training. InternVL3.5 [37] also propose a cascade RL framework that consists of a mixed preference optimization and an online RL stage. However, these RL frameworks merely rely on linguistic reasoning to analyze and interpret video content. While our proposed framework involves not only linguistic reasoning but also multimodal element reasoning.

3. Method

Overview In this section, we propose a two-stage training framework to build and enhance versatile video capabilities for base MLLMs. The main idea is to fully explore the various video capabilities to enhance video understanding, with the expectation that the model can effectively leverage its pre-trained knowledge to perform these tasks. The framework consists of three steps: (1) a multi-task cold start (Section 3.2); (2) a multi-task RL (Section 3.3), and (3) an efficient video inference pipeline (Section 3.4). See Figure 1 for the overview.

3.1. Background of GRPO

Group Relative Policy Optimization (GRPO) [31] is proposed to save the training costs of reinforcement learning. It foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q, GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{G}\sim \pi_{\theta_{old}}(O|q)]} \frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta \| \pi_{ref}\right)\right), \tag{1}$$

$$\mathbb{D}_{KL}\left(\pi_\theta || \pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{2}$$

where ϵ and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \dots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}. \tag{3}$$

3.2. Multi-task Supervised Fine-tuning as a Cold Start

A base MLLM is built upon a large language model and trained with multimodal pretraining to align visual signals with language tokens, such as Qwen2-VL [33], which processes 1.4 trillion tokens during pretraining. Considering that the base model is exposed to large-scale and diverse linguistic and visual scenarios during the pre-training stage, it is expected that the model can utilize this prior knowledge for various visual tasks, including complex video understanding. To this end, we innovatively design three core tasks for the MLLM to learn: video question answering, video event grounding, and keyframe detection. However, “keyframes” are difficult to define, and current MLLMs still struggle to predict them accurately. To address this, we reformulate keyframe detection as key element generation and employ a visual encoder to retrieve the corresponding video frames.

To support the model’s adaptability to the three aforementioned tasks, we curate datasets specifically tailored for each task. For video question answering, which is the most commonly used task for training MLLMs, we adopt multi-choice QA data from the training sets of [12]. For video event grounding, we use the temporal grounding training set from [14]. For key element generation, we collect raw videos from [58] and employ a proprietary model [15] to generate key elements, constructing paired video and textual key element annotations. For each input video X_v and instruction (comprising a system prompt and a video query),

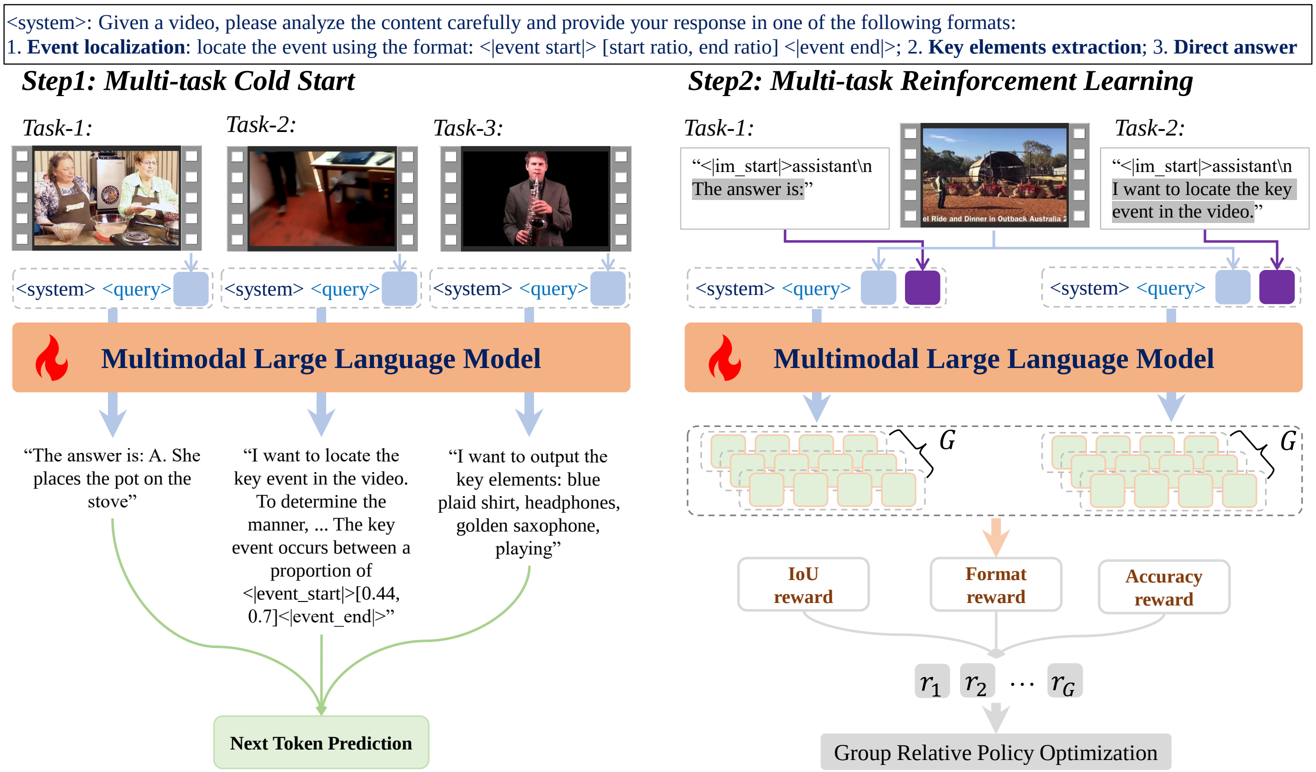
<system>: Given a video, please analyze the content carefully and provide your response in one of the following formats:
1. **Event localization**: locate the event using the format: <|event start|> [start ratio, end ratio] <|event end|>; 2. **Key elements extraction**; 3. **Direct answer**
Step1: Multi-task Cold Start
Step2: Multi-task Reinforcement Learning
Task-1:
Task-2:
Task-3:
<system> <query>
Multimodal Large Language Model
"The answer is: A. She places the pot on the stove"
"I want to locate the key event in the video. To determine the manner, ... The key event occurs between a proportion of <|event_start|>[0.44, 0.7]<|event_end|>"
"I want to output the key elements: blue plaid shirt, headphones, golden saxophone, playing"
Next Token Prediction
Task-1:
"<|im_start|>assistant\n The answer is:"
Task-2:
"<|im_start|>assistant\n I want to locate the key event in the video."
<system> <query>
G
IoU reward
Format reward
Accuracy reward
r_1 r_2 ⋯ r_G
Group Relative Policy Optimization

Figure 1. The proposed training paradigm aims to build and enhance versatile video reasoning capabilities for MLLMs, including a multi-task Cold Start and a multi-task RL with high efficiency.

the target answer corresponds to one of three types, as illustrated in the left part of Figure 1. The most distinctive feature of these three responses lies in their prefixes: "The answer is:", "I want to locate the key event in the video.", and "I want to output the key elements:". All tasks share the same system prompt, which specifies the principles the model should follow, as detailed in Figure 1.

For a sequence of length L, the probability of various target answers is defined as follows:

$$p(X_a|X_v, X_{instruct}) = \prod_{i=1}^{L} \pi_\theta(x_i|X_v, X_{instruct}, X_{a,<i}) \tag{4}$$

where θ is the trainable parameter of MLLM. All parameters of the visual encoder and the language model backbone are trainable. $X_{instruct}$ and $X_{a,<i}$ are instruction and answer tokens before the current prediction token x_i, and the whole response (including the prefix) is used to compute the loss for next token prediction.

For video event grounding, we observe that using absolute numerical predictions [53] makes the model's outputs highly dependent on the training distribution. For instance, if the model is trained on short videos, it tends to predict very small time values and has difficulty handling event localization in long videos. To address this issue, we propose a relative numerical prediction for event grounding, i.e., predicting the time ratio for durations `[start_ratio, end_ratio]`. We also insert two learnable special tokens `<|event_start|>` and `<|event_end|>` to make the model stably predict the grounding results when performing this task.

We employ only pre-trained MLLMs for multi-task fine-tuning instead of adopting post-trained SFT models. Although the latter typically achieve stronger performance, they also exhibit stronger preferences or inductive biases and demand larger-scale multi-task datasets for effective fine-tuning. In our setting, we utilize approximately 3k training samples in total, with each task accounting for around 1k samples.

3.3. Multi-task Reinforcement learning

The Cold Start enables the base MLLM to adapt to the responses of various tasks, but it is more about proficient format output rather than truly effective learning of these capabilities. To enhance these capabilities, we utilize the commonly used reinforcement learning algorithm GRPO [31] to incentivize the diverse capabilities of the model. Although recently there have been some works focusing on RL for

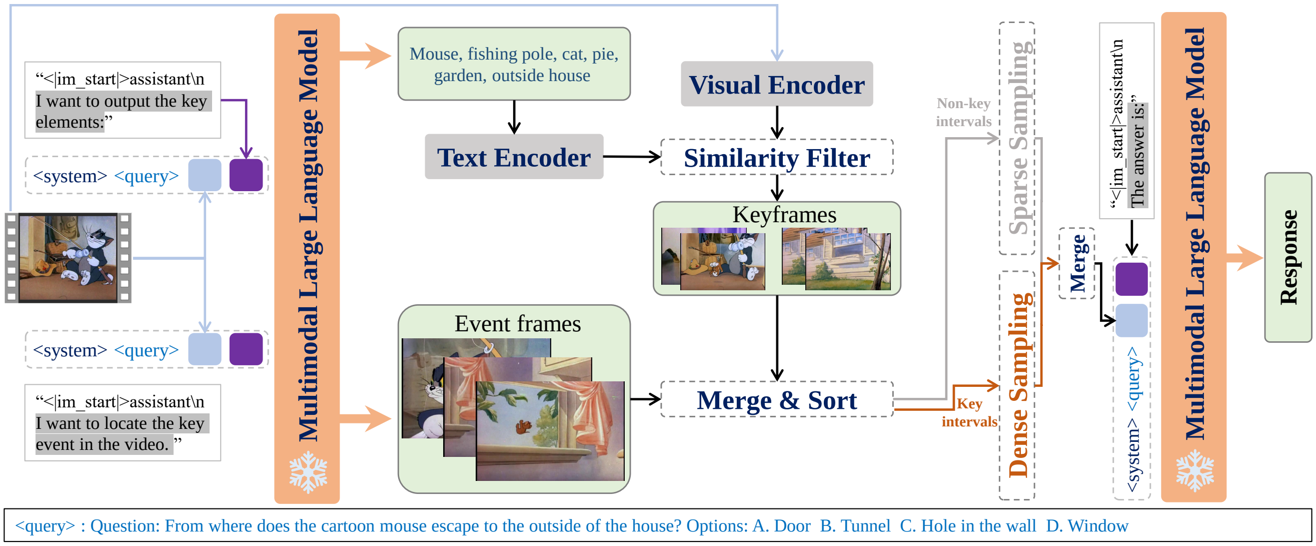
“<|im_start|>assistant\n I want to output the key elements:”
<system> <query>
<system> <query>
“<|im_start|>assistant\n I want to locate the key event in the video. ”
Multimodal Large Language Model
Mouse, fishing pole, cat, pie, garden, outside house
Text Encoder
Visual Encoder
Similarity Filter
Keyframes
Event frames
Merge & Sort
Non-key intervals
Key intervals
Sparse Sampling
Dense Sampling
Merge
“<|im_start|>assistant\n The answer is:”
<system> <query>
Multimodal Large Language Model
Response
<query> : Question: From where does the cartoon mouse escape to the outside of the house? Options: A. Door B. Tunnel C. Hole in the wall D. Window

Figure 2. Inference for video understanding using three capabilities of the enhanced model.

video understanding [12, 42], they only use GRPO for a single task, such as video QA or temporal grounding. In this work, we propose a novel multi-task GRPO algorithm. At the model level, the model can generate rollouts of multiple tasks and separate optimized policy models based on the relative advantages between groups obtained from the rollouts of different tasks. At the data level, given the same video-query pair, we use different prefix hints to prompt the model to roll out different tasks for the same query, effectively improving the utilization efficiency of data.

As shown in the right part of Figure 1, we select two tasks—event grounding and video QA—to construct the multi-task GRPO for video training. Given the same video and query, we prepend task-specific prefixes after the default assistant generation prompt (e.g., `<assistant>`). Specifically, the event grounding prefix is: “I want to locate the key event in the video.”, and the video QA prefix is: “The answer is:”. Importantly, we do not need to manually construct such data. The dataset of [46] provides both the answer metadata and reference time intervals for the same video-query pairs. Although [46] introduces a new video QA task, we instead repurpose this dataset to build the training set for multi-task GRPO.

Unlike the Cold Start stage, where the model is required to predict task prefixes, in GRPO, the tokens corresponding to task prefixes are not used to compute the loss. Given a video X_v, and a query q, generation prompt g, and a task prefix in $\{p_1, p_2\}$, a multimodal query is defined as $m = [X_v, q, g, p]$, where $[\cdot]$ denotes token concatenation. The objective of multi-task GRPO is defined as:

$$\mathcal{J}_{M\text{-}GRPO}(\theta) = \mathbb{E}_{[p\sim\{p_1,p_2\},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|m)} \frac{1}{G}\sum_{i=1}^{G} \Bigg[\min\left(\frac{\pi_\theta(o_i|m)}{\pi_{\theta_{old}}(o_i|m)}A_i, \quad \text{clip}\left(\frac{\pi_\theta(o_i|m)}{\pi_{\theta_{old}}(o_i|m)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\Bigg] \tag{5}$$

where KL divergency $\mathbb{D}_{KL}$ and advantage A_i are are the same as defined in Equation 2 and Equation 3, respectively.

Reward Modeling The definition of the reward r_i guides the model's learning objective. Our goal is to enable the model to find the direction for optimization within its own search space while maintaining proficiency in multiple tasks. At the same time, since we choose not to predict task prefixes, we prevent the model from getting stuck on which format to output or overfitting to a single format during rollout and optimization. Instead, it performs autoregressively to continue generating subsequent tokens for a given specific prefix. To better leverage the role of the two tasks in enhancing video comprehension ability, we have designed three rewards, including an IoU reward $r_{\texttt{IoU}}$, a format reward $r_{\texttt{form}}$, and an accuracy reward $r_{\texttt{acc}}$.

- **IoU Reward** $r_{\texttt{IoU}}$: this reward is designed for the event grounding task. It is computed as the IoU between the predicted interval ratio and the ground-truth interval ratio.
- **Format Reward** $r_{\texttt{form}}$: this reward is used for the event grounding task to evaluate whether the model correctly predicts the two special tokens,

- **Accuracy Reward** $r_{\mathtt{acc}}$: this reward is for the video QA task, and it takes a value of either 0 or 1.

For the collaborative optimization of the two tasks, GRPO only computes importance weights of the rollout tokens, targeting the probability of the model's response to the input sequence. Thus, although query data from different tasks are sampled in the same batch, the optimization objectives of each task do not interfere with each other. For the event grounding task, the objective is to optimize the model's response quality given the input video, query, and task prefix. In this case, the accuracy reward can be set to 0, allowing the model to focus on improving the format reward $r_{\mathtt{form}}$ and IoU reward $r_{\mathtt{IoU}}$. For the Video QA task, the goal is to optimize the model's response quality for this task input. We set the format reward and IoU reward to 0 and let the model improve the accuracy reward $r_{\mathtt{acc}}$. The overall reward function is defined as their sum:

$$r(o) = r_{\mathtt{IoU}} + r_{\mathtt{form}} + r_{\mathtt{acc}} \tag{6}$$

For the training data of this stage, we use only 5k video queries. The training at this stage is data-efficient and does not require long rollouts, resulting in high efficiency.

3.4. Inference with Versatile Video Reasoning Capabilities

After reinforcement learning, the event grounding and video understanding abilities are enhanced, and we try to combine the two abilities and keyframe detections to build an efficient video question-answering process. As shown in Figure 2, we first input the video along with two types of task prefixes to the MLLM, prompting the model to output the duration of the related event and key elements in parallel. For key elements, we use a text encoder [3] to extract the textual embeddings, and feed the uniformly sampled video frames into a visual encoder [3] to obtain video embeddings, from which keyframes with high similarity are selected.

Specifically, the event grounding process outputs the most important event interval ratio $[S_r, E_r]$, which is multiplied by the video duration to obtain the absolute time interval $[S_t, E_t]$. The keyframe detection process generates a series of keyframe positions. Based on the original segments division, we obtained a series of time segments $\{[S_{k_1}, E_{k_1}], [S_{k_2}, E_{k_2}], \cdots, [S_{k_n}, E_{k_n}]\}$. Then, we merge and sort all the selected time segments to obtain the key intervals. We adopt high fps dense sampling to fully utilize key visual information. Meanwhile, to prevent the model from ignoring global information, we also took into account other non-key time regions and implemented sparse sampling using a low fps. We merge and sort the sampled frames of these two parts, and input them together with the video QA prefix into the MLLM to obtain the final response.

4. Experiment

4.1. Setup

Qwen2-VL-7B-Base [33] is used as the base model, with all parameters of the MLLM fully fine-tuned. For analysis experiments, we also finetuned Qwen2-VL-7B-Instruct [33] and Qwen2.5-VL-7B-Instruct [2]. All experiments are conducted on H20 GPUs. We sample 64 frames per video for both training stages and inference. The 3k training data for the cold start stage are sampled from [12, 14, 58], and the 5k training data for RL training are sampled from [46]. Evaluation is performed on 7 video benchmarks: Video-MME [13], LongVB [45], MLVU [61], LVBench [35], VideoEval-Pro [26], VSI-Bench [50], and MMVU [60]. The temporal grounding benchmark uses Charades-STA [14]. More details refer to the Appendix.

4.2. Main Results

Quantitative Results Table 1 presents a comparison with previous models. Compared with current state-of-the-art models. When using Qwen2.5-VL-7B-Instruct as baseline, compared with Video-R1 [12], our multi-task RL training is more efficient (use much less training data) and effective (achieve more significant and stable improvements on all benchmarks. When using Base model as baseline, RL has achieved improvements on all benchmarks. By using our proposed reasoning method, it has further surpassed or achieved comparable results to the latest Qwen2.5-VL-7B-Instruct model. Compared with Qwen2-VL-Base, Qwen2.5-VL-Instruct adopts a better visual encoding design and uses large-scale data in the Instruct tuning stage. Our method only uses 8k data, which highlights the potential of our proposed method.

Table 2 presents a comparison on video reasoning benchmarks. Qwen2.5-VL-7B-Instruct with our RL achieves the best results on both benchmarks. With our RL training, Qwen2.5-VL-7B-Base improves by 4.8% on VSI-Bench and 1.3% on MMVU, surpassing Qwen2-VL-7B-Instruct on VSI-Bench. Compared with Video-R1, our multi-task RL approach shows consistent and stable performance improvements on video reasoning tasks.

Table 3 presents a comparison of temporal grounding on Charades-STA.Notably, although our baseline model, Qwen2-VL-7B, does not support temporal grounding, following cold-start training it outperforms Qwen2.5-VL-7B-Instruct. After RL training, our models achieve the best results in mIoU, R1@0.3, and R1@0.5, surpassing state-of-the-art models. These results demonstrate that our framework can develop and enhance the emerging temporal grounding capability of base MLLMs.

Qualitative Results Figure 3 shows that our method combines multiple inherent capabilities to perform video ques-

Table 1. Performance on public video benchmarks compared to previous models.

Model \Benchmark	Video-MME				LongVB	MLVU	LVBench	VideoEval-Pro
	Short	Medium	Long	Overall	Val	M-Avg	Overall	MCQ
LLaVA-Video-72B [58]	81.4	68.9	61.5	70.6	62.4	71.3	46.1	50.1
InterlVL2.5-72B [8]	82.8	70.9	62.6	72.1	-	-	37.9	-
PLLaVA-7B [48]	-	-	-	-	40.2	-	-	-
LongVA [56]	61.4	50.9	45.0	52.4	-	56.3	-	38.0
Long-LLaVA [32]	61.9	51.4	45.4	52.9	-	-	-	36.9
Kangaroo-8B [24]	66.1	55.3	46.6	56.0	54.2	-	-	-
VideoTree [43]	-	-	-	56.1	52.3	-	-	-
InternVL2-8B [9]	68.0	52.0	48.9	56.3	54.6	56.3	-	39.9
ViLA-1.5-8B [23]	-	-	-	58.2	56.3	56.7	-	-
Qwen2-VL-7B-Instruct [33]	70.7	57.6	50.2	59.3	55.2	61.7	39.7	39.6
LongVILA [6]	-	-	-	60.1	-	-	-	-
MiniCPM-V-2.6 [52]	71.3	59.4	51.8	60.9	54.9	-	-	-
LongVILA-R1 [7]	-	-	-	62.4	-	-	-	-
Baseline: Qwen2.5-VL-7B-Instruct	**74.6**	61.2	51.4	62.4	**59.3**	63.0	37.7	40.3
+ Video-R1 [12]	72.2	59.4	47.0	59.5	49.7	62.0	38.6↑	42.2↑
+ RL (ours)	73.2	**64.0**↑	51.2	**62.8**↑	59.0	64.3↑	39.6↑	41.7↑
Baseline: Qwen2-VL-7B-Base	68.3	57.0	49.6	58.3	53.7	61.4	36.8	39.2
+ RL (ours)	72.1	58.1	52.3	60.8	53.9	63.0	38.4	41.0
+ VideoReasoner (ours)	72.9	60.6	**53.0**	62.0	55.0	**64.6**	**44.6**	**44.1**
Δ	+4.6↑	+3.6↑	+3.4↑	+3.7↑	+1.3↑	+3.2↑	+7.8↑	+4.9↑

Table 2. Performance on public video reasoning benchmarks compared to previous models

Model \Benchmark	VSI-Bench	MMVU
	Overall	MCQ
LLaVA-OV-7B [21]	32.4	49.2
ViLA-1.5-8B [23]	28.9	49.2
VideoTree [43]	-	54.2
LLaVA-Video-7B [58]	36.2	60.2
Qwen2-VL-7B-Instruct [33]	33.4	63.4
Baseline: Qwen2.5-VL-7B-Instruct [2]	38.1	67.5
+ Video-R1 [12]	37.8↓	64.3↓
+ RL (ours)	**39.1**↑	**68.0**↑
Baseline: Qwen2-VL-7B-Base [33]	28.9	61.1
+ RL (ours)	33.7↑	62.4↑

Table 3. Performance on temporal grounding task, † denotes models use Charades-STA's training sets.

Model	Charades-STA [14]			
	mIoU	R1@0.3	R1@0.5	R1@0.7
InternVideo2† [40]	-	-	70.0	**48.9**
TimeSuite† [53]	-	79.4	67.1	43.0
Qwen2.5-VL-7B-Instruct [2]	43.6	76.1	42.9	26.2
Baseline: Qwen2-VL-7B-Base	NA			
+ Cold Start† (ours)	54.1	78.6	62.4	35.8
+ RL† (ours)	**59.1**	**81.9**	**70.4**	45.7

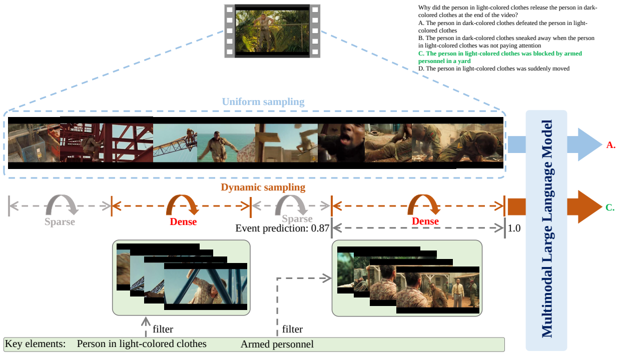
Why did the person in light-colored clothes release the person in dark-colored clothes at the end of the video?
A. The person in dark-colored clothes defeated the person in light-colored clothes
B. The person in dark-colored clothes sneaked away when the person in light-colored clothes was not paying attention
C. The person in light-colored clothes was blocked by armed personnel in a yard
D. The person in light-colored clothes was suddenly moved
Uniform sampling
Dynamic sampling
Sparse
Dense
Sparse
Dense
Event prediction: 0.87
1.0
filter
filter
Key elements: Person in light-colored clothes
Armed personnel
Multimodal Large Language Model
A.
C.

Figure 3. Qualitative Results for video question answering.

tion answering. Compared with the baseline model that uses uniform sampling for responses, the reasoning process can be visualized and is accurate. More cases refer to Appendix.

4.3. Ablation Study

Comparison of different settings in training Table 4 presents the detailed performance of cold start and RL training under single-task and two-tasks settings. First, the cold start consistently improves baseline performance. For RL training, using only temporal grounding data yields improvements only in the long-video scenario of Video-MME, while degrading performance on other tasks. The reason might be that this type of data does not use the final answer as the reward, which affects the accuracy of the model's response. Training using both video QA data and temporal grounding data simultaneously achieved better results than using only video QA data, demonstrating the effectiveness of multi-task RL training.

Table 4. Ablation Study for our proposed method.

Model \Benchmark	Video-MME				LongVB	MLVU	LVBench	VideoEval-Pro
	Short	Medium	Long	Overall	Val	M-Avg	Overall	MCQ
Baseline	68.3	57.0	49.6	58.3	53.7	61.4	36.8	39.2
- Cold Start	70.4	58.1	50.2	59.6	53.8	61.1	38.0	40.3
- RL training								
w/ QA	72.1	57.0	51.7	60.2	51.6	61.3	37.5	41.0
w/ TG	70.4	56.8	51.5	59.5	53.1	60.8	37.4	39.5
w/ QA & TG	72.1	58.1	52.3	60.8	53.9	63.0	38.4	41.0
- Inference								
w/ Event	72.5	60.3	52.4	61.7	54.3	62.4	43.3	43.5
w/ Keyframe	72.6	58.9	52.4	61.3	54.8	63.4	**44.6**	43.0
w/ Event & Keyframe	**72.9**	**60.6**	**53.0**	**62.0**	**55.0**	**64.6**	42.8	**44.1**

Table 5. Analysis of our proposed multi-task Cold Start and multi-task RL training.

Model \Benchmark	Video-MME				LongVB	MLVU	LVBench	VideoEval-Pro
	Short	Medium	Long	Overall	Val	M-Avg	Overall	MCQ
Qwen2.5-VL-7B-Instruct	74.6	61.2	51.4	62.4	59.3	63.0	37.7	40.3
+ Cold Start	74.6	64.0	51.2	63.3	57.3	62.6	38.8	40.2
+ RL	73.2	64.0↑	51.2	62.8↑	59.0	64.3↑	39.6↑	41.7↑
Qwen2-VL-7B-Instruct	70.7	57.6	50.2	59.3	55.2	61.7	39.7	39.6
+ Cold Start	71.5	57.7	49.6	59.6	57.2	62.6	39.6	40.2
+ RL	71.2↑	58.7↑	48.9	59.6↑	54.3	38.3	39.3	38.2
Qwen2-VL-7B-Base	68.3	57.0	49.6	58.3	53.7	61.4	36.8	39.2
+ Cold Start	70.4	58.1	50.2	59.6	53.8	61.1	38.0	40.3
+ RL	72.1↑	58.1↑	52.3↑	60.8↑	53.9↑	63.0↑	38.4↑	41.0↑

Comparison of different settings in inference The last three rows in Table 4 show the model's performance using different multimodal reasoning results during inference. Using either event grounding or keyframe detection results improves output accuracy, with comparable performance across benchmarks. Notably, combining both multimodal elements yields further gains, highlighting their complementary roles and importance in video reasoning.

Comparison of different baselines in training Table 5 presents the detailed training results of three baselines, including Qwen2.5-VL-7B-Instruct, Qwen2-VL-7B-Instruct/Base. As discussed in Sec 1, Instruct Models have stronger preference and bias than Base Models. To ensure a fair comparison, all three baselines are trained using the same dataset. Evaluation results are indicated with arrows, denoting metrics where the RL-trained model outperforms the baseline. When Qwen2.5-VL-7B-Instruct serves as the baseline, improvements are observed in 5 out of 8 indicators. Qwen2-VL-7B-Instruct as the baseline, 3 out of 8 indicators improve, while a sharp decline is observed on MLVU, likely due to instruction disobedience after RL training. Base Models such as Qwen2-VL-7B-Base as the baseline, improvements are observed across all indicators.

5. Conclusion

This work aims to establish a stable and efficient RL training framework for video understanding tasks. Unlike previous frameworks, which merely rely on linguistic reasoning for video content, we propose a novel framework that involves multimodal element reasoning, and our goal is to build and enhance versatile video reasoning capabilities on MLLMs. During the training phase, we proposed a multi-task cold start and a multi-task reinforcement learning. The collaboration of the two training stages can continuously improve the performance of the model and the ability of multimodal element reasoning. Based on the multimodal element reasoning capabilities, in the inference phase, we leverage multimodal reasoning and dynamic sampling to further improve the performance. We verified the efficiency of the proposed framework on a base MLLM. Through cold-start with 3k data and reinforcement learning training with 5k data, the final model significantly outperforms the base model on seven public video benchmarks, and even surpasses the state-of-the-art models trained by large-scale supervised fine-tuning.

References

[1] Anthropic. Introducing the next generation of claude, 2024, 2024. https://www.anthropic.com/news/claude-3-5-sonnet. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6, 7

[3] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 6

[4] ByteDance. Introduction to techniques used in Seed1.6, 2025. https://seed.bytedance.com/en/seed1_6. 2

[5] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02. 1

[6] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 7

[7] 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. 1, 7

[8] 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. 2, 7

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 7

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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. 2

[11] 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. 3

[12] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 1, 2, 3, 5, 6, 7

[13] 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. 2, 6

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 3, 6, 7

[15] 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*, 2025. 1, 2, 3

[16] 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*, 2025. 1

[17] 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. 1

[18] 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. 2

[19] 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. 1, 3

[20] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. 1

[21] 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*, 2024. 7

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[23] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024. 7

[24] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 7

[25] 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*, 2025. 3

[26] 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. 6

[27] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 1

[28] OpenAI. Introducing openai o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025. 1

[29] 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. 3

[30] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024. 2

[31] 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. 1, 3, 4

[32] Yin Song, Chen Wu, and Eden Duthie. aws-prototyping/long-llava-qwen2-7b, 2024. 7

[33] 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*, 2024. 3, 6, 7

[34] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. 1

[35] 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*, 2024. 2, 6

[36] 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*, 2025. 1, 3

[37] 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*, 2025. 3

[38] Xiaodong Wang and Peixi Peng. Open-r1-video. `https://github.com/Wang-Xiaodong1899/Open-R1-Video`, 2025. 1, 2, 3

[39] Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, et al. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*, 2025. 1

[40] 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 *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 7

[41] 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*, 2024. 3

[42] Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. *arXiv e-prints*, pages arXiv–2503, 2025. 5

[43] 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 Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025. 7

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[45] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 2, 6

[46] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 5, 6

[47] 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. 3

[48] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 7

[49] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

[50] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multi-

modal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024. 2, 6

[51] 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*, 2025. 3

[52] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 7

[53] 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 International Conference on Learning Representations*, 2025. 2, 4, 7

[54] 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*, 2025. 2

[55] 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*, 2025. 1

[56] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 7

[57] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. 1

[58] 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*, 2024. 3, 6, 7

[59] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025. 1

[60] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025. 2, 6

[61] 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. 6

[62] 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. 2