

VIDEOP2R: Video Understanding from Perception to Reasoning

Yifan Jiang^{1*}, Yueying Wang², Rui Zhao^{2†}, Toufiq Parag^{3*}, Zhimin Chen²
Zhenyu Liao², Jayakrishnan Unnikrishnan²
¹USC, ²Amazon, ³Keystone AI

¹yifjia@isi.edu, ²{yueyingw, zhaori, zhiminch, zyliao, jayunn}@amazon.com

³toufiq.parag@gmail.com

Abstract

Reinforcement fine-tuning (RFT), a two-stage framework consisting of supervised fine-tuning (SFT) and reinforcement learning (RL) has shown promising results on improving reasoning ability of large language models (LLMs). Yet extending RFT to large video language models (LVLMs) remains challenging. We propose VIDEOP2R, a novel process-aware video RFT framework that enhances video reasoning by modeling perception and reasoning as distinct processes. In the SFT stage, we develop a three-step pipeline to generate VIDEOP2R-CoT-162K, a high-quality, process-aware chain-of-thought (CoT) dataset for perception and reasoning. In the RL stage, we introduce a novel process-aware group relative policy optimization (PA-GRPO) algorithm that supplies separate rewards for perception and reasoning. Extensive experiments show that VIDEOP2R achieves state-of-the-art (SotA) performance on six out of seven video reasoning and understanding benchmarks. Ablation studies further confirm the effectiveness of our process-aware modeling and PA-GRPO and demonstrate that model’s perception output is information-sufficient for downstream reasoning.

1. Introduction

Understanding visual information in video has long been a core challenge in computer vision [15, 32, 48, 49]. Reinforcement fine-tuning (RFT) [34], a two-stage framework of supervised fine-tuning (SFT) followed by reinforcement learning (RL), has recently emerged as a powerful approach for enhancing the reasoning capabilities of large language models (LLMs) [1, 4, 21]. Within the RFT framework, the SFT stage leverages Chain-of-Thought (CoT) [59] annotations to warm up models with explicit guidance, while the RL stage further improves reasoning quality through reward-driven opti-

mization. Motivated by the success of RFT, recent studies [7, 14, 26, 56] have adapted RFT to video reasoning by designing diverse CoT generation pipelines and developing RL variants, with particular attention to Group Relative Policy Optimization (GRPO) [17] for its simple rule-based rewards without a critic model. These adaptations have demonstrated clear advantages over conventional SFT across multiple video benchmarks [14, 52].

However, exploration of adapting RFT from text to video remains preliminary, partly due to simply treating video as another alternative modality to text [50]. This simplicity assumption ignores the decomposability of visual reasoning, which consists of two different processes [43, 66]. The first one is *perception*, which extracts salient information from the visual input, and *reasoning*, which organizes the visual evidence and draws inferences. Each process can introduce error in distinct ways (e.g., missed visual evidence, faulty inference), and can impact video understanding [2, 22]. Yet existing video RFT frameworks are process-agnostic: they collapse perception and reasoning into a single procedure and assign a single final reward to the whole trajectory, which blurs credit assignment. As Fig. 1 (top right) shows, a perception error (e.g., turning back) induces the reasoning error, but without process awareness, the entire trajectory is evaluated as incorrect. Therefore, assigning a single reward to the entire process can prevent the model from effectively correcting mistakes that emerge in different processes.

The challenge to extend RFT to explicitly account for different processes in visual reasoning is two-fold. (i) *Lack of process-aware CoT data*: existing CoT annotations conflate perception and reasoning rather than explicitly disentangling the two; and (ii) *Coarse rewards*: training typically collapses feedback into a single final reward for the entire reasoning process [14, 52, 56], hindering credit assignment across processes. To address these challenges, we propose **VIDEOP2R**, a novel process-aware video RFT framework that models perception and reasoning as distinct processes to

*Work done while at Amazon. †Corresponding author.

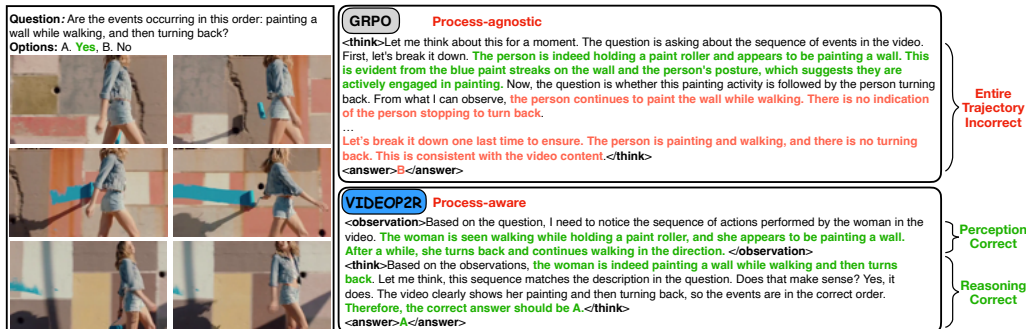


Figure 1. Comparison between GRPO-based video RFT framework (process-agnostic) and VIDEOP2R (process-aware).

enhance video reasoning. Same as conventional RFT, our framework consists of two stages of training. **In the SFT stage**, we construct a three-step CoT generation pipeline that produces high-quality perception and reasoning traces. Given visual question-answer (VQA) samples, our pipeline first generates visual perception and reasoning traces in order; the perceptions are then fed to a reasoning-capable LLM [5] to verify whether they contain sufficient visual evidence to reach the correct answer. Running this pipeline on 260K VQA pairs [14] yields **VIDEOP2R-CoT-162K** after filtering low-quality samples. We use this dataset in SFT to warm up the model, encouraging process separation during inference and providing a strong initialization for subsequent RL. **In the RL stage**, we propose **PA-GRPO**, a process-aware variant of GRPO. Unlike GRPO, which assigns a single reward to the entire trajectory, PA-GRPO supplies two separate rewards specific for perception and reasoning and assigns them to the corresponding output segments: 1) an LLM-judged [16] perception reward, which evaluates whether the model’s perception captures the necessary information from the video, and 2) a rule-based reasoning reward verifying the correctness of the final answer. As illustrated in Fig. 1, after two-stage training, VIDEOP2R enables LVLMs to conduct process-aware inference with calibrated perception and reasoning.

We conduct comprehensive experiments on seven widely used video understanding and reasoning benchmarks [15, 20, 25, 32, 40, 62, 72], comparing VIDEOP2R with representative process-agnostic video RFT baselines (e.g., Video-R1 [14] and VideoRFT [52]). Results show that VIDEOP2R achieves SotA on six out of seven benchmarks, with robust gains of 1.9%–9.1% average accuracy over base models across benchmarks. Ablation studies further validate the effectiveness of the process-aware modeling and the PA-GRPO. In addition, we provide fine-grained analysis of VIDEOP2R’s perception and PA-GRPO’s improvements over GRPO to support future process-aware research in the video domain. Our key contributions are: (i) A novel process-

aware video RFT framework, **VIDEOP2R**, that models perception and reasoning separately to enhance video reasoning. (ii) A process-aware RL algorithm based on GRPO, **PA-GRPO**, that provides separate rewards for perception and reasoning, improving credit assignment in RL. (iii) An automatic three-step CoT generation pipeline that produces perception and reasoning annotations, yielding **VIDEOP2R-CoT-162K** for warm start in SFT. (iv) Comprehensive evaluation confirms VIDEOP2R’s SotA performance, with ablations verifying the effectiveness of process-aware modeling and PA-GRPO. Our project page is available at <https://videop2r.github.io/videop2r/>.

2. Related work

Video Perception and Reasoning in LVLMs. Video reasoning poses coupled spatial–temporal challenges beyond text/image only settings [15, 32, 48], demanding coordinated perception and reasoning [43, 66]. Early approaches handle this by introducing modular perception preprocessors (e.g., frame captioners [53, 57, 68] or spatio-temporal scene graphs [27, 47, 61]) to aid downstream reasoning. However, because these preprocessors are typically frozen, they bottleneck the pipeline and cannot be improved to curb error propagation or information loss [50]. Recent methods organize perception–reasoning into pre-defined stages, either by attributes (frames/objects/actions) [13, 18, 37, 41, 55, 58] or steps (planning/grounding) [33, 45]. However, rigid designs often limit generalization across scenarios [52], and mixing perception and reasoning at each stage makes it vulnerable to perception errors [2, 22]. In contrast, VIDEOP2R separates perception and reasoning, aligning with the two-process modeling of video reasoning [43] for better generalization, while preserving an end-to-end training pipeline for continual refinement.

Reinforcement Learning for LVLMs. Building on GRPO, recent work adapts RL/RFT to video understanding and reasoning [7, 14, 26, 36, 39, 52, 56, 71]. Time-R1 uses timestamp-aware and template re-

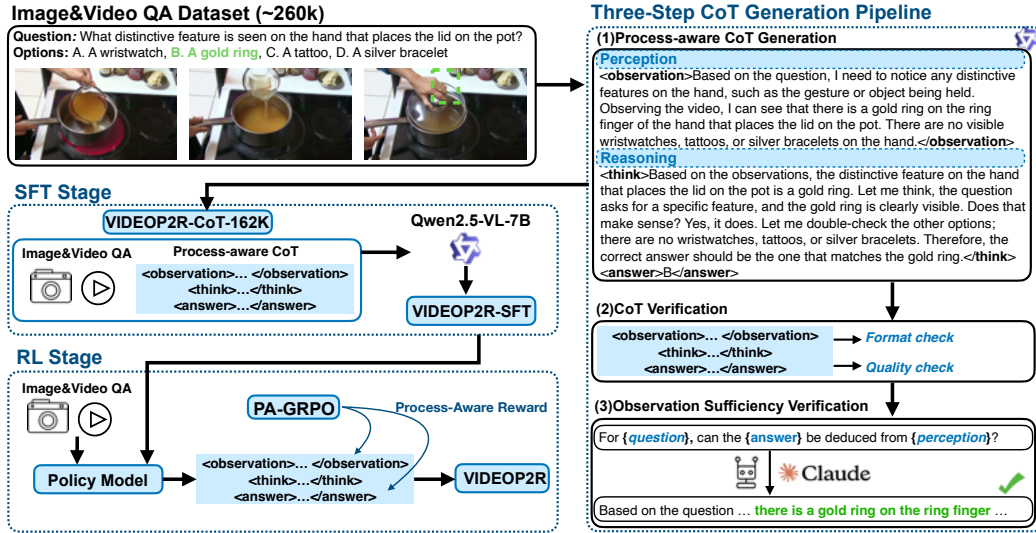


Figure 2. Illustration of overall VIDEOP2R RFT framework (left) and the three-step CoT generation pipeline (right).

wards [56]; Video-R1 and STAR-R1 reward sensitivity to correct temporal order [14, 28]; Videochat-R1 and VersaVid-R1 adopt task-specific rewards [7, 26]; VideoRFT adds a stage-aware semantic reward [52]. However, most prior efforts model video reasoning as a single trajectory and apply a single final reward over the entire output sequence, with no process-aware distinction between perception and reasoning. This also holds for recent image-domain studies on perception–reasoning separation [29, 60, 64]. VIDEOP2R instead uses PA-GRPO to assign separate perception and reasoning rewards to their token segments, providing more clear signals and more precise error attribution.

3. VIDEOP2R RFT Framework

In this section, we introduce the overall design of the VIDEOP2R RFT framework (Fig. 2 left), which follows the standard RFT setup [34] with a specific focus on modeling video reasoning into perception and reasoning: (1) **SFT stage**, we use a three-step CoT generation pipeline to construct a process-aware CoT dataset, **VIDEOP2R-CoT-162K**. We train the base model on this dataset to enhance its perception and reasoning capabilities while warming up the model for the RL stage. (2) **RL stage**, we propose a process-aware reinforcement learning scheme, **PA-GRPO**, which refines the model’s reasoning by providing separate rewards for perception and reasoning, enabling the model to move beyond supervised learning boundaries [38].

3.1. Process-aware CoT Dataset

To address the challenge of lacking process-aware CoT dataset, we develop a strategy to curate CoT data at scale, which is then used for fine-tuning LLM

in SFT. Particularly, we first standardize a process-aware CoT template that explicitly disentangles perception from reasoning in different segments as follows:

```
<observation>...</observation>
<think>...</think><answer>...</answer>
```

`(observation)` segment represents the perception process, where the model extracts relevant visual evidence based on the question. `(think)` and `(answer)` segment captures the reasoning process, where the model reasons (`(think)`) over the extracted visual evidence and states the final answer (`(answer)`). All generated CoTs follow this template, and the model is trained to adhere to it at inference.

3.1.1. Three-Step CoT Generation Pipeline

Building upon the proposed template, we design a three-step pipeline (Fig. 2 right) to generate process-aware CoT data containing both perception and reasoning traces for a diverse set of VQA tasks. The overall workflow comprises three major steps (Details in the Supplementary). **1) Process-aware CoT Generation.** For each VQA sample, we use Qwen2.5-VL-72B-Instruct [4] to generate an initial CoT trace for both perception and reasoning in corresponding segments. **2) CoT Verification.** To ensure consistency and correctness, we evaluate the final answer of each generated response with task-specific metrics (e.g., exact word match and word error rate for generation tasks), discarding samples that yield low-quality answers or deviate from the expected CoT template. **3) Observation Sufficiency Verification.** We further filter generated data using a cross-modal validation strategy to validate the `(observation)` segment in isolation from raw visual inputs [52]. Specifically, for each sample, we provide only the `(observation)` segment, along with the corresponding question and an-

swer, to Claude 3.7 Sonnet [3], which assesses whether the visual evidence in `<observation>` are adequate to support the final reasoning process.

3.1.2. Data Construction and Statistics

To ensure reproducibility and fair comparison, we apply this pipeline to a public and representative image- and video-based QA dataset [14], which includes multiple question types such as multiple-choice, numerical QA, Optical Character Recognition (OCR), free-form QA, and regression. Applying our pipeline on 260K VQA data [14] produces 162K high-quality process-aware CoT data with perception and reasoning traces, termed **VIDEOP2R-CoT-162K**. We provide detailed analyses of VIDEOP2R-CoT-162K (e.g., embedding visualization and word-frequency statistics) in the Supplementary, which show that our annotations inherently separate perception from reasoning.

3.2. Process-Aware Reinforcement Learning

After the SFT stage, we further refine the model through reinforcement learning. Building upon GRPO [17], we introduce a process-aware variant, **PA-GRPO** (Fig. 3), which provides separate reward signals for perception and reasoning processes to encourage more structured and efficient policy optimization [30]. This section first revisits the standard GRPO framework and then presents our process-aware extension designed to align reward signals with perception–reasoning separation.

3.2.1. Group Relative Policy Optimization (GRPO)

GRPO [17] extends Proximal Policy Optimization [42] by removing the dependency on a learned critic model and directly comparing responses within groups. Given a question q , the policy model π_θ samples G candidate responses $o = \{o_1, o_2, \dots, o_G\}$ as a group, each assigned a rule-based reward r_i . Rewards are then normalized within the group to yield the relative advantage:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}. \quad (1)$$

where A_i represents the relative advantage of all tokens in the i -th response within the group. With the relative advantage computed, the GRPO overall optimization objective is formulated as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}} \left[\frac{1}{G} \sum_{i=1}^G \min(\rho_i A_i, \text{clip}(\rho_i, 1 - \epsilon, 1 + \epsilon) A_i) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right]. \quad (2)$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the likelihood ratio between updated policy π_θ and old policy $\pi_{\theta_{\text{old}}}$, and π_{ref} is a fixed reference model (e.g., a frozen copy of policy model) after SFT, providing KL regularization weighted β . This

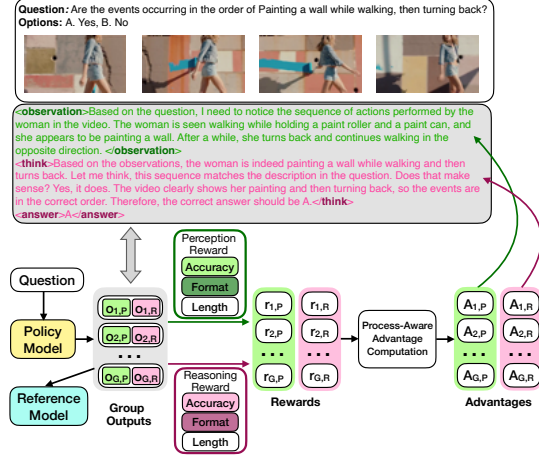


Figure 3. The illustration of the PA-GRPO algorithm.

formulation constrains large policy deviations while promoting high-reward samples, ensuring stable optimization during reinforcement learning.

3.2.2. Process-Aware Group Relative Policy Optimization (PA-GRPO)

While GRPO performs well on textual reasoning, its single scalar reward provides a limited training signal for the two-process based visual reasoning [43, 66]. To address this, we propose **PA-GRPO**, which introduces separate **perception** and **reasoning** rewards for tokens in each process, enabling fine-grained credit assignment during reinforcement learning, as illustrated in Fig. 3. Formally, for each question q , the sampled response from the policy model π_θ is represented as $o = \{(o_{1,P}, o_{1,R}), (o_{2,P}, o_{2,R}), \dots, (o_{G,P}, o_{G,R})\}$, where $o_{i,P}$ denotes perception process tokens within the `<observation>` segment, and $o_{i,R}$ denotes reasoning process tokens within the `<think>` and `<answer>` segments. For tokens within each process, **PA-GRPO** supplies separate accuracy rewards to provide reliable supervision. We further introduce length and format rewards to encourage well-structured, concise outputs, following prior RFT frameworks [14, 17, 26, 52]. We demonstrate the accuracy reward design, and our configurations for the format and length rewards below.

Perception Accuracy Reward ($R_{\text{acc,P}}$). The perception accuracy reward evaluates whether the model correctly perceives visual information from video input. We adopt an LLM-as-Judge evaluation [16] in a text-only setting, a procedure shown to be reliable in various scenarios [9, 12]. Concretely, we provide only the `<observation>` segment, along with the corresponding question and answer, to Claude 3.7 Sonnet [3], which judges whether the `<observation>` segment is sufficient to supplant the correct answer. Formally,

$$R_{\text{acc,P}} = 1 \text{ (judged sufficient); } 0 \text{ otherwise.} \quad (3)$$

Reasoning Accuracy Reward ($R_{\text{acc,R}}$). The reasoning accuracy reward evaluates whether the model produces accurate reasoning outcomes. We apply task-specific evaluation metrics to accommodate different question types, including exact word match for categorical tasks, ROUGE-based similarity for open-ended generation, and error-based scores for numerical or regression problems. Formally,

$$R_{\text{acc,R}} = \text{Acc}_t(o_{i,\text{R}}, y_{\text{true}}), \quad (4)$$

where $\text{Acc}_t(\cdot) \in [0, 1]$ denotes the task-specific accuracy metric for task type t .

Format Reward (R_{form}) and Length Reward (R_{len}). To ensure clear reward assignment, the format and the length reward in PA-GRPO are also provided separately for each process ($o_{i,\text{P}}$ and $o_{i,\text{R}}$). We use the format reward (R_{form}) to encourage adherence to the process-aware inference template. Specifically, perception process tokens ($o_{i,\text{P}}$) must appear within the `<observation>`, the reasoning process tokens ($o_{i,\text{R}}$) must present the reasoning trace under `<think>`, and the final answer must be provided within `<answer>`. We verify compliance using regular expression matching and assign a binary reward ($\{0,1\}$) accordingly. The length reward (R_{len}) is included to favor concise yet informative responses while avoiding overthinking. The reward is assigned only if both accuracy and format rewards are non-zero and the model’s response in each segment ($o_{i,\text{P}}$ and $o_{i,\text{R}}$) falls within the target length ($[l_{\text{min}}, l_{\text{max}}]$). In line with prior RFT work [14, 52], we fix $R_{\text{len}} = 0.2$, $l_{\text{min}} = 128$ and $l_{\text{max}} = 320$ for $o_{i,\text{P}}$, $l_{\text{min}} = 320$ and $l_{\text{max}} = 512$ for $o_{i,\text{R}}$.

3.2.3. Process-Aware Reward Assignment

The overall reward of perception tokens ($o_{i,\text{P}}$) and reasoning tokens ($o_{i,\text{R}}$) are defined as

$$R_{i,k} = R_{\text{acc},k} + R_{\text{form},k} + R_{\text{len},k}, \quad k \in \{\text{P}, \text{R}\}. \quad (5)$$

Unlike Eq. (1), which normalizes all rewards within a single group, we split rewards into separate groups (perception process vs. reasoning process) and normalize each to get process-aware advantage, since their scales and distributions are not directly comparable [67]:

$$A_{i,k} = \frac{R_{i,k} - \text{mean}(\{R_{j,k}\}_{j=1}^G)}{\text{std}(\{R_{j,k}\}_{j=1}^G)}, \quad k \in \{\text{P}, \text{R}\}. \quad (6)$$

We assign each process-aware advantage only to its corresponding tokens (e.g., the perception advantage $A_{i,\text{P}}$ is applied to $o_{i,\text{P}}$). The overall optimization objective of PA-GRPO is then formulated as:

$$\mathcal{J}_{\text{PA-GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}} \left[\frac{1}{G} \sum_{i=1}^G \sum_{k \in \{\text{P}, \text{R}\}} \min(\rho_{i,k} A_{i,k}, \text{clip}(\rho_{i,k}, 1-\epsilon, 1+\epsilon) A_{i,k}) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \quad (7)$$

where $\rho_{i,k} = \frac{\pi_\theta(o_{i,k} | q)}{\pi_{\theta_{\text{old}}}(o_{i,k} | q)}$ denotes the likelihood ratio for perception ($k=\text{P}$) or reasoning ($k=\text{R}$), and $A_{i,k}$ is the process-aware advantage.

4. Experiment Setup

Two-stage Training. We adopt Qwen2.5-VL-7B-Instruct [4] as the base LVLm in our training pipeline. Following the same training setups used in prior video-RFT studies [14, 52], we perform one epoch of SFT on VIDEO2R-CoT-162K, followed by 1K RL updates over original 260K visual QA data using PA-GRPO. The model obtained after the SFT stage is referred to as **VIDEO2R-SFT**, and the final model after the RL stage is denoted as **VIDEO2R**.

Benchmarks. Following prior works [15, 52], we evaluate our approach on seven benchmarks, including four video reasoning datasets (VSI-Bench [62], VideoMMU [20], MMVU [72], and VCR-Bench [40]) and three video understanding datasets (MVBench [25], TempCompass [32], and VideoMME [15]). These benchmarks jointly cover spatial reasoning, knowledge-intensive QA, temporal logic, and general video understanding. We follow the official evaluation protocols of each benchmark.

Baselines. (1) **RFT on Qwen2.5-VL-7B.** We compare against recent video RFT approaches built upon Qwen2.5-VL-7B using GRPO or its variants, including Video-R1 [14], Time-R1 [56], VideoRFT [52], VideoChat-R1 [26], and VersaVid-R1 [7] (2) **Open-Source Models.** We further include Qwen2.5-VL-7B along with other 7B-scale models for a comprehensive evaluation: LLaVA-OneVision [24], LongVA [69], Video-UTR [63], and VideoLLaMA2 [8]. We follow the prompt templates in each baseline’s official publication. We provide all setup details in the Supplementary.

5. Results

Our experiments focus on addressing five research questions from Section 5.1 to 5.5: (1) How does VIDEO2R perform across various video understanding benchmarks? (2) What is the contribution of each component in VIDEO2R? (3) Does the perception representations of VIDEO2R effectively support downstream reasoning? (4) Does the process-aware reward design of PA-GRPO improve RL efficiency and reliability? (5) What are the success and failure mode of VIDEO2R?

5.1. Main Results

The main evaluation results of VIDEO2R and other baselines are shown in Tab. 1. Compared with prior video RFT approaches, VIDEO2R achieves highly competitive performance across seven benchmarks, set-

Table 1. Performance comparison on video reasoning and understanding benchmarks. Best/second-best result of each column is in **bold/underline**. Missing entries indicate unreported results (all numbers unit in %).

Model	Video Reasoning				Video Understanding			Avg
	VSI.	VideoMMMU	MMVU	VCR.	MV.	TempCom.	VideoMME	
Open-Source 7B Models								
LLaVA-OneVision-7B	32.4	33.8	49.2	–	56.7	–	58.2	–
LongVA-7B	29.2	23.9	–	–	–	56.9	52.6	–
Video-UTR-7B	–	–	–	–	58.8	59.7	52.6	–
VideoLLaMA2-7B	–	–	44.8	–	54.6	–	47.9	–
Qwen2.5-VL-7B	30.1	48.1	60.0	44.3	59.0	72.6	56.6	52.9
RFT on Qwen2.5-VL-7B								
Video-R1	<u>35.8</u>	52.3	63.8	49.0	63.9	73.2	59.3	56.8
VideoChat-R1	33.9	<u>54.0</u>	63.0	49.0	<u>67.9</u>	72.5	57.7	56.9
Time-R1	29.0	51.0	62.9	49.6	63.1	73.7	59.3	55.5
VersaVid-R1	33.7	51.9	64.3	49.8	62.9	<u>74.0</u>	58.8	56.5
VideoRFT	36.8	51.1	68.5	49.6	62.1	73.7	<u>59.8</u>	<u>57.4</u>
VIDEOP2R (Ours)	36.8	55.0	<u>65.4</u>	51.0	68.1	74.5	60.0	58.7

ting **SotA results on six of them** and ranking second on the remaining one. In contrast to previous RFT methods that often bring improvement on specific datasets (e.g., Video-R1 ranks second on VSI-Bench, while VideoRFT is SotA on MMVU but last on MVBench), VIDEOP2R delivers **consistent gains across all benchmarks**, surpassing the previous SotA by 1.3% in average accuracy. This consistency underscores the effectiveness and generalizability of modeling visual reasoning through distinct perception and reasoning processes. Compared to the base model Qwen2.5-VL, VIDEOP2R exhibits clear and steady improvements, with accuracy gains ranging from **1.9% to 9.1%** across benchmarks. More broadly, a performance gap exists between models trained with RFT and those trained only with SFT/Instruction Tuning [31] (i.e., open-source models), highlighting the superiority of RFT for expanding capability boundaries. We further analyze VIDEOP2R’s performance drop on MMVU (Sec. 5.5) and attribute it to the lack of domain-specific knowledge (e.g., chemistry) in our training data.

5.2. Ablation Study

The success of VIDEOP2R underscores the importance of decomposing visual reasoning into distinct process stages. To further analyze the contribution of each process-aware component, we perform an ablation study on the **two-stage training** in the RFT framework, **process-aware modeling**, and **reward design** (Tab. 2). **Two-stage Training:** SFT-only and RL-only improves the baseline by 2.7% and 3.1% respectively. But combining both yield a more significant 5.8% improvement, suggesting that single-stage training is insufficient, whereas using both stages can further extend the model’s capability.

Process-aware Modeling: We evaluate process-aware

modeling against a process-agnostic counterpart in both SFT and RL. In SFT, the process-aware variant (same as VIDEOP2R-SFT) yields an average accuracy gain of 2.1% over the process-agnostic variant, which uses reasoning-only segments. In RL, the agnostic variant follows GRPO—assigning a single reasoning reward across all output tokens, while the process-aware variant (same as VIDEOP2R) again leads on six benchmarks with 2.3% on average. Under identical visual inputs, adding perception annotations in SFT and using process-aware credit assignment in RL provide clearer training signals, improving video understanding and reasoning.

Reward design: We examine the PA-GRPO reward function by ablating each reward component and the separation reward assignment. Removing any component causes a notable drop, sometimes even below the SFT baseline (e.g., removing the perception reward yields worse performance than SFT on VideoMME). This indicates that all components are necessary and that their joint design in PA-GRPO enables fine-grained credit assignment in RL. When we remove separation and assign both perception and reasoning rewards to all output tokens, the results remain competitive but still lag behind PA-GRPO, indicating that separation reward assignment provides clearer and more effective training signals. Notably, removing R_{len} improves performance on VSI-Bench. Further analysis (Supplementary) indicates that its questions often require long, fine-grained descriptions to ensure sufficient perception, where the length reward becomes counter-productive. We propose to have a dynamic length reward [51] in future work. Finally, we further ablate different judge models in the Supplementary for perception accuracy reward assignment, and find that VIDEOP2R remains robust.

Table 2. Ablation studies of VIDEOP2R on two-stage training, process-aware modeling and reward design (all numbers unit in %).

Ablation Factor Model Variant	Video Reasoning				Video Understanding			Avg.
	VSI.	VideoMMMU	MMVU	VCR.	MV.	TempCom.	VideoMME	
<i>Two-stage Training</i>								
- VIDEOP2R (Ours)	36.8	55.0	65.4	51.0	68.1	74.5	60.0	58.7 ^{+5.8}
- SFT-only (VIDEOP2R-SFT)	35.2	53.7	61.6	46.9	62.3	72.4	57.2	55.6 ^{+2.7}
- RL-only	35.8	54.6	64.6	46.3	60.8	73.8	55.9	56.0 ^{+3.1}
<i>Process-aware Modeling</i>								
- VIDEOP2R (Ours)	36.8	55.0	65.4	51.0	68.1	74.5	60.0	58.7 ^{+5.8}
- process-agnostic RL (GRPO)	37.4	53.6	62.8	48.3	63.8	73.3	55.4	56.4 ^{+3.5}
- process-aware SFT (no RL)	35.2	53.7	61.6	46.9	62.3	72.4	57.2	55.6 ^{+2.7}
- process-agnostic SFT (no RL)	34.3	48.9	61.6	47.3	59.0	69.7	54.0	53.5 ^{+0.6}
<i>Reward Design</i>								
- VIDEOP2R (Ours)	36.8	55.0	65.4	51.0	68.1	74.5	60.0	58.7 ^{+5.8}
- without R_R	36.0	51.6	60.3	46.8	62.1	72.5	57.9	55.3 ^{+2.4}
- without R_P	37.4	53.6	62.8	48.3	63.8	73.3	55.4	56.4 ^{+3.5}
- without R_L	40.0	52.7	63.2	48.4	65.5	73.9	60.0	57.7 ^{+4.8}
- without <i>separation</i>	37.1	53.2	64.9	48.8	65.0	73.2	59.7	57.4 ^{+4.5}
Baseline: Qwen2.5-VL-7B	30.1	48.1	60.0	44.3	59.0	72.6	56.6	52.9

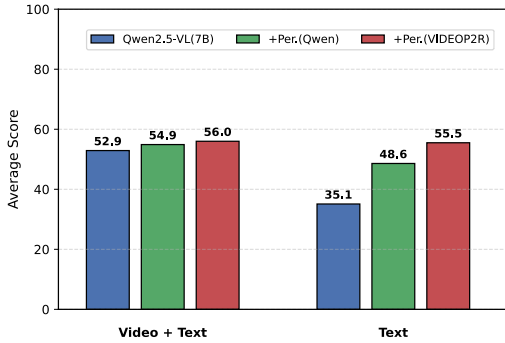


Figure 4. Effect of perception on downstream reasoning

5.3. Effectiveness of Perception Representations

The ablation study highlights the effectiveness of process-aware modeling in VIDEOP2R. In this section, we further examine whether the perception outputs produced by VIDEOP2R can enhance the reasoning ability of generic LLMs (Examination details in the Supplementary). Specifically, we compare Qwen2.5-VL-7B’s zero-shot performance on (i) text-only questions, (ii) text plus video inputs. For each scenario, we further augment question text with a perception segment generated either by VIDEOP2R or by pre-trained Qwen. The results, as shown in Fig. 4, yield two key insights. **First**, when Qwen2.5-VL-7B is provided only with text questions augmented by VIDEOP2R’s perceptions segments, its performance (55.5%) even surpasses that under raw video input (52.9%), indicating that the textual perceptions generated by VIDEOP2R capture semantically rich perceptual information that directly supports reasoning. **Second**, in both text and video modality, when we augment the prompt with a perception seg-

ment, using VIDEOP2R’s perceptions consistently outperforms using Qwen’s own perceptions, indicating that VIDEOP2R’s perceptions provide more accurate and useful evidence for downstream reasoning.

5.4. Advantages of PA-GRPO over GRPO

Figure 6(a-b) illustrates the training dynamics of VIDEOP2R. Both perception and reasoning accuracy rewards exhibit an increasing trend, which, combined with the quantitative gains in Tab. 2, confirms that PA-GRPO is more effective than standard GRPO in improving the model’s perception and reasoning traces. In the following, we further analyze the underlying reasons for PA-GRPO’s advantages over standard GRPO. **(1) Training Efficiency.** In advantage-based policy gradient methods [17, 42, 54], including GRPO, when all sampled responses (o_1, o_2, \dots, o_G) for a given prompt receive nearly identical rewards, this leads to *advantage collapse* [70], where the advantages shrink toward zero, leaving little to no effective learning signal and causing updates to stagnate [65]. PA-GRPO mitigates this by decomposing the overall reward into perception and reasoning components, so even when reasoning rewards saturate, perception rewards can still provide non-zero gradients. We visualize the number of samples in a batch with *advantage collapse* for PA-GRPO and GRPO during the RL stage in Fig. 6(c). Compared with GRPO, PA-GRPO consistently exhibits fewer advantage collapse samples, indicating better utilization of training samples and improved training efficiency.

(2) Mitigating Think-Answer Mismatch. Reasoning-augmented models often exhibit *Think-Answer Mismatch* [44], where generated reasoning traces diverge from the actual decision process yet still produce cor-

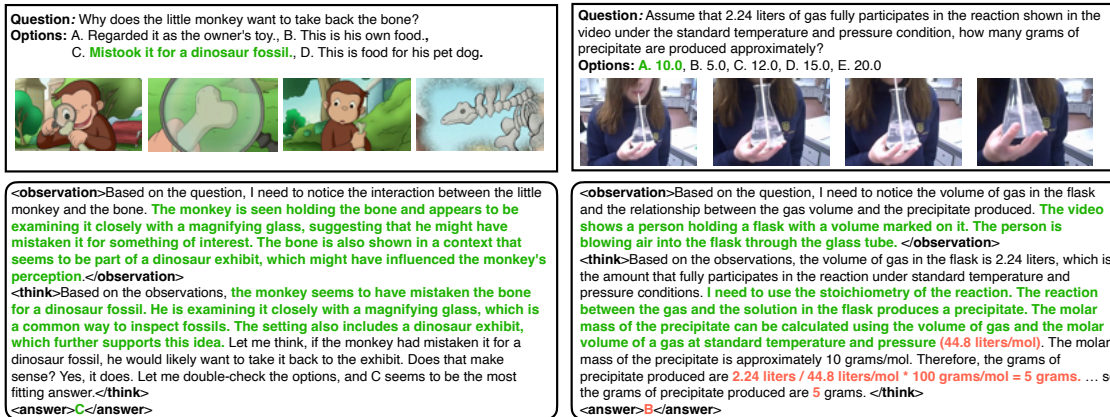


Figure 5. Success (Left) and Failure (Right) case of VIDEOP2R. Correct statement and incorrect statement are colored.

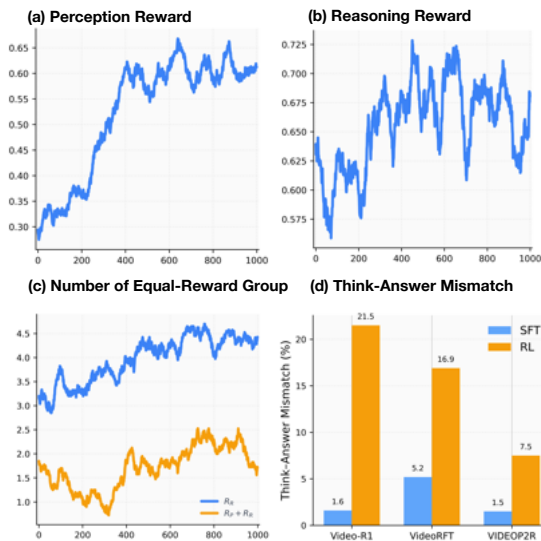


Figure 6. Training Dynamics and Think-Answer Mismatch Analysis of VIDEOP2R. Details in Section 5.4.

rect answers (e.g., “the man is on the right side, thus the answer is B. left”). In GRPO, such inconsistencies can lead to reward hacking [10, 46], as a single final reward can reinforce unfaithful reasoning traces that coincidentally yield correct outcomes. To quantify this issue, we perform an alignment check (Details in Supplementary) using Claude 3.7 Sonnet to extract answers from the `<think>` segments and compare them with the final output answer. Figure 6(d) reports mismatch rates across VIDEOP2R and two single-reward trained models: while all SFT models maintain stable reasoning consistency ($\leq 5\%$), both Video-R1 and VideoRFT’s RL models show significant degradation ($\geq 16\%$). In contrast, VIDEOP2R shows notably lower mismatch, indicating that PA-GRPO’s process-aware rewards, which separately encourage faithful perception traces and correct final answers, effectively mitigate Think–Answer

Mismatch and promote more reliable reasoning.

5.5. Qualitative Results of VIDEOP2R

We present one success and one failure case of VIDEOP2R in Fig. 5 to illustrate both its strengths and areas of improvement (More examples in the Supplementary). The left example shows an *Aha Moment* [17], where VIDEOP2R performs process-aware inference by accurately describing visual cues, such as the monkey’s actions and its imagination, and reasoning over them to reach the correct answer. In contrast, the right example depicts a failure case: although the model correctly identifies relevant visual details about the person and her behavior, it produces an incorrect conclusion due to missing domain-specific knowledge (e.g., the molar volume of a gas should be 22.4). Overall, while VIDEOP2R exhibits strong capabilities in general video understanding and reasoning, its performance can be further improved by injecting factual and domain-specific knowledge.

6. Conclusion

In this work, we introduced VIDEOP2R, a process-aware RFT framework that models perception and reasoning as distinct processes for video understanding. Through a three-step CoT generation pipeline, we constructed VIDEOP2R-CoT-162K, a large-scale process-aware dataset enabling fine-grained supervision in the SFT stage. In the RL stage, we proposed PA-GRPO, a process-aware extension of GRPO that provides separate rewards for perception and reasoning to improve credit assignment. Experiments across seven benchmarks demonstrate SotA performance and strong generalization, while ablations verify the effectiveness of process-aware modeling and PA-GRPO.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Kian Ahrabian, Zhivar Sourati, Kexuan Sun, Jiarui Zhang, Yifan Jiang, Fred Morstatter, and Jay Pujara. The curious case of nonverbal abstract reasoning with multi-modal large language models. *arXiv preprint arXiv:2401.12117*, 2024. 1, 2
- [3] Anthropic. Claude 3, 2024. 4
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3, 5
- [5] Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houlliston, et al. Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*, 2025. 2
- [6] Mustafa Chasmai, Gauri Jagatap, Gouthaman KV, Grant Van Horn, Subhransu Maji, and Andrea Fanelli. Moment sampling in video llms for long-form video qa. *arXiv preprint arXiv:2507.00033*, 2025. 6
- [7] Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Bohan Zeng, Yang Shi, Sihang Yang, Pengfei Wan, Qiang Liu, Liang Wang, and Tieniu Tan. Versavid-r1: A versatile video understanding and reasoning model from question answering to captioning tasks. *arXiv preprint arXiv:2506.09079*, 2025. 1, 2, 3, 5
- [8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 5
- [9] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolaos Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024. 4
- [10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 8
- [11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 4
- [12] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024. 4
- [13] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: step-by-step video reasoning from perception to cognition. In *Proceedings of the 41st International Conference on Machine Learning*, pages 13109–13125, 2024. 2
- [14] 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, 4, 5, 8
- [15] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 1, 2, 5, 3, 7
- [16] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024. 2, 4
- [17] Daya Guo and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 4, 7, 8
- [18] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videospresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191, 2025. 2
- [19] Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. Self-adaptive sampling for accurate video question answering on image text models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2522–2534, 2024. 6
- [20] 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. 2, 5, 3, 7
- [21] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Alexander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 1
- [22] Yifan Jiang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, Jay Pujara, et al. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592, 2024. 1, 2
- [23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 3
- [24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei

- Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5
- [25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2, 5, 3
- [26] 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. 1, 2, 3, 4, 5
- [27] Yiming Li, Xiaoshan Yang, Bing-Kun Bao, and Changsheng Xu. Graph prompts: Adapting video graph for video question answering. 2025. 2
- [28] Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*, 2025. 3
- [29] Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025. 3
- [30] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023. 4
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 6
- [32] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8731–8772, 2024. 1, 2, 5, 3
- [33] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025. 2
- [34] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024. 1, 3
- [35] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 2
- [36] Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025. 2
- [37] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024. 2
- [38] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3
- [39] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*, 2025. 2
- [40] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025. 2, 5, 3
- [41] Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. Step: Enhancing video-llms’ compositional reasoning by spatio-temporal graph-guided self-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3284–3294, 2025. 2
- [42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4, 7
- [43] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011, 2020. 1, 2, 4
- [44] Si Shen, Peijun Shen, Wenhua Zhao, and Danhao Zhu. Mitigating think-answer mismatch in llm reasoning through noise-aware advantage reweighting. *arXiv preprint arXiv:2508.05928*, 2025. 7
- [45] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8523–8533, 2025. 2
- [46] Joar Skalse, Nikolaus Howe, Dmitrii Krashennikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022. 8
- [47] Zihan Song, Xin Wang, Zi Qian, Hong Chen, Longtao Huang, Hui Xue, and Wenwu Zhu. Modularized self-reflected video reasoner for multimodal llm with application to video question answering. In *Forty-second International Conference on Machine Learning*. 2
- [48] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007. 1, 2
- [49] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large lan-

- guage models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1
- [50] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1, 2
- [51] Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*, 2025. 6, 7
- [52] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorf: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025. 1, 2, 3, 4, 5
- [53] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 2
- [54] Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *Uncertainty in artificial intelligence*, pages 113–122. PMLR, 2020. 7
- [55] Yueqian Wang, Yuxuan Wang, Kai Chen, and Dongyan Zhao. Stair: spatial-temporal reasoning with auditable intermediate results for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19215–19223, 2024. 2
- [56] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025. 1, 2, 3, 5
- [57] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497, 2022. 2
- [58] Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. *arXiv preprint arXiv:2506.06097*, 2025. 2
- [59] 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. 1
- [60] Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025. 3
- [61] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022. 2
- [62] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 2, 5, 3, 7
- [63] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025. 5
- [64] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025. 3
- [65] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 7
- [66] Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. Perception matters: Detecting perception failures of vqa models using metamorphic testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16908–16917, 2021. 1, 2, 4
- [67] Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821*, 2025. 5
- [68] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023. 2
- [69] 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. 5
- [70] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*, 2025. 7
- [71] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller llms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025. 2
- [72] 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, 5, 3