Reinforcing Structured Chain-of-Thought for Video Understanding

Peiyao Wang[1*] Haotian Xu[2] Noranart Vesdapunt[2] Rui Hou[2] Jingyi Zhang[2]
Haibin Ling[1†] Oleksandr Obiednikov[2] Ning Zhou[2] Kah Kuen Fu[2†]

[1]Stony Brook University [2]Amazon

Abstract

*Multi-modal Large Language Models (MLLMs) show promise in video understanding. However, their reasoning often suffers from thinking drift and weak temporal comprehension, even when enhanced by Reinforcement Learning (RL) techniques like Group Relative Policy Optimization (GRPO). Moreover, existing RL methods usually depend on Supervised Fine-Tuning (SFT), which requires costly Chain-of-Thought (CoT) annotation and multi-stage training, and enforces fixed reasoning paths, limiting MLLMs' ability to generalize and potentially inducing bias. To overcome these limitations, we introduce **S**ummary-**D**riven **R**einforcement **L**earning (**SDRL**), a novel single-stage RL framework that obviates the need for SFT by utilizing a Structured CoT format: Summarize → Think → Answer. SDRL introduces two self-supervised mechanisms integrated into the GRPO objective: 1) Consistency of Vision Knowledge (CVK) enforces factual grounding by reducing KL divergence among generated summaries; and 2) Dynamic Variety of Reasoning (DVR) promotes exploration by dynamically modulating thinking diversity based on group accuracy. This novel integration effectively balances alignment and exploration, supervising both the final answer and the reasoning process. Our method achieves state-of-the-art performance on seven public VideoQA datasets.*

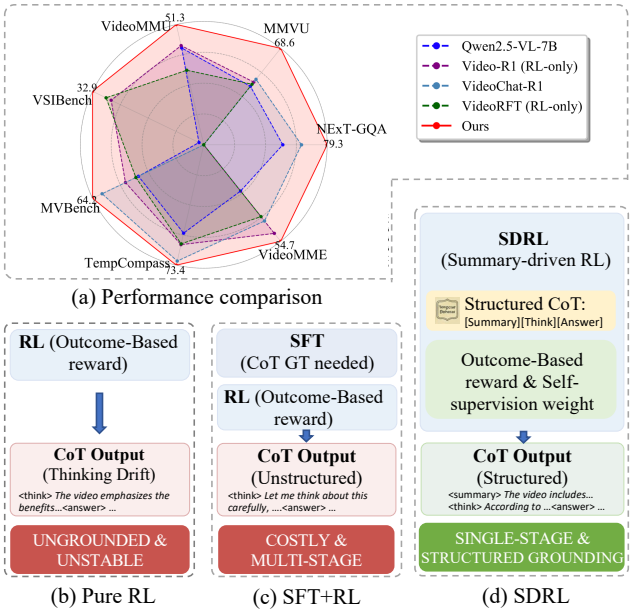


Figure 1. Performance comparison and training paradigms for video reasoning models. (a) Performance comparison among several benchmarks. (b)-(d) Training paradigm analysis: (b) Pure RL often yields ungrounded and unstable Chain-of-Thought (CoT) outputs. (c) SFT+RL is costly and complex. In contrast, (d) SDRL (Summary-driven RL) utilizes a Structured CoT and self-supervision to achieve stable and grounded video reasoning.

1. Introduction

Multimodal Large Language Models (MLLMs) have significantly advanced the frontier of video understanding, enabling open-ended reasoning over dynamic visual scenes [1, 5, 7, 18, 40]. The capability of MLLMs is further amplified by the Chain-of-Thought (CoT) prompting technique [30, 34]. By explicitly introducing intermediate reasoning steps, CoT enhances both interpretability and logical reasoning, allowing models to "think before answering." However, realizing the full potential of CoT often requires high-quality reasoning data for effective training [10, 35, 45].

A significant recent advancement involves leveraging Reinforcement Learning (RL), such as Group Relative Policy Optimization (GRPO) [31], to enhance MLLMs' complex reasoning abilities [44]. By optimizing models with reward signals based on verifiable outcomes (*e.g.*, final answer correctness), RL offers a scalable path to elicit benefi-

*This work was conducted during an internship at Amazon. Email: peiyaowang@cs.stonybrook.edu.

†Corresponding authors: Haibin Ling (hling@cs.stonybrook.edu) and Kah Kuen Fu (kahkuen@amazon.com). Haibin Ling was involved in this work while affiliated with Stony Brook University.

cial problem-solving strategies without the need for extensive CoT labels. However, this outcome-driven solution is fundamentally limited in complex video tasks: (1) *Thinking drift from unconstrained reasoning*: Relying solely on the final reward leaves intermediate reasoning steps unconstrained. This often leads to thinking drift [25], where the model generates verbose or reasoning irrelevant with the visual evidence, significantly hindering result stability. (2) *Weak temporal reasoning*: MLLMs frequently represent video as stacked or averaged frame embeddings, hence ignoring fine-grained temporal dependencies. As recent studies demonstrate [6], such lack of temporal awareness causes significantly poor performance on temporally-sensitive VideoQA tasks.

An alternative to direct RL is to add imitation learning, often implemented via Supervised Fine-Tuning (SFT) on expert demonstrations [2, 15, 42]. SFT is used to instill targeted reasoning behaviors, compensating for the random exploration phase in RL. For instance, [26, 36] use SFT to inject explicit spatio-temporal information or integrate descriptive captions into CoT to enhance grounding. While SFT can distill valuable reasoning behaviors, its next-token prediction objective enforces rigid, token-level imitation, limiting the model's generalization beyond training data. Consequently, long and complex demonstrations often lead to overfitting and shallow reasoning. Moreover, most approaches with such SFT→RL pipeline require costly annotations, are time-consuming, and may potentially constrain the base model's intrinsic reasoning potential [23].

This critical research gap, the need for efficient, structurally supervised, and temporally grounded training, motivates our work. We introduce Summary-Driven Reinforcement Learning (SDRL), a novel single-stage framework designed to enhance the temporal action order fidelity and interpretability of MLLMs without the need for prior SFT. Our core innovation lies in the direct integration of a Structured CoT into the RL objective. Specifically, we propose a Summarize→Think→Answer structure. The Summarize stage explicitly mandates the correct temporal action order, serving as a robust, structure-based anchor that grounds the subsequent reasoning. To effectively optimize this structure with RL, we leverage a novel GRPO-based objective that balances alignment and exploration. This objective enhances group consistency across sampled summaries while encouraging dynamic diversity during the Think stage.

Our main contributions are as follows:

- We propose Summary-Driven Reinforcement Learning (SDRL), which utilizes a Structured CoT format (Summarize→Think→Answer) to enhance temporal and factual reasoning fidelity, effectively recuding the reliance on SFT.
- We introduce two complementary mechanisms: Consistency of Vision Knowledge (CVK) to enforce factual grounding via group-level summary alignment, and Dynamic Variety of Reasoning (DVR) to promote exploration by modulating reasoning diversity.
- We simplify a two-stage SFT+RL pipeline to a single-stage RL-only framework, leading to superior performance across seven public VideoQA benchmarks.

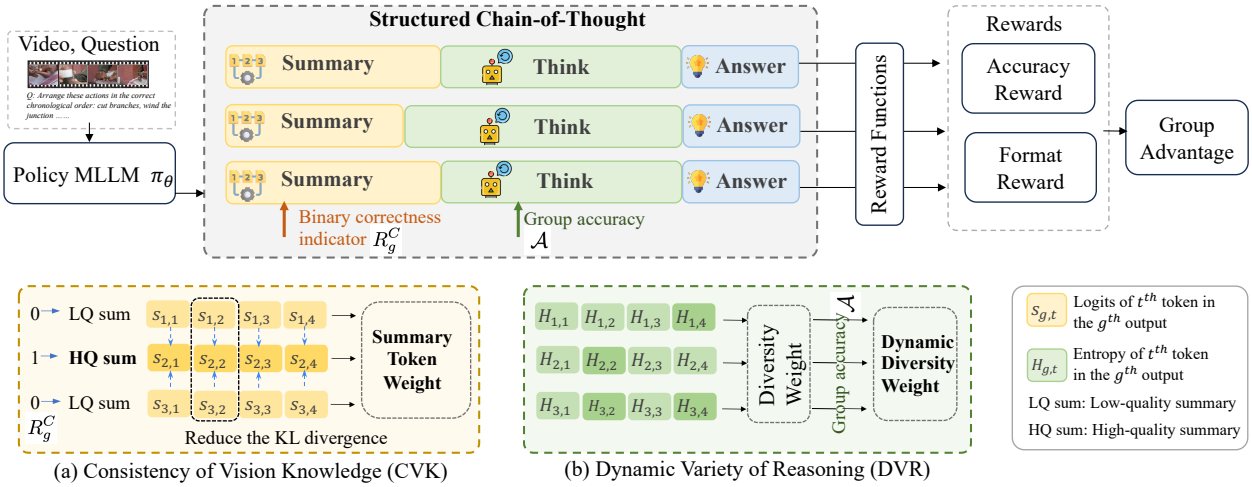
2. Related Work

Reinforcement Learning for MLLMs in Video Understanding. Reinforcement Learning, especially GRPO, has been widely applied to improve LLMs' reasoning capacity [9, 12, 31, 32]. Recent works extend GRPO to spatiotemporal reasoning in videos. R1-Omni [47], Video-R1 [7], and AoT [38] reveal the benefits of temporal consistency and implicit reasoning rewards, while Video-VER [25] grounds reasoning evidentially. However, they rely on multi-stage SFT+RL pipelines for best performance, which potentially constrain exploration and lead to overfitting to training reasoning patterns. VideoChat-R1 [18] introduces an IoU-based soft reward for grounding, TW-GRPO [5] focuses on temporal credit assignment, and GRPO-CARE [3] enforces group-level consistency. These efforts highlight the potential of process-aware RL for robust and coherent video reasoning.

Process Supervision and Verification of CoT. Recent progress in large language models (LLMs) has shifted from outcome-based optimization toward process-level supervision, which explicitly monitors and evaluates intermediate reasoning steps rather than only final predictions. Early studies [20, 27] proposed process reward models (PRMs) that assign feedback to each reasoning step, improving interpretability and reasoning fidelity. Subsequent extensions, *e.g.*, PSPO* [16], LongRePS [49] and ThinkPRM [13], incorporated non-linear reward shaping, long-context reasoning, and automatic verification to scale process supervision without exhaustive human annotation. While these methods enhance a model's reasoning capacity, they typically require expensive annotation of intermediate reasoning steps, and they frequently suffer from noisy or mis-aligned process-step rewards [41, 46]. In contrast, our proposed method provides self-supervision on the reasoning process without the need for explicit process annotations.

3. Method

Our method, Summary-Driven RL (SDRL), extends GRPO to realize the Structured CoT paradigm, enabling robust reasoning without supervised instruction tuning. Addtionally, SDRL proposes two self-supervised objectives: Consistency of Vision Knowledge (CVK) (Sec. 3.2), which aligns summaries with factual content, and Dynamic Variety of Reasoning (DVR) (Sec. 3.3), which promotes diverse reasoning paths. These constraints are unified under the Struc-



(a) Consistency of Vision Knowledge (CVK) (b) Dynamic Variety of Reasoning (DVR)

Figure 2. Overview of the SDRL Framework, which introduces Structured Chain-of-Thought. The Policy Model (π_θ) generates G reasoning sequences, each structured as Summary, Think, and Answer. The framework introduces two structured objectives implemented via token-wise weight: (a) Consistency of Vision Knowledge (CVK) and (b) Dynamic Variety of Reasoning (DVR). These structured weights, along with standard rewards (Accuracy, Format), are combined to derive the group advantage for policy optimization.

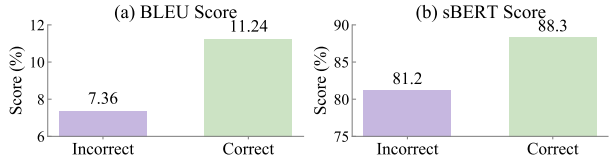


Figure 3. BLEU and sBERT scores between different predictions.

tured Policy Objective (Sec. 3.4) via token-wise weighting over Summary and Thinking segments.

3.1. Structured CoT for Top-down Reasoning

What constitutes an effective CoT in video understanding? The quality of a model's final prediction is tightly coupled with the fidelity of its intermediate reasoning steps. In video understanding, an effective CoT must explicitly capture the key actions and their correct temporal order. To verify this dependence, we analyzed the similarity (using BLEU [28] and sBERT [29] scores) between the model-generated CoT and ground-truth CoT[1] under both correct and incorrect final predictions. As shown in Fig. 3, correct predictions consistently exhibit higher CoT-to-ground-truth similarity than incorrect predictions. This analysis reveals a trend: superior performance correlates with CoT sequences that more closely align with the underlying factual reasoning. Therefore, we argue that an effective CoT for video understanding can explicitly capture two core components: **(1) the key actions or events**, and **(2) the temporal order** in which these events unfold. Motivated by this, we introduce a Structured CoT format enforcing a top-down process: **Summary** (salient, ordered events) → logical **Thinking** → final **Answer**.

How should the summary be obtained? The procedure for generating the summary depends on annotation availability. (1) With Ground-Truth Annotations: When fine-grained temporal labels (e.g., action segments with boundaries) are available, the summary is deterministically constructed from the ground truth. (2) Without Ground-Truth (Self-Exploration): When such annotations are unavailable, the model leverages its intrinsic ability to extract high-level temporal cues through self-exploration. To facilitate this, according to the empirical results in Figure 4, we prepend a dedicated <summary>[2] tag in the prompt.

Specifically, for an input $x = (\text{Video}, \text{Question})$, the model π is used to generate a group of G sampled outputs:

$$\{\mathcal{O}_g\}_{g=1}^{G} \sim \pi(\cdot|x)\ . \tag{1}$$

Each individual output $\mathcal{O}_g$ is a sequence of token logits, which follows the three-part structured trajectory (Summarize, Think, Answer). We segment the sequence into contiguous parts for clarity in the subsequent discussion:

$$\mathcal{O}_g = \{o_{g,t}\}_{t=1}^{T}, \quad \text{where } \mathcal{O}_g = \{S_g, P_g, Y_g\}, \tag{2}$$

where V denotes the vocabulary size, $o_{g,t} \in \mathbb{R}^V$ represents the unnormalized logit vector of the t-th output token, and $\mathcal{O}_g$ is segmented into the Summary (S_g), Thinking (P_g),

[1]The ground-truth CoT sequences were constructed by annotating key actions and their correct temporal order.

[2]<summary> is placed before <think> to enable explicit supervision on the summary segment.

and Answer (Y_g) segments. Each segment are defined as:

$$\begin{aligned} \textbf{Summary: } S_g &= \{s_{g,t}\}_{t=1}^{T^s} = \{o_{g,t}\}_{t=1}^{T'}, \\ \textbf{Thinking: } P_g &= \{p_{g,t}\}_{t=1}^{T^p} = \{o_{g,t}\}_{t=T'+1}^{T''}, \\ \textbf{Answer: } Y_g &= \{y_{g,t}\}_{t=1}^{T^y} = \{o_{g,t}\}_{t=T''+1}^{T}. \end{aligned} \tag{3}$$

Here, T denotes the total number of output tokens. T', T'' are the boundary indices that separate the summary, thinking, and answer segments, respectively, corresponding to the sequential token indices $[1, T']$, $[T'+1, T'']$, and $[T''+1, T]$.

3.2. Consistency of Vision Knowledge (CVK)

Group consistency of Summarization in RL. The generated summaries are not always tightly grounded in the visual content. Instead of relying on an SFT process, we directly introduce a structural supervision signal within the RL framework to enforce visual fidelity in the summary.

We begin with the assumption that the underlying visual content of a video is *fixed and factual*. Therefore, for a given input x, any set of sampled high-level action summaries, $\{S_g\}_{g=1}^{G}$, generated by a robust model should exhibit **strong semantic consistency** across different generations. Formally, all sampled summaries S_g must be drawn from a highly concentrated distribution $\mathcal{D}_S$:

$$\forall g \in [1, G], \quad S_g \sim \mathcal{D}_S(\cdot|x) \quad , \tag{4}$$

where $\mathcal{D}_S$ represents a distribution tightly aligned with a singular, factual semantic anchor.

Based on this assumption, we propose a group-level consistency objective which enforces summary alignment across the G outputs generated from the same input. Specifically, this objective aims to minimize the semantic dispersion of all sampled summaries S_g around a common consistency anchor $\hat{S}$, which represents the ideal factual summary for the group.

To this end, the objective function must be formulated to either maximize the average similarity or minimize the average consistency cost (dissimilarity) among the group members relative to $\hat{S}$. We integrate this objective within the GRPO framework, as it naturally provides a pipeline to sample G responses for each video–question pair, enabling the direct computation of group-level metrics. It needs two key factors: (1) The Consistency Anchor ($\hat{S}$): This serves as the reference representation that guides the alignment among all group members. (2) The Similarity/Dissimilarity Metric: This measures how closely each generated summary aligns with the anchor in the semantic space.

GT Supervised Summary Alignment. When ground-truth (GT) summaries S^{gt} are available, we use them as the definitive consistency anchor, i.e., $\hat{S} = S^{\text{gt}}$. To accurately assess the fidelity of each generated summary, we employ a composite similarity metric that integrates both semantic and lexical signals via sBERT and BLEU, respectively. Formally, the similarity between the g-th sampled summary S_g and the GT anchor is defined as:

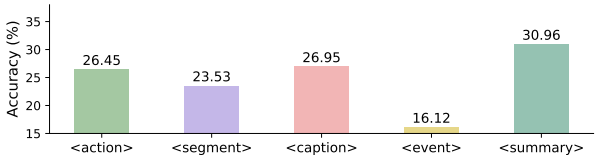
$$\text{Sim}(S_g, \hat{S}) = \alpha \cdot \text{sBERT}(S_g, S^{\text{gt}}) + \beta \cdot \text{BLEU}(S_g, S^{\text{gt}}), \tag{5}$$

where α and β are weights to balance the two components and ensure the score lies in $[0, 1]$.

We incorporate this similarity score into the RL objective by augmenting the original answer-based reward R_g with a summary-consistency reward, resulting in:

$$R'_g = \gamma_1 R_g + \gamma_2 \text{Sim}(S_g, \hat{S}) \, , \tag{6}$$

where γ_1 and γ_2 are scaling factors.


Accuracy (%)
35
30
25
20
15
26.45
23.53
26.95
16.12
30.96
<action>

<caption>
<event>
<summary>

Figure 4. Accuracy comparison across different tag types under training-free, inference-only settings for selecting an appropriate structural format prior to RL optimization.

Self-Supervised Summary Alignment. Obtaining GT summaries is costly and time-consuming, and strict alignment to GT annotations may constrain the model's expressive capacity or induce overfitting and bias. Therefore, we extend the group consistency objective to a self-supervised setting that eliminates the need for manual annotations. We dynamically derive the consistency anchor ($\hat{S}$) from the model's own predictions and identify high-quality summaries based on the binary correctness indicator R_g^C:

$$R_g^C = \begin{cases} 1, & Z_g = Z^{\text{gt}}, \\ 0, & Z_g \neq Z^{\text{gt}}, \end{cases} \tag{7}$$

where Z_g and the Z^{gt} denote the predicted and ground-truth answers, respectively. The position-wise center S^C, which serves as our consistency anchor $\hat{S}$, is then computed by aggregating the token representations of all selected high-quality summaries. Let $S^C = \{s_t^C\}_{t=1}^{T^s}$ denote the consistency center, and $S_g = \{s_{g,t}\}_{t=1}^{T^s}$ denote the g^{th} summary's token representation. For each position $t = 1, 2, \ldots, T^s$, the position-wise center s_t^C is computed as:

$$s_t^C = \frac{1}{\sum_{g=1}^{G} R_g^C} \sum_{g=1}^{G} R_g^C \cdot s_{g,t} \, . \tag{8}$$

To implement the self-supervised objective, we utilize Kullback–Leibler (KL) divergence as our dissimilarity metric to quantify the position-wise inconsistency within the

group. The KL divergence $\mathcal{D}_t$ at position t is calculated as:

$$\mathcal{D}_t = \frac{1}{G}\sum_{g=1}^{G} D_{\text{KL}}(s_{g,t}||s_t^C)\ . \tag{9}$$

To align with the GRPO Policy Objective, we convert this inconsistency measure into the Summary Token Weight (ω_t^S). Since a larger divergence ($\mathcal{D}_t$) indicates lower consistency (less agreement), the policy should assign a smaller weight to that token position. This encourages the model to focus on learning stable and consistent parts of the summary.We define the Summary Token Weight ω_t^S as:

$$\omega_t^S = 1 - \lambda \cdot \frac{\mathcal{D}_t - \mathcal{D}_{\min}}{\mathcal{D}_{\max} - \mathcal{D}_{\min}}\ . \tag{10}$$

Here, $\lambda \in [0, 1]$ controls the scaling intensity. This weight ω_t^S is then applied to the Summary Section ($1 \le t \le T'$) of the total GRPO Policy Objective, as illustrated in $\mathcal{J}_{\text{total}}(\theta)$, to realize the Self-Supervised Consistency goal.

3.3. Dynamic Variety of Reasoning (DVR)

While summary consistency stabilizes the factual grounding of the model, excessive uniformity in subsequent reasoning paths can be detrimental. Therefore, we introduce the Dynamic Variety of Reasoning (DVR) objective to encourage diversity in the <think> stage of structured CoT.

Specifically, we encourage diversity in the subsequent Thinking Segment (P) by focusing on the entropy of the token distribution at each position. The Diversity Weight $\omega_{g,t}^d$ is directly proportional to the measured entropy $H_{g,t}$, which is calculated over the predicted token distribution $p_{g,t}$:

$$H_{g,t} = -\,p_{g,t}^{\top}\log p_{g,t}. \tag{11}$$

This measured entropy is then normalized to define the base Diversity Weight $\omega_{g,t}^d$:

$$\begin{aligned} \omega_{g,t}^d &= 1 + \lambda' \cdot \frac{H_{g,t} - H_{\min}}{H_{\max} - H_{\min}}, \\ H_{\min} = \min_{g,t}\{H_{g,t}\}, &\quad H_{\max} = \max_{g,t}\{H_{g,t}\}\ , \end{aligned} \tag{12}$$

where λ' is a scaling hyperparameter.

Simply maximizing diversity can be detrimental to performance as the policy converges. When a group yields a high number of positive samples, it indicates the existing reasoning paths are effective and less exploration is needed. In such high-accuracy groups, excessive diversity may introduce noise.Therefore, we introduce a dynamic modulation coefficient based on the group's performance to adjust the diversity incentive. We define the group accuracy $\mathcal{A}$ as the fraction of correct answers:

$$\mathcal{A} = \frac{\sum_{g=1}^{G} \mathbf{1}_{Z_g = Z^{\text{gt}}}}{G}. \tag{13}$$

The base Diversity Weight $\omega_{i,t}^d$ is then reweighted by the factor $(1 - \mathcal{A})$ to yield the Dynamic Diversity Weight $\omega_{i,t}^{d'}$:

$$\omega_{i,t}^{d'} = \omega_{i,t}^d \cdot (1 - \mathcal{A}). \tag{14}$$

This formulation ensures that the diversity incentive is strongest for groups with low overall accuracy. Conversely, it is minimized for highly accurate groups, preserving stable reasoning paths. This dynamic weight $\omega_{i,t}^{d'}$ is used to inject the diversity signal into the final policy objective $\mathcal{J}_{\text{total}}(\theta)$.

3.4. Structured Policy Objective

We define the final policy objective $\mathcal{J}_{\text{total}}(\theta)$ as maximizing the expected augmented reward, which is integrated with our structural consistency and diversity constraints and the GRPO-specific policy regularization terms.The overall objective is to maximize the following expression:

$$\mathcal{J}_{\text{total}}(\theta) = \mathcal{J}_{\text{grpo}}^{\text{SCoT}}(\theta) - \mathcal{J}_{\text{reg}}(\theta)\ . \tag{15}$$

The structured GRPO objective $\mathcal{J}_{\text{grpo}}^{\text{SCoT}}(\theta)$ is designed to maximize the token-wise weighted advantage, the weights $W_{g,t}$ enforce the consistency constraints in the summary and dynamic diverstiy in the thinking. This objective is defined as:

$$\begin{aligned} \mathcal{J}_{\text{grpo}}^{\text{SCoT}}(\theta) =& \mathbb{E}\Big[\frac{1}{G}\sum_{g=1}^{G}\big(\sum_{t=1}^{T} W_{g,t}\cdot\min\big(r_{g,t}(\theta)\cdot A_{g,t}, \\ & \text{clip}(r_{g,t}(\theta), 1-\epsilon, 1+\epsilon)\cdot A_{g,t}\big)\big)\Big]\ , \\ r_{g,t}(\theta) =& \frac{\pi_\theta(o_{g,t}|q, o_{g,<t})}{\pi_{\text{old}}(o_{g,t}|q, o_{g,<t})}. \end{aligned} \tag{16}$$

The term $A_{g,t}$ represents the overall relative advantage (computed by mean-variance normalization of the enhanced reward R_g'), and the Token-Wise Weights $W_{g,t}$ modulate the policy update strength at each token position:

$$W_{g,t} = \begin{cases} \omega_t^S\ , & 1 \le t \le T' \\ \omega_{g,t}^{d'}\ , & T'+1 \le t \le T \end{cases} \tag{17}$$

3.5. Dataset Construction

Previous work has consistently highlighted a critical limitation in MLLMs: the inability to robustly capture and utilize fine-grained temporal information [6]. To facilitate effective video understanding training, which is a necessity for our proposed SDRL framework, we introduce EventFlowQA, a comprehensive video question-answering dataset focused on intricate action sequencing and temporal causality. In total, EventFlowQA comprises 53K high-quality QA pairs (50K for training and 3K for validation). The final distribution is over 15 focused temporal aspects, serving as the core benchmark for all ablation studies. The detailed methodology for dataset construction and analysis is provided in the Supplementary Material.

Table 1. Ablation study of CVK and DVR modules on the EventFlowQA. Rows (a–g) use GT supervision, while (h–l) adopt self-supervision. Adding semantic (sBERT, BLEU) and distributional (KL, Entropy, Dynamic) constraints improves accuracy.

	CVK		DVR			Accuracy
	sBERT	BLEU	KL	Entropy	Dynamic	
Orig.	-	-	-	-	-	42.37
(a)		✓				43.85
(b)	✓					46.32
(c)	✓	✓				48.56
(d)	✓	✓	✓			46.71
(e)	✓	✓	✓		✓	49.13
(f)	✓	✓		✓		50.09
(g)	✓	✓		✓	✓	52.22
(h)	self supervision					54.28
(i)	self supervision		✓			53.34
(j)	self supervision		✓		✓	55.78
(k)	self supervision			✓		54.13
(l)	self supervision			✓	✓	56.10

4. Experiment

Training Setup. We adopt an RL-only training paradigm to isolate the effect of policy optimization from SFT. We use Qwen2.5-VL-Instruct-7B as the backbone. Each training sample consists of 16 uniformly sampled frames at a resolution of 128×28×28, and inference is conducted under the same 16-frame setting for consistency. For the ground-truth–supervised consistency objective, we set weighting coefficients to $\alpha = 0.7$ and $\beta = 0.3$, with $\gamma_1 = 1$ and $\gamma_2 = 1$. For the self-supervised CVK and DVR objectives, we use $\lambda = 0.5$ and $\lambda' = 0.7$. Training is performed on 32 NVIDIA A100 GPUs with a GRPO group size of 8 for a total of 1,000 RL iterations.

We conduct two types of experiments in this paper. (1) All ablation studies are conducted on our proposed *EventFlowQA* dataset, which provides ground-truth action sequence annotations and enables controlled analysis of both GT-supervised and self-supervised summary consistency and diversity mechanisms. (2) Benchmark comparisons (Table 2) are conducted on seven public VideoQA benchmarks. For fair comparison, SDRL is trained on the same Video-R1-260K training data as prior RL-based methods, unless otherwise specified (e.g., Ours† trained on EventFlowQA). Further implementation details are provided in the supplementary material.

Benchmarks. We evaluate our model comprehensively across seven widely used video understanding benchmarks: NExT-GQA [37], MMVU [48], VideoMMMU [11], VSI-Bench [39], MVBench [17], TempCompass [24], and VideoMME [8]. These benchmarks can be broadly categorized into two groups: (1) *Video Reasoning Benchmarks* (NExT-GQA, MMVU, VideoMMMU, VSIBench) are designed to assess a model's temporal and causal reasoning capabilities, including multi-choice question answering, compositional inference, and long-range dependency understanding. (2) *General Video Understanding Benchmarks* (MVBench, TempCompass, VideoMME) focus on holistic video comprehension, integrating perception-level understanding (e.g., object, action, and event recognition) with high-level reasoning abilities.

4.1. Comparision with State of the Art method

Table 2 presents a comprehensive comparison between our method and recent Video MLLMs across both video reasoning and general understanding benchmarks. Overall, our RL-only framework (Ours) achieves consistent state-of-the-art performance, surpassing both SFT-only and SFT+RL pipelines. On reasoning benchmarks such as NExT-GQA, MMVU, and VideoMMMU, our model outperforms the SFT+RL method (VideoRFT*) by +4.2%, +1.6%, and +0.7%, respectively. On general benchmarks including MVBench, TempCompass, and VideoMME, SDRL further yields gains of +5.3%, +3.9%, and +5.7%, demonstrating strong generalization and temporal reasoning ability. Moreover, compared to other single-stage RL methods (e.g., VideoChat-R1, TW-GRPO), SDRL consistently achieves higher accuracy across all metrics. When compared with the base model Qwen2.5-VL, our approach yields accuracy improvements of up to +6.3% and +9.5% points on VSIBench (as indicated by the green numbers), even under different training data settings. Using EventFlowQA, which is approximately half the size of the Video-R1 training dataset, SDRL achieves the highest performance on TempCompass, highlighting the efficiency of our dataset. Distinct from previous approaches that rely on SFT followed by RL fine-tuning, our RL-only strategy stabilizes optimization through structured reasoning supervision, delivering both higher accuracy and greater training efficiency.

4.2. Analysis in Consistency of Summarization

Ground-Truth Supervision vs. Self-Supervision. Although GT supervision intuitively provides stronger guidance by anchoring predictions to human references, our results reveal a nuanced trend influenced by model scale and pre-training. As shown in Table 1, the larger 7B model benefits more from self-supervision ($+11.91\%$) than from GT supervision ($+6.19\%$), likely due to catastrophic forgetting that strict GT alignment with limited human summaries can over-constrain optimization and suppress useful semantic priors from pre-training. In contrast, self-supervision exploits semantic consistency among predictions, enhancing temporal and factual reasoning without destabilizing parameters. For the smaller 3B model (Table 3), GT super-

Table 2. Evaluation of Video MLLMs showing performance (Accuracy %) across various benchmarks, categorized by their training strategy (SFT, RL, SFT+RL). Our model achieves state-of-the-art results by employing the RL-only strategy, demonstrating superior performance across most metrics. 'Ours†' denotes the variant trained using the EventFlow dataset. VideoRFT* indicates evaluation with 16-frame input.

Models	Training	Video Reasoning Benchmark				Video General Benchmark		
		NExT-GQA	MMVU	VideoMMMU	VSIBench	MVBench	TempCompass	VideoMME
LLaMA-VID [19]	None	-	-	-	-	41.9	45.6	-
VideoLLaMA2 [4]	None	-	44.8	-	-	54.6	-	47.9
LongVA-7B [43]	None	-	-	23.9	29.2	-	56.9	52.6
VILA-1.5-8B [21]	None	-	-	20.8	28.9	-	58.8	-
Video-UTR-7B [40]	None	-	-	-	-	58.8	59.7	52.6
LLaVA-OneVision-7B [14]	None	-	49.2	33.8	32.4	56.7	-	**58.2**
Kangaroo-8B [22]	None	-	-	-	-	61.1	62.5	56.0
Qwen2.5-VL-7B [1]	None	75.9	65.4	48.4	29.1	63.3	72.5	56.5
Qwen2.5-VL-7B (video-r1 CoT) [1]	None	-	59.2	47.8	27.7	57.4	72.2	53.1
Qwen2.5-VL-7B (ours CoT) [1]	None	73.6	63.2	49.3	26.6	58.9	69.5	49.0
Video-R1 [7]	SFT	-	51.3	47.4	31.8	59.4	69.2	52.8
VideoRFT* [33]	SFT	-	60.5	48.5	31.7	57.0	68.4	54.1
Video-R1 [7]	SFT+ RL	74.3	64.2	**52.4**	34.6	62.7	72.6	57.4
VideoRFT* [33]	SFT+ RL	75.1	67.3	50.6	35.7	61.4	73.1	58.1
Video-R1 [7]	RL	-	63.8	49.5	31.8	60.4	70.9	53.8
VideoChat-R1 [18]	RL	76.0	64.2	-	-	63.1	72.9	52.4
VideoRFT* [33]	RL	-	63.5	47.4	32.1	59.2	70.8	51.9
TW-GRPO [5]	RL	76.1	65.8	-	-	63.3	73.3	55.1
SDRL (Ours†)	RL	77.3 (+3.7)	64.8 (+1.6)	51.1 (+1.8)	**36.1** (+9.5)	63.3 (+4.4)	**74.4** (+4.9)	53.1 (+4.1)
SDRL (Ours)	RL	**79.3** (+5.7)	**68.6** (+5.4)	51.3 (+2.0)	32.9 (+6.3)	**64.2** (+5.3)	73.4 (+3.9)	54.7 (+5.7)

Table 3. Effect of model size under different supervision types (Accuracy %). Smaller model gains more under GT supervision.

Model size	Original	GT sup.	Self sup.
3B	41.46	44.47 (+3.01)	43.86 (+2.40)
7B	42.37	48.56 (+6.19)	54.28 (+11.91)

Table 4. Comparison of BLEU and sBERT scores based on CVK.

	Accuracy (%)	BLEU (%)	sBERT (%)
w/o	42.37	8.84	70.33
w	54.28	12.57	79.76

vision yields slightly higher gains, indicating that smaller models depend more on explicit human guidance, whereas larger ones benefit from self-consistency regularization.

Metric for Consistency Constraint under GT Supervision. To determine the most effective metric for enforcing summary consistency under ground-truth (GT) supervision, we compare BLEU and sBERT, which evaluate similarity between predicted and GT summaries from complementary perspectives. As shown in Table 1, using (b)sBERT alone outperforms (a) BLEU alone(46.32% vs. 43.85%), indicating that semantic-level supervision provides stronger and more stable guidance than surface-level token matching. Moreover, combining BLEU and sBERT (c) yields the highest accuracy of 46.71%, surpassing either metric individually. This result suggests that BLEU and sBERT offer complementary benefits, and their integration thus delivers a more balanced supervision signal.

4.3. Analysis in the diversity of the thinking

Is diversity necessary within the group? As shown in Table 1, when diversity enhancement is applied statically to all groups, such as in (d) and (f), the model achieves accuracies of 49.13% and 55.78%. By dynamically applying diversity to encourag exploration only in uncertain or incorrect groups while maintaining stability in confident ones, the accuracy consistently rises to 52.22% and 56.10%. These results reveal that indiscriminately enforcing diversity can degrade performance when the model is already confident. In contrast, the dynamic strategy *focuses exploration on uncertain groups*, fostering richer reasoning where necessary while *preserving stability* for confident ones. By suppressing diversity in "solved" groups and amplifying it in "unsolved" ones, the model receives stronger, more informative learning signals, leading to higher overall accuracy.

Which metric is more effective in promoting diversity? We further examine which metric better promotes reasoning

Figure 5. Comparison of CoTs and final answers generated by VideoChat-R1 and our proposed SDRL method. SDRL demonstrates superior grounding and logic flow, evidenced by higher BLEU and sBERT scores relative to the ground truth summary.

diversity during GRPO training. As shown in Table 1, we compare two diversity objectives, *i.e.*, KL divergence and entropy regularization, under both GT and self-supervised consistency settings. Entropy consistently yields higher accuracy and more stable performance than KL. For example, under GT supervision, replacing KL (49.13%, (d)) with entropy (50.09%, (e)) improves accuracy, and a similar gain is observed in the self-supervised setting (55.78% $\rightarrow$ 56.10%). This performance gap arises from their intrinsic difference: KL divergence enforces *local, position-dependent alignment* across token distributions, which can suppress global variability, while entropy regularization acts as a *global uncertainty control*, encouraging balanced exploration without collapsing into deterministic reasoning. Consequently, entropy better preserves semantic diversity while stabilizing the reasoning process.

5. Visualization

Figure 5 illustrates qualitative comparisons of the generated summaries and reasoning outputs without GT supervision. In Figure 5(a), VideoChat-R1 produces an incorrect answer due to poor temporal reasoning. Its generated summary fails to preserve the correct action order, resulting in a lower BLEU score (5.2) and weak alignment with the ground-truth sequence. In contrast, SDRL generates a summary that closely follows the true action sequence and correctly predicts the answer. This indicates that the proposed summary consistency constraint helps the model capture accurate temporal dependencies and maintain coherent action ordering. Even when both models produce the correct answer in Figure 5(b), SDRL achieves a higher BLEU score, indicating that its generated summary more faithfully reflects the action sequence and provides clearer temporal organization. Table 4 further shows that CVK consistently improves both summary consistency and task accuracy.

6. Conclusion

In conclusion, we introduced Summary-Driven Reinforcement Learning (SDRL), a framework that relies policy optimization paradigm for MLLMs in video understanding. By integrating a Structured Chain-of-Thought (Summarize $\rightarrow$ Think $\rightarrow$ Answer) and leveraging self-supervised Token-Wise Weighting (via CVK and DVR), SDRL alleviates the critical challenges of thinking drift and the multi-Stage training pipeline inherent in existing SFT+RL methods. Our framework forces the model's intermediate reasoning by anchoring it to the explicit Summary segment, enabling robust and interpretable decision-making without reliance on human-annotated CoT data. We anticipate that this work will pave the way for future research into single-stage, data-efficient, and structurally robust reasoning frameworks for general MLLMs.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 7

[2] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *Transactions on Machine Learning Research*, 2025. 2

[3] Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*, 2025. 2

[4] 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. 7

[5] Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *arXiv preprint arXiv:2505.24718*, 2025. 1, 2, 7

[6] Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *Advances in Neural Information Processing Systems*, 2025. 2, 5

[7] 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. *Advances in Neural Information Processing Systems*, 2025. 1, 2, 7

[8] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 24108–24118, 2025. 6

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 2025. 2

[10] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 26181–26191, 2025. 1

[11] 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.(2025). *arXiv preprint arXiv:2501.13826*, 2025. 6

[12] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai-o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 2

[13] Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. In *Conference on Language Modeling*, 2025. 2

[14] 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. *Transactions on Machine Learning Research*, 2025. 7

[15] Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, et al. Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*, 2025. 2

[16] Jiawei Li, Xinyue Liang, Junlong Zhang, Yizhe Yang, Chong Feng, and Yang Gao. Pspo*: An effective process-supervised policy optimization for reasoning alignment. *arXiv preprint arXiv:2411.11681*, 2024. 2

[17] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 6

[18] 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. *Advances in Neural Information Processing Systems*, 2025. 1, 2, 7

[19] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 7

[20] 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 *International Conference on Learning Representations*, 2023. 2

[21] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 7

[22] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *International Journal of Computer Vision*, 2025. 7

[23] Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Uft: Unifying supervised and reinforcement fine-tuning. *Advances in Neural Information Processing Systems*, 2025. 2

[24] 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*, pages 8731–8772, 2024. 6

[25] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video

reasoning. *Advances in Neural Information Processing Systems*, 2025. 2

[26] Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. *Advances in Neural Information Processing Systems*, 2025. 2

[27] OpenAI. Improving mathematical reasoning with process supervision. OpenAI Technical Report, 2023. 2

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, pages 311–318, 2002. 3

[29] Nils Reimers and Iryna Gurevych. Sentence-bert:sentence embeddings using siamese bert-networks. *Conference on Empirical Methods in Natural Language Processing*, 2019. 3

[30] 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. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024. 1

[31] 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. *International Conference on Learning Representations*, 2024. 1, 2

[32] 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. 2

[33] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in Neural Information Processing Systems*, 2025. 7

[34] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025. 1

[35] 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

[36] Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025. 2

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

[38] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *Advances in Neural Information Processing Systems*, 2025. 2

[39] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10632–10643, 2025. 6

[40] Enze Yu, Kai Lin, Liang Zhao, Yifan Wei, Zhihao Zhu, Haoran Wei, Jian Sun, Zhitao Ge, Xinyu Zhang, Jingdong Wang, et al. Unhackable temporal rewarding for scalable video mllms. *International Conference on Learning Representations*, 2025. 1, 7

[41] Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *International Conference on Machine Learning*, 2025. 2

[42] Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in Neural Information Processing Systems*, 37:110935–110971, 2024. 2

[43] Peng Zhang, Kai Zhang, Bo Li, Guoqiang Zeng, Jianwei Yang, Yujing Zhang, Zhen Wang, Hao Tan, Chong Li, and Zhiqiang Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025. 7

[44] Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *Advances in Neural Information Processing Systems*, 2025. 1

[45] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024. 1

[46] Yao Zhang, Yu Wu, Haowei Zhang, Weiguo Li, Haokun Chen, Jingpei Wu, Guohao Li, Zhen Han, and Volker Tresp. Groundedprm: Tree-guided and fidelity-aware process reward modeling for step-level reasoning. *arXiv preprint arXiv:2510.14942*, 2025. 2

[47] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025. 2

[48] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8475–8489, 2025. 6

[49] Dawei Zhu, Xiyu Wei, Guangxiang Zhao, Wenhao Wu, Haosheng Zou, Junfeng Ran, Xun Wang, Lin Sun, Xiangzheng Zhang, and Sujian Li. Chain-of-thought matters: Improving long-context language models with reasoning path supervision. *Findings of the Association for Computational Linguistics*, 2025. 2