

---

# VIDEO-RFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning

---

Qi Wang<sup>1,2</sup>, Yanrui Yu<sup>2</sup>, Ye Yuan<sup>2,3</sup>, Rui Mao<sup>4</sup>, Tianfei Zhou<sup>2,3,\*</sup>

<sup>1</sup>Beijing Institute of Technology, Zhuhai <sup>2</sup>Beijing Institute of Technology

<sup>3</sup>State Key Laboratory of Environment Characteristics and Effects for Near-space

<sup>4</sup>Shenzhen University

<https://github.com/QiWang98/VideoRFT>

## Abstract

Reinforcement fine-tuning (RFT) has shown great promise in achieving human-level reasoning capabilities of Large Language Models (LLMs), and has recently been extended to MLLMs. Nevertheless, reasoning about videos, which is a fundamental aspect of human intelligence, remains a persistent challenge due to the complex logic, temporal and causal structures inherent in video data. To fill this gap, we propose **VIDEO-RFT**, a novel approach that extends the RFT paradigm to cultivate human-like video reasoning capabilities in MLLMs. **VIDEO-RFT** follows the standard two-stage scheme in RFT: supervised fine-tuning (SFT) with chain-of-thought (CoT) annotations, followed by reinforcement learning (RL) to improve generalization. A central challenge to achieve this in the video domain lies in the scarcity of large-scale, high-quality video CoT datasets. We address this by building a multi-expert-driven, cognition-inspired CoT curation pipeline. First, we devise a cognition-inspired prompting strategy to elicit a reasoning LLM to generate preliminary CoTs based solely on rich, structured, and literal representations of video content. Subsequently, these CoTs are revised by a MLLM conditioned on the actual video, ensuring visual consistency and reducing visual hallucinations. This pipeline results in two new datasets, *i.e.* VideoRFT-CoT-102K for SFT and VideoRFT-RL-310K for RL. To further strengthen the RL phase, we introduce a novel semantic-consistency reward that explicitly promotes the alignment between textual reasoning and visual evidence. This reward encourages the model to produce coherent, context-aware reasoning outputs grounded in visual input. Extensive experiments show that **VIDEO-RFT** achieves state-of-the-art performance on six video reasoning benchmarks.

## 1 Introduction

The ability to reason about complex videos lies at the core of human cognitive development [36]. Humans, even infants, exhibit a remarkable capacity to understand videos – recognizing what has happened, inferring what will happen next, and explaining why events occur. Replicating this capability in AI systems has become a central goal in video understanding, and has been extensively studied in the field of computer vision over the past decade [4, 29, 47, 57]. Despite the progress, most AI models remain limited to perceptual-level understanding and struggle to reason about video content with the depth, efficiency, and interpretability that are characteristic of human cognition.

Recently, the rapid development of video MLLMs, such as Video-ChatGPT [28], VILA [24], Video-LLaVA [23], has significantly advanced the capabilities of video understanding. However, these

---

\*Corresponding Author

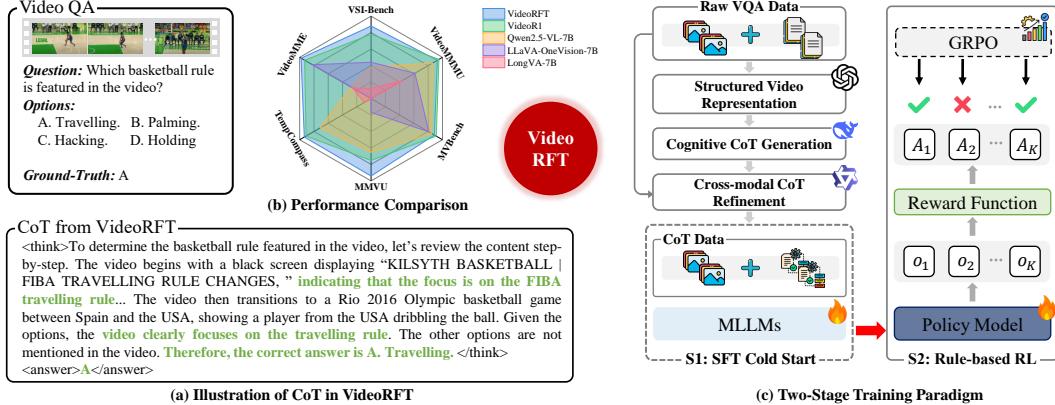


Figure 1: **Overview of VIDEORFT.** (a) An example of CoT derived from VIDEORFT. (b) VIDEORFT achieves leading performance in six datasets. (c) The two-stage RFT underpins the training of VIDEORFT.

models are predominantly answer-driven, *i.e.*, they produce answers without explicitly revealing the reasoning process. VoT [8] overcomes this by introducing a human-like reasoning framework that structures the task of video reasoning into five predefined steps using fixed templates. Nevertheless, such a rigid, template-based approach stands in contrast to the flexibility of human cognition, which enables adaptive reasoning processes based on perceptual inputs [2, 6].

In contrast, the very recent advancements, *e.g.*, OpenAI-o1 [17], DeepSeek-R1 [11], and Kimi-1.5 [38], have shifted focus towards building LLMs that think before answering. These models show strong proficiency in interpreting complex problems, performing multi-step reasoning, and ultimately arriving at correct answers. A key enabler to such capabilities is reinforcement fine-tuning (RFT) [27], which typically commences with a warm-up phase of supervised fine-tuning using CoTs, and subsequently refines the model through reinforcement learning algorithms (*e.g.*, PPO [32], GRPO [34]). Beyond the language domain, pioneering efforts have extended RFT to MLLMs to enhance image-based capabilities [15, 26, 37, 44, 49, 54], and some works [9, 20, 51] that concurrently with ours, show the potential of RFT in the video domain. However, there is a critical challenge remaining unsolved: current video CoT datasets lack the complexity and granularity necessary for advanced video reasoning, which fundamentally limits the ability of models to emulate human-level cognitive capabilities. Moreover, how to ensure that reasoning outputs are faithfully grounded in visual evidence remains underexplored in these works.

Motivated by the above analysis, we propose **VIDEORFT**, a novel reinforcement fine-tuning framework to incentivize the video reasoning capability in MLLMs (see Fig. 1). To overcome the scarcity of video CoTs, we develop a scalable, cognitively inspired pipeline that integrates multiple expert models to collaboratively construct high-quality video CoT datasets. Specifically, we first employ a specialized MLLM to extract structured textual descriptions from videos, capturing fine-grained visual details. These descriptions are then processed by a reasoning-capable LLM (*e.g.*, DeepSeek-R1), which generates initial CoTs through blind reasoning—relying solely on textual input. However, due to the lack of direct visual grounding, such CoTs often contain inconsistencies and hallucinations [14]. To mitigate this issue, we introduce a cross-modal revision stage, wherein a MLLM refines the initial CoTs by incorporating the original video, ensuring consistency with visual evidence. Based on this pipeline, we construct two large-scale datasets, *i.e.*, VideoRFT-CoT-102K and VideoRFT-RL-310K, which together support the RFT process in **VIDEORFT**.

Furthermore, to strengthen the RL phase, we develop a novel semantic-consistency reward that explicitly enhances the visual faithfulness of reasoning outputs in MLLMs. Our key observation is that the reasoning traces of MLLMs are typically structured into three consecutive parts: *question parsing*, *video describing*, and *abstract reasoning*. While the *question parsing* and *abstract reasoning* components are not necessarily grounded in the visual input, the *video describing* part should be closely aligned with the actual visual semantics. Based on this insight, our semantic-consistency reward measures the alignment between the token representations of the *video description* part and the visual features of the input video. This reward is integrated into the GRPO algorithm to guide MLLMs toward generating visually grounded outputs.

**Contributions of this work.** We propose **VIDEORFT**, a novel framework that extends RFT to MLLMs so as to emulate human-like video reasoning capabilities. To achieve this, we first establish a CoT foundation for video RFT by designing a cognitively inspired pipeline to curate large-scale, high-quality video CoT annotations. Furthermore, we introduce a novel semantic-consistency reward to explicitly guide the reasoning trajectories of MLLMs grounded in visual evidence, which enhances the effectiveness of RFT in cross-modal reasoning. Built on these contributions, **VIDEORFT** favorably outperforms advanced competitors on a series of challenging video reasoning benchmarks.

## 2 VIDEORFT CoT Dataset

We first present the construction of VideoRFT-COT and VideoRFT-RL to support RFT in MLLMs.

### 2.1 Data Collection

We extensively collect video question-answer data covering diverse modalities, task types, and cognitive skills. Given the scarcity of high-quality video data in certain domains (*e.g.*, mathematics, science), we additionally incorporate carefully curated image-based instances. The final dataset contains 310K samples in total, supporting diverse answer formats, including multiple-choice (mc), numerical (num), free-form text generation (free), optical character recognition (ocr), and regression (reg). As shown in Fig. 2, the samples are categorized into five groups according to the type of cognitive skills involved in the reasoning process:

- **General:** Commonsense reasoning in open-domain temporal and causal contexts.
- **Mathematics:** Symbolic reasoning and spatial alignment for multi-step logic tasks.
- **Science:** Domain-specific reasoning in physics, chemistry, and medicine, emphasizing causal reasoning and conceptual abstraction.
- **Document:** Targets structured visual parsing and information extraction from complex layouts.
- **Spatiotemporal:** Involves motion prediction, spatial transformation, and relational reasoning.

### 2.2 Cognitively Inspired CoT Generation

To enable MLLMs to acquire human-like reasoning abilities, it is essential to construct a high-quality, cognitively grounded video CoT dataset. We propose an automated pipeline for generating such CoT data. As illustrated in Fig. 3, the pipeline comprises three major stages, and *all the prompts used in the pipeline are provided in the supplementary material*.

**Structured Video Representation.** For each video  $v$ , we generate semantically rich textual descriptions by prompting GPT-4o-mini [16]. The prompt  $P_{\text{rep}}$  is carefully crafted to guide the model to (i) summarize video content with a high-level caption, and (ii) produce analytical, frame-level metadata for uniformly sampled video frames. Each frame is structured in a predefined JSON schema that includes timestamped captions and key visual elements such as objects, actions, scenes, spatial relations, and potential interactions. We denote the structured representation of  $v$  as  $S_v$ .

**Cognitively Inspired CoT Generation.** Given the representation  $S_v$  and a corresponding question  $q$ , we invoke a LLM, *e.g.*, DeepSeek-R1 to answer the question and extract its step-by-step reasoning outputs as the initial CoT, *i.e.*,  $\text{CoT}_v^{(0)}$ :

$$\text{CoT}_v^{(0)} = \text{LLM}(q, S_v, P_{\text{cog}}). \quad (1)$$

Here,  $P_{\text{cog}}$  denotes a composite prompt consisting of five sub-prompts *i.e.*,  $P_{\text{cog}} = [p_s, p_t, p_a, p_v, p_r]$ , each guiding a distinct stage of the reasoning trajectory in a manner that mimics human cognitive processing. Concretely, these sub-prompts are defined as follows. (1) *Simulated observation prompt*

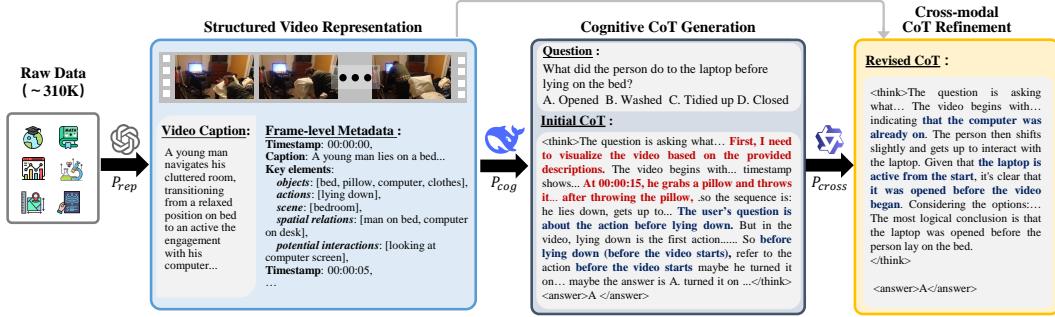


Figure 3: Illustration of the pipeline for cognitively inspired CoT generation.

$(p_s)$ : Instruct the model to simulate viewing the entire video and form an initial high-level understanding. (2) *Task understanding prompt* ( $p_t$ ): Encourage analysis of the question  $q$  to infer the task type (e.g., fact, reason, causal relationship). (3) *Selective focus prompt* ( $p_a$ ): Direct attention to specific temporal segments of the video relevant to  $q$ . (4) *Visual reasoning prompt* ( $p_v$ ): Ground the reasoning process in visual content, encouraging analysis over objects, actions, spatial-temporal relations, and event transitions. (5) *Reflective answering prompt* ( $p_r$ ): Guide the model to derive the final answer, optionally incorporating self-verification or reflection to ensure reasoning quality.

**Cross-modal CoT Refinement.** A key limitation in the initial CoTs is that they might suffer from visual hallucinations due to the lack of visual cues in Eq. 1. To resolve this issue, we introduce a cross-modal refinement strategy to revise the CoT so that it aligns better with the actual video input. Specifically, we prompt a MLLM, *i.e.*, Qwen2.5-VL [1], to compare the initial CoT with the video  $v$ , identify inconsistencies, and perform necessary revisions:

$$\text{CoT}_v = \text{MLLM}(v, \text{CoT}_v^{(0)}, P_{\text{cross}}). \quad (2)$$

Here the prompt  $P_{\text{cross}}$  is designed to guide the MLLM to: (i) verify the cross-modal alignment of  $\text{CoT}_v^{(0)}$  with the content of video  $v$ , (ii) localize and explain any visual-textual inconsistencies, and (iii) revise the CoT to enhance visual grounding while preserving its original logical structure.

Finally, we apply a filtering stage to ensure the factual correctness of the resulting CoT annotations. For structured tasks with clear ground-truth labels, we directly exclude samples with incorrect final answers. For open-ended tasks, we remove samples exhibiting low semantic consistency (measured by CLIP [31]) between the generated answer and the reference answer. This filtering process ensures that the resulting CoT dataset maintains high quality and factual reliability. After filtering, 102K high-confidence samples are retained from the initial pool of 310K, forming VideoRFT-CoT-102K for supervised fine-tuning.

## 2.3 Data Analysis

Fig. 4 presents a comparative analysis of the CoTs in our proposed VideoRFT-CoT-102K and Video-R1 [9]. As shown in Fig. 4 (a), the CoTs in our dataset exhibit a broader distribution and longer average token length compared to those in Video-R1 (Fig. 4(c)), indicating that our VideoRFT-CoT-102K contains more elaborate, fine-grained, and nuanced reasoning processes. Additionally, the word cloud in Fig. 4 (b) reveals that CoTs in VideoRFT-CoT-102K are dominated by dynamic, video-centric concepts such as “video”, “main”, “happen”, and “first”. The lexical profile reflects an emphasis on narrative structure and temporal progression, which are the key characteristics of complex video understanding. In contrast, Video-R1 (Fig. 4 (d)) features frequent refer-

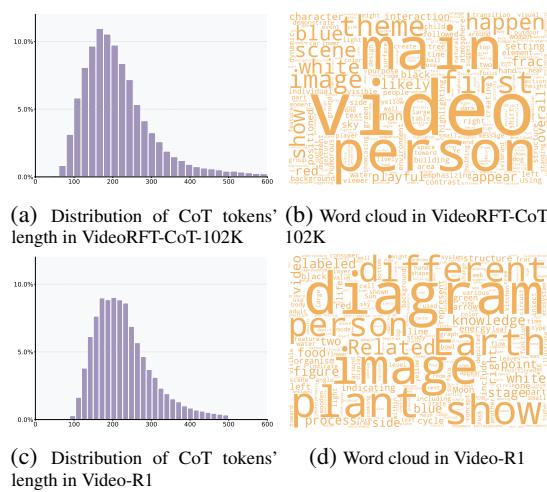


Figure 4: Comparison of CoT dataset in VideoRFT-CoT-102K and Video-R1.

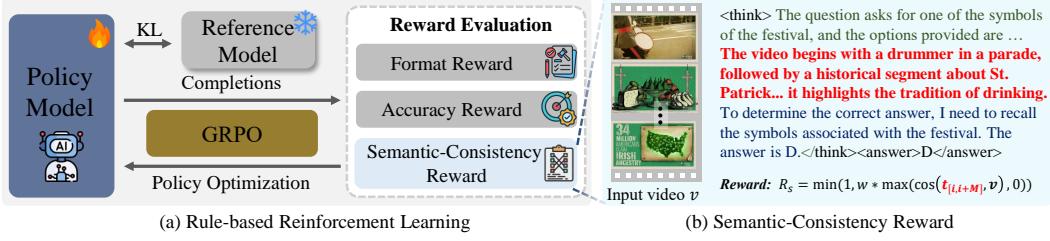


Figure 5: Illustrations of (a) rule-based RL, and (b) the computation of semantic-consistency reward  $R_s$ . The reasoning outputs are color-coded to highlight *question parsing* (green), *video description* (red) and *abstract reasoning* (blue). Only the red part is involved in the computation of  $R_s$  (see §3.2).

ences to static or declarative content, *e.g.*, “*diagram*”, “*image*”, “*plant*”, and “*Earth*”, suggesting a stronger bias toward factual descriptions rather than deep reasoning. These results highlight that VideoRFT-CoT-102K offers greater expressiveness in reasoning depth and aligns more closely with the demands of real-world video reasoning tasks. Hence, it provides a better foundation for training video MLLMs with advanced reasoning capabilities.

### 3 Video Reinforcement Fine-Tuning

This section presents our approach for video reinforcement fine-tuning as shown in Fig. 5. We first provide a brief overview of Group Relative Policy Optimization (GRPO) [34] in §3.1, and then elaborate on the proposed rule-based reward for efficient reinforcement fine-tuning in §3.2.

#### 3.1 Group Relative Policy Optimization

GRPO [34] is a computationally efficient rule-based RL algorithm designed specifically for training large reasoning models. Unlike traditional RL methods such as PPO [32], which require four models (policy, value, reward, and reference), GRPO simplifies the approach by eliminating the value model, significantly reducing memory requirements and training complexity. GRPO operates by generating  $K$  candidate responses  $\{o_1, o_2, \dots, o_K\}$  for each query  $q$ . These responses are then evaluated using defined reward functions, yielding rewards  $\{r_1, r_2, \dots, r_K\}$ . Afterwards, these rewards are normalized to calculate the advantage  $A_i$  for each response as:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_K\})}{\text{std}(\{r_1, r_2, \dots, r_K\})}, \quad (3)$$

where  $\text{mean}$  and  $\text{std}$  denote the mean and standard deviation of the rewards, respectively. Subsequently, the model is optimized through maximization of the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{[q, \{o_i\}]} \frac{1}{K} \sum_{i=1}^K \left[ \min \left( \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}} A_i, \text{clip} \left( \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) \right], \quad (4)$$

where  $\theta$  denotes model parameters to be updated,  $\pi_\theta$  and  $\pi_{\theta_{\text{old}}}$  are the current and old policy model,  $\pi_{\text{ref}}$  indicates the reference policy,  $\beta$  is the KL divergence regularization coefficient, and  $\epsilon$  is a regularization coefficient that prevents the policy from deviating too far from the reference model.

#### 3.2 Rule-based Reward Modeling in VIDEORFT

The rewards in Eq. 3 are derived from rule-based reward functions, which represent a foundational step in rule-based RL by simply evaluating whether model predictions exactly match ground-truth answers. Two highly resilient rule-based rewards are the Format Reward and Accuracy Reward, which are consistently utilized in DeepSeek-R1 and its follow-ups. However, in the context of cross-modal reasoning, these rewards are insufficient to provide explicit guidance to MLLMs towards visually grounded reasoning. To address this limitation, we introduce a semantic-consistency reward, which enforces the grounding of generated reasoning content in the visual input.

### 3.2.1 Semantic-Consistency Reward

The reward is motivated by the observation that the reasoning trace generated by video MLLMs typically consists of three distinct parts, *i.e.*, *question parsing*, *video describing* and *abstract reasoning*, as shown in Fig. 5. Among them, the *video describing* stage represents the model’s understanding of visual content, which is the foundation for subsequent reasoning. Therefore, the reward is designed to selectively promote alignment between this stage and the input video.

Formally, to isolate the *video describing* sentence from the generated response, we apply a regular expression to locate the first full stop. Empirically, the text following this full stop corresponds to the model’s interpretation of visual content. From this point, we extract a fixed-length span of  $M$  tokens, denoted  $t_{[i, i+M]}$ , and encode it using SigLIP [48]’s text encoder:  $t_{[i, i+M]} = \text{SigLIP}_{\text{text}}(t_{[i, i+M]})$ . Additionally, we uniformly sample  $F$  frames  $\{v^{(0)}, \dots, v^{(F-1)}\}$  from video  $v$ , and compute the visual representation of each frame  $v^{(i)}$  via SigLIP’s image encoder:  $v^{(i)} = \text{SigLIP}_{\text{image}}(v^{(i)})$ . Then the final video representation  $v$  is naturally obtained by averaging the frame embeddings:  $v = \frac{1}{F} \sum_{i=0}^F v^{(i)}$ . We then define the semantic-consistency reward as:

$$R_s = \min(1, w \times \max(\cos(t_{[i, i+M]}, v), 0)), \quad (5)$$

where  $\cos(\cdot, \cdot)$  denotes cosine similarity, and  $w = 2$  is a scaling constant. The  $\max(\cdot, 0)$  ensures non-negativity of the reward, while the  $\min(\cdot, 1)$  stabilizes training by bounding the reward. This stage-aware formulation allows us to reward only the part of reasoning tied to visual comprehension, without penalizing abstract reasoning that appropriately extends beyond the visual scope. The result is enhanced semantic fidelity, reduced hallucinations, and improved alignment during RL.

## 3.3 Overall Reward

**VIDEORFT** uses three types of rewards for RL:

- **Format Reward.** During RL, we incorporate the widely-used format reward to guide the model in generating its reasoning process and final answer in a structured format. This reward, denoted as  $R_f$ , ensures that the model’s output adheres to a predefined structure: the reasoning process must be enclosed within `<think>...</think>` tags, and answers within `<answer>...</answer>` tags. Compliance is verified via regular expression matching, and a binary reward is assigned accordingly.
- **Accuracy Reward.** To provide reliable supervision across heterogeneous tasks, we adopt task-specific accuracy metrics: Exact Match for multiple-choice and numerical questions, ROUGE for open-ended generation, Word Error Rate for OCR tasks, and a scaled relative accuracy for regression problems. These tailored evaluations ensure the reward  $R_a$  aligns with each task.
- **Semantic-Consistency Reward.** The semantic-consistency reward  $R_s$ , defined in Eq. 5, promotes alignment between the reasoning text and the input visual information.

The overall reward  $R$  for a sample is computed as follows:

$$R = R_f + R_a + \mathbb{1}[R_a > 0] \cdot R_s, \quad (6)$$

where  $\mathbb{1}[R_a > 0]$  is the indicator function that returns 1 if  $R_a > 0$  and 0 otherwise. This indicator function acts as a gate to ensure  $R_s$  is activated only when  $R_a$  is non-zero, thus avoiding the reinforcement of semantically plausible but factually incorrect reasoning.

## 4 Experiment

### 4.1 Experimental Setup

**Benchmark and Metric.** Following previous works [9, 18, 51], we evaluate our approach on six video reasoning and understanding benchmarks: VSI-Bench [43], VideoMMM [13], MMVU [55], MVbench [19], TempCompass [25], and VideoMME [10], covering spatial reasoning, knowledge-intensive video QA, temporal logic, and general video understanding. Following conventions, we only use the subset of multiple-choice samples in MMVU, and VideoMME is evaluated without subtitles. Average accuracy is adopted as the evaluation metric.

Table 1: **Performance Comparison.** The best results are highlighted in **bold**.  $^\dagger$ : Results are obtained using larger input resolutions, up to  $768 \times 28 \times 28$  and 768 sampled frames, while ours are  $256 \times 28 \times 28$  and 32.

Model	Pub	Video Reasoning			Video Understanding		
		VSI.	VideoMMU	MMVU	MV.	TempC.	VideoMME
<b>• Proprietary Models</b>							
GPT-4o [16]	-	34.0	61.2	75.4	-	-	71.9
<b>• Open-Source Models</b>							
LLaMA-VID [21]	ECCV 24	-	-	-	41.9	45.6	-
ShareGPT4Video [3]	NeurIPS 24	-	-	-	51.2	-	39.9
VideoLLaMA2 [5]	arXiv 24.06	-	-	44.8	54.6	-	47.9
LongVA-7B [50]	TMLR 24	29.2	23.9	-	-	56.9	52.6
VILA-1.5-8B [24]	CVPR 24	28.9	20.8	-	-	58.8	-
LLaVA-OneVision-7B [18]	TMLR 24	32.4	33.8	49.2	56.7	-	58.2
mPLUG-Owl3-8B [46]	ICLR 25	-	-	-	54.5	-	53.5
Qwen2.5-VL-7B [1]	arXiv 25.02	31.8	47.4	61.3	59.4	69.2	52.8
<b>• Concurrent R1-based Models</b>							
Video-R1 [9]	arXiv 25.03	35.8	<b>52.3</b>	63.8	63.9	73.2	59.3
TinyLLaVA-Video-R1 [51]	arXiv 25.04	-	-	46.9	-	49.5	46.6
VideoChat-R1 [20]	arXiv 25.04	-	-	-	<b>67.9</b> <sup>†</sup>	-	-
<b>VIDEORFT</b>	-	<b>36.8</b>	51.1	<b>68.5</b>	62.1	<b>73.7</b>	<b>59.8</b>

**Model Training.** We follow the RFT to train **VIDEORFT** in two stages: the warm-up SFT stage and the rule-based RL stage. Specifically, the SFT stage equips the model with the ability to generate correct responses for diverse questions, and is trained based on VideoRFT-CoT-102K. The rule-based RL stage is based on VideoRFT-RL-310K using the reward in Eq. 6 to optimize structured reasoning and ensure factual validity. The RL training is implemented using the HuggingFace TRL library [39], and our codebase is built upon Open-R1 [7].

**Implementation Details.** We use Qwen2.5-VL-7B [1] as the base model and train **VIDEORFT** on 8 NVIDIA A800 GPUs, with 80GB each. For efficiency, the video input is limited to 16 frames, with each frame processed into  $128 \times 28 \times 28$  resolution during training, where  $28 \times 28$  is the size of each image patch, and 128 denotes the number of patches. During inference, we increase the number of frames to 32 and the resolution to  $256 \times 28 \times 28$ . For efficiency, we use a lightweight version of SigLIP with 400M parameters in computing the semantic-consistency reward. The entire model is trained for one epoch of SFT followed by 1K steps of RL.

## 4.2 Main Result

As shown in Table 1, we compare **VIDEORFT** against a variety of baselines, including proprietary models (*i.e.*, GPT-4o [16]), Open-Source MLLMs (*e.g.*, Qwen2.5-VL [1], VILA [24], LongVA [50]), and contemporaneous models (*e.g.*, Video-R1 [9], TinyLLaVA-Video-R1 [51], VideoChat-R1 [20]).

Several key observations can be drawn from the results. First, compared to our base model, *i.e.*, Qwen2.5-VL-7B, **VIDEORFT** achieves significant improvements across all six benchmarks, *e.g.*, **+5.0%** on VSI-Bench, **+7.2%** on MMVU, and **+7.0%** on VideoMME. This demonstrates the effectiveness of our approach in incentivizing video reasoning capabilities in MLLMs. Moreover, **VIDEORFT** consistently outperforms all non-RL Open-Source MLLMs. Second, **VIDEORFT** surpasses the proprietary GPT-4o on VSI-Bench by **+2.8%**, highlighting the strong potential of RFT in bridging the performance gap with closed-source models in the task of video reasoning. Third, when compared to contemporaneous works, our model delivers the best overall performance, **ranking first on four out of six** benchmarks. This validates the superiority and generalization ability of our method in comparison with recent endeavors.

## 4.3 Diagnostic Experiment

To gain deeper insights into **VIDEORFT**, we conduct a set of diagnostic experiments, as in Table 2.

Table 2: Diagnostic experiments for **VIDEORFT**.

Model	Video Reasoning			Video Understanding		
	VSI-Bench	VideoMMU	MMVU	MVBench	TempCompass	VideoMME
<b>• Training Data</b>						
w/o CoT Refinement	34.5	48.1	64.8	58.3	72.4	52.8
<b>• Training Paradigm</b>						
SFT only	31.7	48.5	60.5	57.0	68.4	54.1
RL only	32.1	47.4	63.5	59.2	70.8	51.9
<b>• Reward Modeling</b>						
$R = R_f + R_a$	33.2	49.1	66.4	61.1	72.4	58.5
$R = R_f + R_a + R_s$	34.6	50.2	65.2	61.4	<b>73.9</b>	56.3
<b>• Full Model</b>						
<b>VIDEORFT</b>	<b>36.8</b>	<b>51.1</b>	<b>68.5</b>	<b>62.1</b>	73.7	<b>59.8</b>

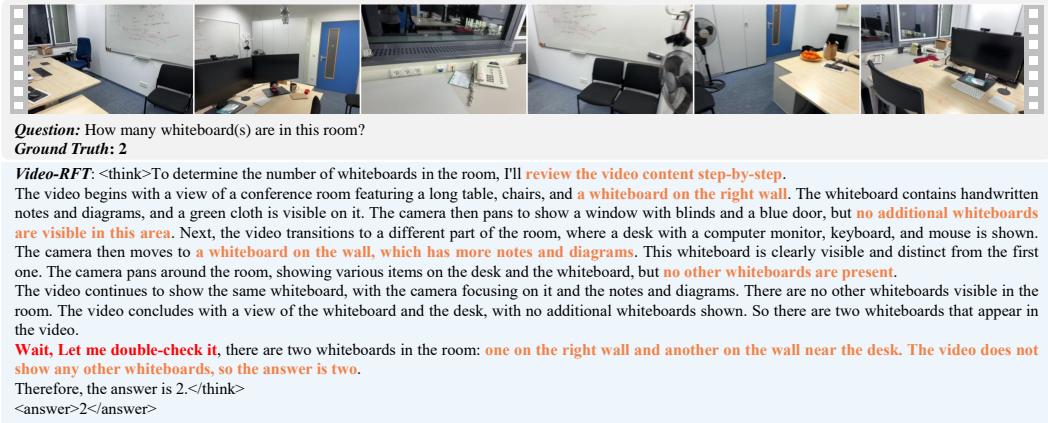


Figure 6: Illustration of reasoning traces derived from **VIDEORFT** in VSI-Bench.

**Training Data.** To assess the impact of the cross-modal refinement in Eq. 2 on data quality, we conduct SFT+RL using the data generated by Eq. 1, denoted as  $\text{CoT}_v^{(0)}$ , resulting in a variant, *i.e.*, without CoT Refinement. As shown in Table 2, this consistently leads to performance drops across all six benchmarks, *i.e.*, **-2.3%** on VSI-Bench, **-3.7%** on MMVU, and **-7.0%** on VideoMME. This demonstrates that our cross-modal refinement effectively mitigates errors and hallucinations in the initial CoTs, thereby enhancing data quality and ultimately improving model performance.

**Training Paradigm.** To validate the effectiveness of the RFT training paradigm in our approach, we build two baseline approaches: *i.e.*, *SFT only* and *RL only*. The former trains the model solely with supervised fine-tuning on VideoRFT-CoT-102K, while the latter, also known as the “zero” model in DeepSeek-R1, relies exclusively on RL without prior SFT. As seen from Table 2, the *RL only* surpasses the *SFT only* counterpart on four out of six datasets, indicating the capability of RL in stimulating more generalized reasoning capabilities. When combining both stages as in RFT, our full model **VIDEORFT** achieves the best results, substantially outperforming the two baselines across all datasets. This highlights the complementary strengths of SFT for stable initialization and RL for reasoning enhancement in tackling video reasoning.

**Reward Modeling.** We further examine the effect of the reward defined in Eq. 6. Specifically, we compare two ablated variants: the first uses only the Format Reward and Accuracy Reward ( $R = R_f + R_a$ ), while the second incorporates all three rewards directly ( $R = R_f + R_a + R_s$ ). As seen from Table 2, adding the semantic-consistency reward  $R_s$  consistently improves performance over the first variant, validating its effectiveness. Finally, our full reward formulation, which conditionally activates  $R_s$  via a gating mechanism (*i.e.*, the indicator function  $\mathbb{1}[R_a > 0]$ ), achieves the best overall results. Notably, it brings substantial gains especially for video reasoning benchmarks, *i.e.*, **+2.2%** on VSI-Bench, **+3.3%** on MMVU, and **+3.5%** on VideoMME.

**Aha Moment in VIDEORFT.** Fig. 6 exhibits an Aha Moment in **VIDEORFT**, where it behaves in a human-like manner by pausing to double-check its inference before finalizing the answer, as seen in the phrase “Wait, let me double-check it”. Such behavior suggests that the model is not simply recalling learned patterns, but is instead engaging in internal feedback loops to re-evaluate evidence and refine its inference.

## 5 Related Work

### 5.1 Multimodal Reasoning in MLLMs

Enabling reasoning in MLLMs has become a central objective in recent research [22, 35, 52, 56]. In the image domain, early works such as MMCOT [52] and DDCoT [56] disentangle perception and reasoning by treating visual understanding as input prompts for subsequent inference. In the video domain, VoT [8] and STEP [30] decompose video reasoning into predefined stages, employing template-based prompting to facilitate multi-step inference. DoraemonGPT [45] models video understanding through symbolic memory and external tool sequences, yet still follows a modular reasoning paradigm. While these methods offer structured supervision, their rigid designs often limit generalization across diverse temporal and causal scenarios. Recently, rule-based RL has emerged as a promising paradigm for promoting multimodal reasoning in MLLMs. Pioneering efforts such as Visual-RFT [26], R1-VL [49], and Reason-RFT [37] directly adapt rule-based RL to image perception tasks. Follow-ups like Vision-R1 [15] and R1-OneVision [44] further demonstrate its effectiveness in enabling CoT reasoning on images. Concurrently, this paradigm has also been explored for video understanding [9, 20, 51]. Despite encouraging progress, these methods face a fundamental bottleneck: the lack of large-scale, high-quality video CoT datasets, which limits the full potential of RFT in the video domain. Our work addresses this gap by proposing a scalable and cognitively inspired pipeline to automatically mine high-quality CoT annotations for videos. Beyond this, we introduce a novel reward modeling strategy based on cross-modal semantic consistency, which explicitly guides MLLMs to generate visually grounded reasoning traces, and proves to be highly effective in improving model performance.

### 5.2 Multimodal CoT Dataset Construction

CoT has proven effective for enhancing the reasoning capabilities of LLMs by encouraging step-by-step reasoning [41, 53]. Constructing high-quality CoT data in multimodal settings, particularly for video reasoning, remains a major challenge due to the temporal complexity and visual ambiguity of video data [12, 33, 40]. Recent works have explored CoT construction in both image and video domains. LLaVA-CoT [42], Vision-R1 [15], and R1-OneVision [44] simply convert visual inputs to textual descriptions before reasoning. This often leads to hallucinations and weak semantic alignment. Video-R1 [9] adopts a simplistic prompting strategy that encourages MLLMs to generate CoT by inserting “let me think”, “wait”, *etc.* into responses. However, such CoTs merely mimic the surface form of human thinking without engaging in genuine reasoning. VideoEspresso [12] generates CoT data by prompting GPT-4o with a small set of selected key frames. Due to the sparse visual context and reliance on a text-only model, the generated CoTs often lack grounding in the actual video content and are prone to hallucinations. In contrast, our CoT data combines the reasoning abilities of reasoning LLMs and the multimodal abilities of MLLMs, ensuring the reasoning depth and visual grounding of CoT data. Moreover, we use cognition-inspired prompts to enable the reasoning model to generate CoT data that is more in line with human cognition.

## 6 Conclusion

In this work, we introduce **VIDEORFT**, a novel approach for incentivizing cognitive video reasoning capabilities in MLLMs through reinforced fine-tuning. To accomplish this, we propose a cross-modal pipeline that generates high-quality cognitive video CoT data simulating human reasoning processes, resulting in two large-scale datasets: VideoRFT-CoT-102K and VideoRFT-RL-310K. Furthermore, to strengthen the RL phase, we develop semantic-consistency guided reward to explicitly encourage the alignment between reasoning traces and visual evidence. Extensive experiments across six benchmarks demonstrate that **VIDEORFT** consistently surpasses a variety of advanced MLLMs. We expect this work to lay a foundation for future efforts in RFT-based video reasoning.

## Acknowledgments

This work is supported by the NSFC (Grant Nos. 62576035, 62225203, 62532007), the National Key R&D Program of China (Grant No. 2022YFB2702100), Beijing Natural Science Foundation (L252036), CAAI-Lenovo Blue Sky Research Fund, the Beijing Municipal Science and Technology Commission and Zhongguancun Science Park Management Committee (Z231100007423003).

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645, 2008.
- [3] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024.
- [4] Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, et al. Meed: Unlocking multi-event causal discovery in video reasoning. In *NeurIPS*, 2024.
- [5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-langs. *arXiv preprint arXiv:2406.07476*, 2024.
- [6] Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013.
- [7] Hugging Face. Open r1: A fully open reproduction of deepseek-r1. <https://github.com/huggingface/open-r1>, 2025.
- [8] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *ICML*, 2024.
- [9] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in llms. *arXiv preprint arXiv:2503.21776*, 2025.
- [10] Chaoyou Fu, Yuhai Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.
- [11] 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.
- [12] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *CVPR*, 2025.
- [13] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.
- [14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):1–55, 2025.
- [15] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [16] 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.

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

[18] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2024.

[19] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping  
Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024.

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

[21] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language  
models. In *ECCV*, 2024.

[22] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing  
advanced visual reasoning ability of large language models. In *EMNLP*, 2024.

[23] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual  
representation by alignment before projection. In *EMNLP*, 2024.

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

[25] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou.  
Tempcompass: Do video llms really understand videos? In *ACL*, 2024.

[26] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang.  
Visual-rft: Visual reinforcement fine-tuning. In *ICCV*, 2025.

[27] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning  
with reinforced fine-tuning. In *ACL*, 2024.

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards  
detailed video understanding via large vision and language models. In *ACL*, 2024.

[29] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring  
modular reasoning models for video question answering. In *CVPR*, 2024.

[30] Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang,  
Yuetong Zhuang, and Tat-Seng Chua. Step: Enhancing video-llms' compositional reasoning by spatio-  
temporal graph-guided self-training. In *CVPR*, 2025.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish  
Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from  
natural language supervision. In *ICML*, 2021.

[32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy  
optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[33] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng  
Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for  
chain-of-thought reasoning. In *NeurIPS*, 2024.

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

[35] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Unlocking video-llm via agent-of-thoughts distillation.  
*arXiv preprint arXiv:2412.01694*, 2024.

[36] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

[37] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang  
Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*,  
2025.

[38] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[39] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.

[40] 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.

[41] 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. In *NeurIPS*, 2022.

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

[43] Jihai Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025.

[44] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[45] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*, 2024.

[46] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

[47] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2020.

[48] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *CVPR*, 2023.

[49] 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.

[50] 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. *Trans. Mach. Learn. Res.*, 2024.

[51] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025.

[52] Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

[53] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR*, 2023.

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

[55] 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 *CVPR*, 2025.

[56] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*, 2023.

[57] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state the main contributions, which are supported throughout the paper, especially in Sections 3 and 4.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Our CoT data generation relies on existing reasoning LLMs and MLLMs, which may introduce biases or hallucinations, and the training requires substantial computational resources.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not contain formal theoretical results or proofs

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 5 provides implementation details including datasets, model architecture, hardware used, and evaluation protocol. And the code, data, and model weights will be open-sourced in the future.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: **[Yes]**

Justification: The constructed datasets and training code with documentation will be open-sourced upon paper acceptance

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: **[Yes]**

Justification: Section 5.1 describes benchmarks, metrics, and model training settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: **[No]**

Justification: Error bars and statistical tests are not currently reported due to time and computational resource constraints, but multiple runs and more robust metrics could be included in the camera-ready version.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: Section 5.1 mentions training was done on 8 A800 GPUs (80GB) and specifies input resolutions and training steps.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: Our work complies with the NeurIPS Code of Ethics. No human subjects or sensitive data were involved, and we avoid harmful generation or misuse.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: The method benefits video understanding but may risk misuse in misinformation or surveillance if deployed irresponsibly.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: We will release the models and datasets under licenses with usage guidelines. Video-based CoT data will be filtered to avoid unsafe content.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: All models and datasets used are cited appropriately with license and source attribution, including SIGLIP, Qwen2.5-VL, gpt-4o-mini, and DeepSeek-R1.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release two new datasets along with structured documentation including data format, source description, and limitations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our research does not involve human subjects or crowdsourcing. All data are derived from public or synthetic video sources.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No IRB approval is required as our work does not involve human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used as core components in our research in two ways: (1) DeepSeek-R1 and GPT-4o-mini are used in our video CoT data generation pipeline, and (2) we use Qwen2.5-VL-7B as our base model for training. We will include the detailed prompts in the supplementary materials.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Additional Experiments

### A.1 Effect of Model Scaling on Reasoning Ability

Table 3: Performance Comparison with Small-scale Models

Model	Video Reasoning			Video Understanding		
	VSI.	VideoMMMU	MMVU	MV.	TempC.	VideoMME
TinyLLaVA-Video-R1-3B	-	-	46.9	-	49.5	46.6
<b>VIDEORFT-3B</b>	32.5	41.1	55.1	59.5	61.0	45.4
<b>VIDEORFT-7B</b>	36.8	51.1	68.5	62.1	73.7	59.8

To further assess the scalability and robustness of our method under limited computational resources, we train an additional 3B variant of our model. While its performance is naturally lower than the original 7B version, the 3B model still performs competitively across benchmarks. Crucially, when compared to TinyLLaVA-Video-R1-3B, a contemporary 3B model specifically optimized for lightweight deployment, our 3B variant outperforms it, with gains of **+8.2%** on MMVU and **+11.5%** on TempCompass. This significant margin confirms the competitiveness of our approach even in small-scale settings, and underscores its strong generalization and reasoning abilities under resource-constrained conditions.

### A.2 Hyperparameter Analysis

Table 4: Hyperparameter experiments for VIDEORFT.

Hyperparameter	Video Reasoning			Video Understanding		
	VSI-Bench	VideoMMMU	MMVU	MVBench	TempCompass	VideoMME
$w = 1$	34.2	49.2	67.9	<b>62.6</b>	73.5	<b>61.1</b>
$w = 2$	<b>36.8</b>	<b>51.1</b>	<b>68.5</b>	62.1	<b>73.7</b>	59.8
$w = 3$	35.6	50.1	67.9	60.8	72.3	60.4
$w = 4$	35.4	49.7	67.6	62.0	73.1	59.1

To evaluate the effect of the scaling factor  $w$  in the semantic reward, we conducted a hyperparameter study across multiple benchmarks. As shown in Table 4, the performance varies with different values of  $w$ , with  $w=2$  consistently achieving the best or near-best results across all benchmarks. This suggests that  $w=2$  provides a favorable balance between semantic alignment and factual precision. In comparison, both smaller ( $w=1$ ) and larger ( $w=3$  or  $w=4$ ) values lead to a slight decline in performance, implying that insufficient or excessive emphasis on semantic similarity may hinder the overall reasoning and understanding capabilities. These findings highlight that while performance is relatively stable across a range of scaling values, setting  $w=2$  consistently yields optimal results, demonstrating the effectiveness of moderate semantic weighting.

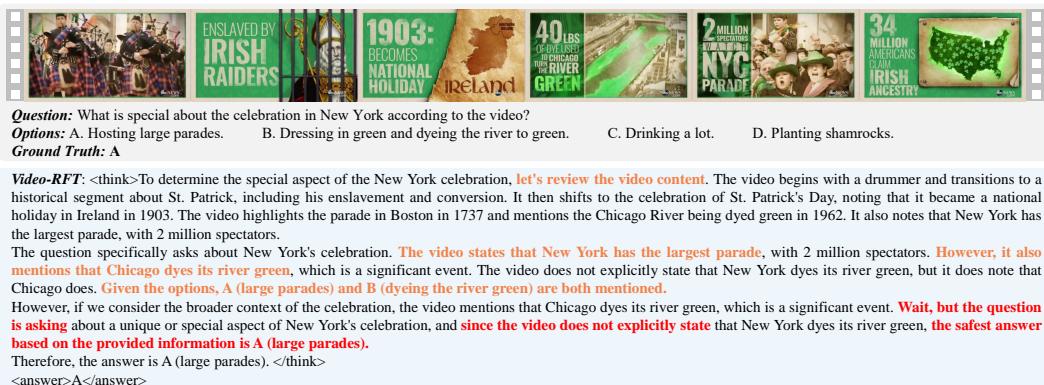


Figure 7: Illustration of reasoning traces derived from VIDEORFT in VideoMME.

### A.3 Aha Moment in **VIDEORFT**

Figure 7 illustrates additional instances of the “Aha Moment” behavior observed in **VIDEORFT**. Prior to reaching a final decision, the model exhibits a human-like tendency to pause and scrutinize its intermediate reasoning steps. This reflective process, indicative of deliberative reasoning, is marked in red within the figure.

## B Limitations and Future Directions

We consider this work a strong foundation for advancing video reasoning research using MLLMs. Several avenues warrant further investigation in future studies:

- **Handling of Challenging Video Scenarios.** While our model demonstrates strong performance across various benchmark tasks, its effectiveness may be affected under complex conditions such as rapid motion or severe visual occlusion. Incorporating finer-grained visual signals and higher frame-rate sampling may help mitigate these challenges, enabling more accurate and robust visual representations.
- **CoT Data Reliance.** The quality of the generated CoT annotations is closely tied to the capabilities of the underlying reasoning language model (e.g., DeepSeek-R1). Future work could explore leveraging more advanced reasoning models to further enhance the quality of CoT data, potentially leading to improved performance of **VIDEORFT**.

## C Potential Social Impacts

- **Positive Impacts.** Enhanced video understanding enabled by **VIDEORFT** can benefit various applications. In education, it facilitates the development of intelligent tutoring systems using video content. In security, it improves the efficiency and accuracy of surveillance video analysis. Moreover, it supports content moderation by aiding in the detection and filtering of inappropriate material.
- **Negative Impacts.** There exists a risk of misuse. Misinterpretation of video content due to over-reliance on automated analysis could lead to the spread of misinformation. Additionally, in surveillance scenarios, the deployment of such systems may raise concerns regarding privacy and ethical use.

## D Detailed Prompt Used in CoT Generation

Here, we provide detailed prompts for each step of the **VIDEORFT** CoT generation process.

### D.1 Video CoT Generation

► **Structured Video Representation:** For all videos, we sample them at 1FPS and input them into GPT-4o-mini, and generate a structured representation according to the prompt  $P_{\text{rep}}$ :

**Structured Video Representation**

► **System Prompt:**  
You are a video analysis assistant designed to produce rich, analytical per-frame captions from video inputs.

► **User Prompt:**  
<Input Video>  </Input Video>

**Task:**

1. **Overall Video Caption:**
  - **video\_caption:** A concise 1-2 sentence (20-30 words) summary capturing the main theme or action of the video.

## Structured Video Representation (Continued table)

### 2. Per-Frame Metadata:

Uniformly sample frames and wrap them into a JSON list. Each element must include:

- **timestamp**: "HH:MM:SS", aligned with the sampling interval.
- **caption**: 2-3 sentences (30-50 words) describing the scene in detail.
- **key\_elements**: an object with fields:
  - **objects**: list of detected objects/entities (strings)
  - **actions**: list of ongoing actions or movements (strings)
  - **scene**: *e.g.* "kitchen", "urban street"
  - **notable\_features**: list of distinctive colors, textures, or patterns (strings)
  - **spatial\_relations**: list of spatial relationships (*e.g.* "cup on table", "person left of car")
  - **human\_attributes**: object or null. If present, include:
    - \* **gender**: "male", "female", "unknown"
    - \* **clothing**: brief description
    - \* **posture**: "standing", "sitting"
  - **potential\_interactions**: list of possible interactions (strings)

### General Instructions:

- **Temporal Consistency**: Reference continuing actions from the previous frame and highlight any changes.
- **Uncertainty**: If confidence < 0.6 or object visibility < 50%, append [Uncertain] to the caption.
- **Implied Actions**: Describe preparatory movements (*e.g.* "hand reaching toward door handle" vs. "holding door handle").
- **Output Requirements**: Wrap all per-frame objects into a single JSON list matching the number of sampled frames.

► **Cognitively Inspired CoT Generation:** We invoke DeepSeek-R1 to answer the question and extracts its step-by-step reasoning outputs with the prompt  $P_{\text{cog}}$  as the initial CoT.

## Cognitively Inspired CoT Generation

### ► System Prompt:

You are an AI assistant helping a user answer questions about a video. When the user asks a question, you respond by imagining you are watching the video with full attention, just like a human would. Your task is to reason visually and logically about the video content to answer the user's question.

Follow this multi-step reasoning approach:

1. **Simulate Browsing the Video**: Imagine you are watching the entire video from beginning to end. Build a general sense of what is happening.
2. **Understand the Question**: Reflect on what the user is asking. Think carefully about what kind of answer is needed (*e.g.*, a fact, a reason, a comparison).
3. **Localize Relevant Moments**: Consider which parts of the video are most related to the question. Focus on those segments in your mental replay.
4. **Visual Reasoning**: Describe what you "see" in those segments using natural visual language (*e.g.*, "The video shows... ", "In the second half of the video... "). Analyze and interpret the visual content to build your answer.
5. **Answer Thoughtfully**: Provide a clear and direct answer. Ensure your reasoning is consistent with the visual events you described.

Guidelines for Responses:

## Cognitively Inspired CoT Generation — Continued

- Don't expose in the output that you are answering based on text information. Use statements imitating watching a video to answer.
- Don't directly refer to any textual metadata such as "captions", "description", "frame-level metadata", "key elements", etc. If you need to mention them, use "visual evidence" instead (e.g., "the video shows...").
- It's okay to double-check or question yourself during the thought process — reflect naturally as a human would.
- Refer to moments in time using broad expressions like: "at the beginning of the video", "around the middle", or "toward the end".

### ► User Prompt:

Video content: <Overall Video Caption>

Frame-level metadata: <Per-Frame Metadata>

Question: <Question>

Please think about this question as if you were a human pondering deeply. It's encouraged to include self-reflection or verification in the reasoning process.

<Corresponding Answer Format Template>

The <Corresponding Answer Format Template> in the prompt is dynamically selected from the following templates based on the question type:

Table 5: Answer Format Templates for Different Question Types

Question Type	Template
Multiple Choice	Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.
Numerical	Please provide the numerical value within the <answer> </answer> tags.
OCR	Please transcribe text from the image/video clearly and provide your text answer within the <answer> </answer> tags.
Free-form	Please provide your text answer within the <answer> </answer> tags.
Regression	Please provide the numerical value within the <answer> </answer> tags.

► **Cross-modal CoT Refinement:** We employ a cross-modal refinement process to ensure the CoT aligns with the video content, using the prompt  $P_{\text{cross}}$ :

## Cross-modal CoT Refinement

### ► System Prompt:

You are a multimodal reasoning expert. Your task is to revise hallucinations and errors in the chain-of-thought (CoT) based on the visual content of the provided video. Do not significantly alter the original CoT logic or content, and ensure the final conclusion remains the same.

### Your task:

1. Carefully examine the video, the question, and the CoT.
2. Identify only the reasoning steps that directly conflict with what is visually shown in the video:
  - Replace any text-based references with direct visual observations
  - Use visual phrasing such as "The video shows...", "I can see...", or "From the visual sequence...".
  - Replace specific timestamps with broader temporal phrases.
  - Do not rewrite steps that are already consistent with the visual content.
  - Only replace or correct parts that visually contradict what is shown.

### Cross-modal CoT Refinement — Continued

3. Ensure the rest of the CoT stays faithful to the original meaning.

► **User Prompt:**

```
<Input Video>  </Input Video>  
Question: <Question>  
Original CoT: <Original CoT>
```

**Output format:**

Strictly follow this format. Return only the revised CoT and no additional explanation:  
<think>[Revised CoT]</think>

## D.2 Image CoT Generation

In addition to the main video data in the VideoRFT-CoT-102K, we have also designed specific prompts specifically for image data, and used the similarly CoT generation process to generate CoT.

► **Structured Image Representation:** Due to the differences among different image datasets (the task focuses of the datasets are different), we designed different prompts for each image dataset in structured image representation phase.

### Common System Prompt for All Image Representations

You are a vision-language expert. Your task is to analyze the provided image and output a detailed, modular description in JSON format.

The following prompts are used for different image datasets, all sharing the above system prompt:

### General Image Representation (A-OKVQA, ShareGPT4V)

► **User Prompt:**

```
<Input Image>  </Input Image>  
You are given a natural photograph. Output only valid JSON with two keys:
```

1. **Overall Image Caption:** a 2-3 sentence narrative describing the scene, objects, actions, and context.
2. **Image Metadata:** an object containing:
  - **objects:** list of objects with id, type, color, size, and bbox
  - **text:** list of text elements with content and bbox
  - **scene\_context:** environment and activity
  - **relations:** list of subject-predicate-object relations

### CLEVR Image Representation

► **User Prompt:**

```
<Input Image>  </Input Image>  
You are given a synthetic 3D scene. Output only valid JSON with:
```

1. **Overall Image Caption:** a brief summary of number of objects and overall layout.
2. **Image Metadata:**
  - **objects:** list of objects with id, shape, color, size, material, and coordinates
  - **spatial\_relations:** list of spatial relations between objects

## STEM Image Representation (Geometry3K, UniGeo, AI2D)

### ► User Prompt:

<Input Image>  </Input Image>

You are given a line-drawing or diagram. Output **only** valid JSON with:

1. **Overall Image Caption:** a concise paragraph describing the diagram's purpose.
2. **Image Metadata:**
  - **primitives:** list of geometric primitives with type, label, and bbox
  - **annotations:** list of relations between primitives
  - **measurements:** list of measurements with primitive IDs and values

## OCR Image Representation (TextVQA, HME100k)

### ► User Prompt:

<Input Image>  </Input Image>

You are given an image containing printed or handwritten text. Output **only** valid JSON with:

1. **Overall Image Caption:** one sentence summarizing the text context.
2. **Image Metadata:**
  - **text\_items:** list of text elements with content, bbox, and style

## Science Image Representation (ScienceQA, PMC-VQA, ArxivQA)

### ► User Prompt:

<Input Image>  </Input Image>

You are given a multi-panel scientific figure. Output **only** valid JSON with:

1. **Overall Image Caption:** a paragraph overviewing the figure's subject.
2. **Image Metadata:**
  - **panels:** list of panels with elements and process arrows

## Chart Image Representation (DVQA, PlotQA, FigureQA)

### ► User Prompt:

<Input Image>  </Input Image>

You are given a chart image. Output **only** valid JSON with:

1. **Overall Image Caption:** a 1-2 sentence summary of chart type and key trend.
2. **Image Metadata:**
  - **chart\_type:** type of chart
  - **axes:** list of axis information
  - **series:** list of data series
  - **legend:** list of legend entries

## Math Image Representation (Multimath-300K, TQA)

### ► User Prompt:

<Input Image>  </Input Image>

You are given a textbook problem image. Output **only** valid JSON with:

## Math Image Representation — Continued

1. **Overall Image Caption:** a summary of the problem context.

2. **Image Metadata:**

- **equations:** list of equations with LaTeX and bbox
- **diagram\_parts:** list of diagram elements
- **givens:** list of given values

## Spatial Image Representation (OpenSpaces, Spacellava)

► **User Prompt:**

<Input Image>  </Input Image>

You are given an indoor scene or floorplan image. Output **only** valid JSON with:

1. **Overall Image Caption:** a 2-3 sentence narrative of the space.

2. **Image Metadata:**

- **rooms\_or\_sections:** list of room information
- **furniture:** list of furniture items with position and orientation
- **annotations:** list of structural elements

► **Cognitively Inspired CoT Generation:** We invoke DeepSeek-R1 to answer the question and extracts its step-by-step reasoning outputs with the prompt  $P_{\text{cog}}$  as the initial CoT.

## Cognitively Inspired CoT Generation for Images

► **System Prompt:**

You are an AI assistant helping a user answer questions about an image. When the user asks a question, you respond by imagining you are looking at the image with full attention, just like a human would. Your task is to reason visually and logically about the image content to answer the user's question.

Follow this multi-step reasoning approach:

1. **Simulate Visual Perception:** Imagine you are looking at the entire image carefully. Build a general understanding of what is shown.
2. **Understand the Question:** Reflect on what the user is asking. Think carefully about what kind of answer is needed (e.g., a fact, a reason, a comparison).
3. **Identify Relevant Elements:** Consider which parts of the image are most related to the question. Focus on those elements in your mental analysis.
4. **Visual Reasoning:** Describe what you “see” using natural visual language (e.g., “The image shows...”, “In the upper part of the image...”). Analyze and interpret the visual content to build your answer.
5. **Answer Thoughtfully:** Provide a clear and direct answer. Ensure your reasoning is consistent with the visual elements you described.

Guidelines for Responses:

- Don’t expose in the output that you are answering based on text information. Use statements imitating looking at an image to answer.
- Don’t directly refer to any textual metadata such as “captions”, “description”, “metadata”, etc. If you need to mention them, use “visual evidence” instead (e.g., “the image shows...”).
- It’s okay to double-check or question yourself during the thought process — reflect naturally as a human would.

## Cognitively Inspired CoT Generation for Images — Continued

- Refer to locations in the image using expressions like: “in the center”, “at the top”, “on the left side”, or “in the background”.

### ► User Prompt:

Image content: <Overall Image Caption>

Image metadata: <Image Metadata>

Question: <Question>

Please think about this question as if you were a human pondering deeply. It's encouraged to include self-reflection or verification in the reasoning process.

<Corresponding Answer Format Template>

► **Cross-modal CoT Refinement:** We employ a cross-modal refinement process to ensure the CoT aligns with the image content, using the prompt  $P_{\text{cross}}$ :

## Cross-modal CoT Refinement for Images

### ► System Prompt:

You are a multimodal reasoning expert. Your task is to revise hallucinations and errors in the chain-of-thought (CoT) based on the provided image. Do not significantly alter the original CoT logic or content, and ensure the final conclusion remains the same.

### Your task:

1. Carefully examine the image, the question, and the CoT.
2. Identify only the reasoning steps that directly conflict with what is shown in the image:
  - Replace all references to textual cues (such as "title", "bbox", "label") with direct visual observations from the image.
  - Use visual phrases such as “The image shows...”, “I can see...”, or “From the visual layout...” instead of text-based observations.
  - Use broader spatial descriptions like “on the left”, “in the background”, or “in the center”, instead of specific coordinates or labeled boxes.
  - Do not rewrite or paraphrase steps that are already visually accurate or consistent with the image.
  - Only replace or correct parts that visually contradict what is shown.
3. Ensure the rest of the CoT (which is either correct or visually consistent) stays faithful to the original meaning. Only revise incorrect or hallucinated steps based on visual evidence.

### ► User Prompt:

<Input Image>  </Input Image>

Question: <Question>

Original CoT: <Original CoT>

### Output format:

Strictly follow this format. Return only the revised CoT and no additional explanation:

<think>[Revised CoT]</think>