VIDEORUBRIC: RUBRIC GRANULARITY MATTERS IN REINFORCEMENT LEARNING FOR VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards is promising for improving video reasoning in multimodal large language models, yet answer-level rewards provide limited guidance on the reasoning process. We introduce VideoRubric, a rubric-guided reinforcement learning framework that studies process supervision for VideoQA at two granularities: specific rubrics describe sample-conditioned evidence and reasoning requirements, while abstract rubrics capture reusable patterns such as temporal consistency, causal modeling, and evidence grounding. VideoRubric constructs specific rubrics with a multimodal teacher, abstracts them into a shared rubric set, and incorporates rubric judgments into GRPO through correctness gating. Experiments with Qwen2.5-VL-7B-Instruct on 1K rubric-augmented samples show that the correctness-gated specific-rubric supervision achieves the strongest average performance across MMVU, TempCompass, VSI-Bench, and VideoMME. Further analyses show that rubric granularity shapes learned reasoning differently: specific rubrics yield clearer separation between correct and incorrect responses, whereas abstract rubrics produce more structured and preferred reasoning outputs.

1 INTRODUCTION

Multimodal large language models (MLLMs) have rapidly progressed from visual recognition toward complex reasoning over images and videos. Video reasoning is particularly demanding because a model must identify relevant visual evidence, track entities and events over time, and integrate spatial, temporal, and causal information before producing an answer. Recent RL advances show that outcome-based optimization can elicit and strengthen reasoning outputs Guo et al. (2025); Yu et al. (2026). This paradigm has subsequently been extended to visual and video reasoning, where reinforcement fine-tuning improves multimodal reasoning, temporal understanding, and evidence grounding Liu et al. (2025); Huang et al. (2025a); Feng et al. (2026); Li et al. (2025); Wang et al. (2026b). Nevertheless, improving how an MLLM reasons about a video remains more difficult than rewarding whether its final answer is correct.

Most reinforcement-learning approaches for reasoning rely on verifiable outcome rewards, such as answer correctness, and formatting constraints. These signals are objective, but collapse multiple reasoning dimensions into one score. In VideoQA, an incorrect answer may arise from missed evidence, temporal confusion, unsupported causal inference, or conclusions inconsistent with the observed outcome. Conversely, correct answers may still follow incomplete or weakly grounded reasoning. Thus, answer-level rewards assign the same supervision to qualitatively different trajectories and provide little guidance on which reasoning failures to correct. Recent work addresses this limitation with visual-perception, semantic-consistency, observation-and-reasoning, and evidence-driven rewards Xiao et al. (2025); Wang et al. (2026b); Zhang et al. (2025); Chen et al. (2026a). However, these methods typically instantiate a particular reasoning-level signal, leaving the effects of different forms of process feedback on learned reasoning outputs largely unexplored.

Rubric-based rewards express desirable and undesirable reasoning behaviors in natural language. Prior work has used human- or model-generated rubrics to extend RL to open-ended tasks, build scalable reward models, supervise chain-of-thought trajectories, or organize learning by model competence Huang et al. (2025b); Liu et al. (2026); Sheng et al. (2026); Jia et al. (2026); Chen et al. (2026b). Yet most studies focus on rubric generation, selection, evolution, or scheduling, leaving

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rubric granularity underexplored. Abstract rubrics can transfer across examples but may also reward plausible reasoning with wrong conclusions, whereas specific rubrics target concrete events, relations, or distinctions but provide more local supervision. Treating them as interchangeable overlooks that they may reward different properties of the same trajectory.

In this work, we introduce VideoRubric, a rubric-guided RL framework for VideoQA that contrasts outcome rewards with process supervision at two levels of granularity. Specific rubrics are sample-conditioned criteria describing the evidence and reasoning needed to justify the correct answer, such as relevant events, temporal or spatial relations, causal connections, or distinctions among plausible options. Abstract rubrics are iteratively derived from specific rubrics into shared high-level criteria, such as temporal consistency, causal modeling, and evidence grounding. During RL, a judge model evaluates generated reasoning against either rubric set. Correctness gating incorporates the resulting rubric scores into GRPO to differentiate correct responses without rewarding plausible but incorrect reasoning.

We evaluate VideoRubric with Qwen2.5-VL-7B-Instruct Bai et al. (2025) on a 1K rubric-augmented VideoQA set. Across MMVU Zhao et al. (2025), TempCompass Liu et al. (2024), VSI-Bench Yang et al. (2025), and VideoMME Fu et al. (2025), the specific-rubric variant with correctness gating performs best, averaging 56.36 versus 55.59 for outcome-only GRPO Shao et al. (2024) and 55.68 for gated abstract-rubric supervision. It also outperforms the reported zero-style baselines on most benchmarks despite using substantially fewer RL samples. Beyond accuracy, our analyses show that rubric granularity shapes reasoning differently: specific rubrics produce clearer rubric score separation between correct and incorrect responses, while abstract rubrics promote reasoning with more explicit organization, sustained comparison and conclusion development, and higher human and LLM preferences. These results suggest that benchmark accuracy and perceived reasoning quality capture different aspects of video reasoning, making rubric granularity a substantive design choice. Our main contributions are as follows:

- We identify rubric granularity as a key design dimension for video-reasoning process supervision and study how instance-grounded and shared abstract supervision differ in their effects on learned reasoning.
- We propose VideoRubric, a rubric-guided RL framework that constructs specific and abstract rubrics and integrates their judgments into GRPO through correctness gating.
- Through ablations on four video benchmarks, we show that specific-rubric supervision achieves the strongest overall performance. Behavioral analyses further show that rubric granularity shapes learned reasoning differently: specific rubrics provide clearer separation between correct and incorrect responses, whereas abstract rubrics produce more elaborated and preferred reasoning outputs.

2 Related Work

2.1 Reinforcement Learning for Video Reasoning

Reinforcement learning with verifiable rewards (RLVR) has been extended from language reasoning to multimodal and video reasoning. DeepSeek-R1 demonstrated that RL can elicit complex reasoning behaviors, while DAPO improved the stability and scalability of RL training Guo et al. (2025); Yu et al. (2026). In the visual domain, Visual-RFT, Vision-R1, and Perception-R1 introduced task-specific, chain-of-thought, or perception-aware rewards for visual reasoning Liu et al. (2025); Huang et al. (2025a); Xiao et al. (2025). Recent video-oriented methods further adapt RL to temporal and spatio-temporal reasoning: Video-R1 uses temporally guided GRPO, VideoChat-R1 applies reinforcement fine-tuning to spatio-temporal perception, VideoRFT introduces semantic-consistency rewards, ReWatch-R1 penalizes video-grounding errors, and Reflect-R1 optimizes self-correction with retrieved evidence Feng et al. (2026); Li et al. (2025); Wang et al. (2026b); Zhang et al. (2025); Chen et al. (2026a). Despite this progress, most methods rely on outcome rewards or a single predefined reasoning-level signal, leaving underexplored how feedback at different abstraction levels shapes learned reasoning behavior.

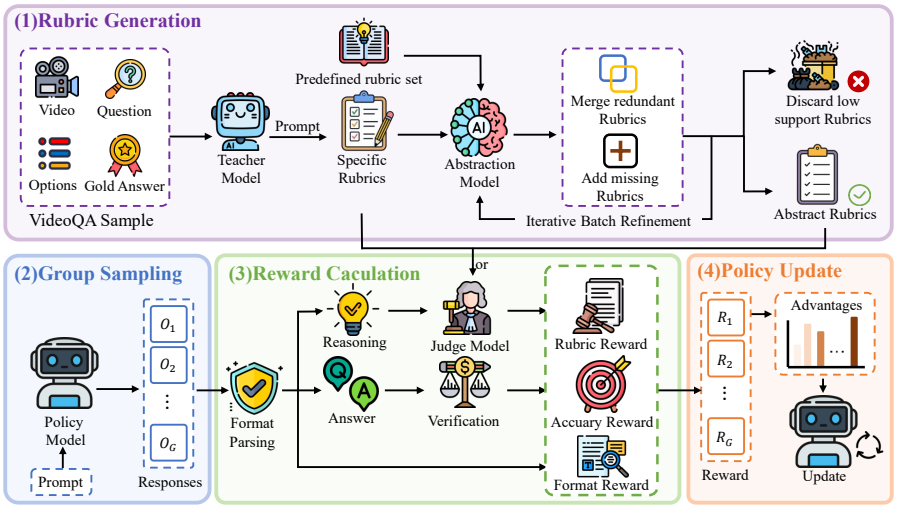

Figure 1: Overview of VideoRubric. A multimodal teacher first generates sample-specific rubrics for each VideoQA instance. These rubrics are then iteratively abstracted and consolidated into a shared abstract rubric set. During reinforcement learning, a judge model evaluates the generated CoT against either the specific rubrics or the abstract rubrics, and the resulting rubric score is combined with format and accuracy rewards within GRPO.

2.2 Rubric-Based Process Supervision

Rubric-based rewards provide structured supervision beyond scalar outcomes. Rubric Anchors uses natural-language criteria as reward anchors for open-ended RL Huang et al. (2025b), while OpenRubrics derives constraints and principles from preference pairs for rubric-based reward modeling Liu et al. (2026). RLCER and AutoRubric-R1V use rubrics to supervise reasoning trajectories, through self-evolving textual criteria and automatically aggregated multimodal reasoning checkpoints, respectively Sheng et al. (2026); Jia et al. (2026). RuCL further combines rubric supervision with curriculum learning by stratifying generalized rubrics according to model competence Chen et al. (2026b). These methods establish rubrics as scalable process-level signals, but mainly study how rubrics are generated, evolved, aggregated, or scheduled. In contrast, we study rubric granularity: abstract rubrics capture transferable reasoning patterns, while specific rubrics capture instance-level evidence requirements. We also examine how these two supervision types affect performance and reasoning behavior.

2.3 Adaptive Supervision in Reinforcement Learning

Another line of work improves RL by adapting supervision to model competence or past experience. VL-Cogito, SPARD, and Temp-R1 adjust task difficulty, reward weights, or data importance according to learning progress Yuan et al. (2025); Zhi et al. (2026); Gong et al. (2026), while Experiential RL and SkillRL convert feedback or past trajectories into reflection and reusable skills Shi et al. (2026); Xia et al. (2026). These methods demonstrate the value of structured adaptation, but mainly regulate when and what the model learns. Our work is complementary: it focuses on how the abstraction level of process supervision changes the reasoning behavior learned by the policy.

3 Method

We present VideoRubric, a rubric-guided reinforcement learning framework that introduces fine-grained reasoning supervision for VideoQA. As illustrated in Figure 1, VideoRubric constructs reasoning rubrics at two levels of granularity and uses them as response-level supervision during reinforcement learning. We first formulate the learning problem, then describe the construction of specific and abstract rubrics, and finally present the rubric-guided reward and training procedure.

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a VideoQA training set, where $x_i = (v_i, q_i)$ consists of a video v_i and its corresponding question q_i with associated answer options, and y_i is the ground-truth answer. Given x_i, the policy π_θ generates a response $o_i = (\hat{c}_i, \hat{y}_i) \sim \pi_\theta(\cdot \mid x_i)$, where $\hat{c}_i$ denotes the generated chain-of-thought reasoning and $\hat{y}_i$ denotes the predicted answer. Each response receives a scalar reward r_i that combines outcome-level feedback on the final answer with rubric-based feedback on the chain-of-thought reasoning. The policy is optimized to maximize the expected response-level reward, $\max_\theta \mathbb{E}_{(x_i,y_i)\sim\mathcal{D},\, o_i\sim\pi_\theta(\cdot|x_i)}\big[r_i\big]$.

3.2 Rubric Construction

3.2.1 Specific Rubric Generation

To provide fine-grained supervision for sample-level reasoning, we introduce specific rubrics, each defined as a sample-conditioned and independently assessable criterion that captures either concrete visual evidence or a reasoning step required to support the correct answer. A rubric may require the response to identify a relevant event, establish a temporal or spatial relation between observations, infer a causal connection, or distinguish the correct option from a visually plausible alternative. Collectively, these rubrics specify the evidence and reasoning needed for each video question without enforcing fixed wording or a predetermined reasoning sequence.

For each sample i, we employ a strong teacher model $\mathcal{M}_{\text{spe}}$ to generate a set of K_i^{spe} specific rubrics, $\mathcal{S}_i = \{s_{i,1}, \ldots, s_{i,K_i^{\text{spe}}}\}$. The teacher takes the input $x_i = (v_i, q_i)$ and ground-truth answer y_i as conditioning information to generate the rubric set:

$$\mathcal{S}_i = \mathcal{M}_{\text{spe}}(x_i, y_i). \tag{1}$$

The teacher examines the video in the context of the question, identifies the evidence supporting the ground-truth answer, and determines the relations required to connect that evidence to the answer. These requirements are then decomposed into clear, self-contained, and independent rubrics. To prevent the rubrics from trivially revealing the answer, the teacher is instructed not to restate y_i or directly encode the correct answer option in the generated rubrics. Examples of generated specific rubrics are provided in the Appendix B (Figure 6).

3.2.2 Abstract Rubric Construction

Specific rubrics provide fine-grained supervision by specifying concrete evidence and sample-conditioned reasoning requirements of individual questions, but their sample-dependent formulation does not explicitly capture the higher-level criteria underlying these requirements. This motivates an abstract level of supervision that focuses on reasoning patterns rather than sample-specific evidence.

Therefore, we introduce abstract rubrics, where each abstract rubric is an independently assessable, sample-agnostic criterion that describes a general property of the reasoning process rather than the specific evidence required by an individual question. Such criteria may assess whether the reasoning maintains temporal consistency, establishes a valid causal relation, or grounds its conclusion in relevant visual evidence. By removing sample-specific objects, events, and answer options while preserving the underlying reasoning requirement, abstract rubrics can correspond to multiple specific rubrics and form a shared high-level supervision space across samples.

The abstract rubric set is denoted by $\mathcal{A} = \{a_1, a_2, \ldots, a_{K^{\text{abs}}}\}$, where K^{abs} is the number of abstract rubrics. Starting from a predefined set $\mathcal{A}^{(0)}$ covering common high-level reasoning rubrics in VideoQA, we iteratively organize specific rubrics into abstract categories. Specifically, the specific rubrics are divided into T batches. At iteration t, the current abstract rubric set and a batch of specific rubrics $\mathcal{B}_t$ are provided to a strong abstraction model $\mathcal{M}_{\text{abs}}$:

$$\mathcal{A}^{(t)} = \mathcal{M}_{\text{abs}}\big(\mathcal{A}^{(t-1)}, \mathcal{B}_t\big). \tag{2}$$

For each specific rubric in $\mathcal{B}_t$, the model identifies the underlying reasoning rubric and assigns it to an existing abstract rubric when an appropriate category exists. Otherwise, a new abstract rubric is introduced by removing sample-specific details while preserving the essential reasoning requirement. After processing all batches, we remove abstract categories supported by fewer than γ

specific-rubric assignments. The remaining set $\mathcal{A}$ is used as shared high-level reasoning supervision during training. Both the initial set $\mathcal{A}^{(0)}$ and the final training set $\mathcal{A}$ are listed in Appendix C.

3.3 Rubric-Guided Reward Design

The constructed rubrics provide fine-grained reasoning process supervision beyond the final-answer outcome. Instead of modifying the optimization algorithm, we convert rubric evaluations into response-level rewards and incorporate them into the standard GRPO framework.

For each training sample i, the policy π_θ samples a group of G responses $\{o_{i,j}\}_{j=1}^{G}$. Each response is required to follow the tagged structure `<think>...</think><answer>...</answer>`. These tags separate chain of thought reasoning $\hat{c}_{i,j}$ from the predicted answer $\hat{y}_{i,j}$ for independent evaluation.. If parsing fails, both the accuracy and rubric rewards are set to zero.

To ensure that the chain-of-thought reasoning and final answer can be reliably extracted before applying the corresponding rewards, we first assign a binary format reward according to whether the response follows the prescribed structure and both fields can be successfully parsed.

$$r_{i,j}^{\text{fmt}} = \begin{cases} 1, & \text{Valid format,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The accuracy reward evaluates whether the extracted answer matches the ground-truth answer:

$$r_{i,j}^{\text{acc}} = \mathbf{1}[\hat{y}_{i,j} = y_i]\,, \tag{4}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function.

To study the effect of rubric granularity, we train two policy variants that differ only in the rubric supervision applied to each response. The specific-rubric variant evaluates each response with the sample-specific set $\mathcal{S}_i$, whereas the abstract-rubric variant uses the shared set $\mathcal{A}$. We denote the active rubric set by $\mathcal{R}_i$ and its size by $K_i = |\mathcal{R}_i|$. For each rubric $\tau_{i,k} \in \mathcal{R}_i$, a judge model$J_\phi$ evaluates whether the generated chain-of-thought reasoning $\hat{c}_{i,j}$ satisfies the corresponding criterion:

$$z_{i,j,k} = J_\phi(\hat{c}_{i,j}, \tau_{i,k}) \in \{0, 1\}, \tag{5}$$

where $z_{i,j,k} = 1$ indicates satisfaction. The rubric judge evaluates whether the generated CoT satisfies the reasoning requirement specified by each criterion. It operates on the generated reasoning and rubric text, while video-dependent evidence requirements in the specific-rubric setting are established upstream by the video-conditioned teacher. The rubric reward is computed as the proportion of satisfied criteria:

$$r_{i,j}^{\text{rub}} = \frac{1}{K_i} \sum_{k=1}^{K_i} z_{i,j,k}. \tag{6}$$

Finally, we combine the three reward components using correctness gating. Answer correctness provides a directly verifiable reward signal, whereas rubric satisfaction relies on model-based judgments and does not necessarily indicate that the reasoning leads to a correct answer. We therefore anchor the reward in answer correctness and use rubric satisfaction as a conditional refinement:

$$r_{i,j} = r_{i,j}^{\text{fmt}} + r_{i,j}^{\text{acc}} \left(1 + r_{i,j}^{\text{rub}}\right). \tag{7}$$

This formulation preserves the accuracy-based incentive of outcome-only GRPO while using rubric supervision to further differentiate correct responses. We also consider an additive formulation that applies rubric rewards regardless of answer correctness, allowing reasoning differences among incorrect responses to contribute to the learning signal. Empirically, the gated formulation performs better overall, suggesting that prioritizing verifiable answer correctness over process-level discrimination among incorrect responses provides a more effective training signal.

4 Experiments

4.1 Datasets and Evaluation

Training Data. We construct a rubric training set containing 1K VideoQA samples collected from five datasets: LLaVA-Video Zhang et al. (2024b), NeXT-QA Xiao et al. (2021), STAR Wu et al.

Table 1: Comparison with general video MLLMs and zero-style video reasoning methods. "RL Data" denotes the reported RL dataset size; "–" indicates unavailable or unreported results. * marks downstream evaluation results from Video-R1 Feng et al. (2026) and SDRL Wang et al. (2026a); other results are from the original papers. MMVU uses the multiple-choice subset, and VideoMME is evaluated without subtitles. Scores are percentages; the best and second-best per column are **bold** and underlined, respectively.

Method	RL Data	MMVU	TempCompass	VSI-Bench	VideoMME
General video MLLMs					
VideoLLaMA2* Cheng et al. (2024)	–	44.80	–	–	47.90
LongVA-7B* Zhang et al. (2024a)	–	–	56.90	29.20	52.60
VILA-1.5-8B* Lin et al. (2024)	–	–	58.80	28.90	–
Video-UTR-7B* Yu et al. (2025)	–	–	59.70	–	52.60
Zero-style video reasoning methods					
VideoChat-R1* Wang et al. (2026a)	18K	<u>64.20</u>	**72.90**	–	52.40
Video-R1-7B-Zero Feng et al. (2026)	260K	63.80	70.90	31.80	<u>53.80</u>
VideoRFT (RL-only) Wang et al. (2026b)	310K	63.50	70.80	<u>32.10</u>	51.90
VideoRubric	1K	**65.60**	<u>72.23</u>	**33.33**	**54.30**

(2024), CLEVRER Yi et al. (2019), and PerceptionTest Patraucean et al. (2023). Each sample is paired with a set of specific rubrics, while all samples share the same abstract rubric set. The data composition and video-duration distribution are provided in Appendix A.2 (Figure 4).

Evaluation Benchmarks. We evaluate VideoRubric on four widely used video-understanding benchmarks: MMVU for knowledge-intensive reasoning, TempCompass for temporal perception, VSI-Bench for spatial reasoning, and VideoMME for general video understanding across diverse domains and durations. We evaluate only the multiple-choice subset of MMVU and evaluate VideoMME without subtitles. All evaluations are performed with greedy decoding.

4.2 Implementation Details

We initialize the policy from Qwen2.5-VL-7B-Instruct and train all models on four NVIDIA RTX PRO 6000 Blackwell GPUs with 96 GB of memory each. During both reinforcement learning and benchmark evaluation, we uniformly sample 16 frames from each video, with a maximum pixel budget of $128 \times 28 \times 28$ for each frame. The GRPO group size is set to $G = 8$. Qwen3-Max is used as the frozen judge model during training and the model for abstract-rubric construction, while Qwen3-VL-Plus serves as the teacher model for specific-rubric construction. For abstract-rubric construction, we set the pruning threshold to $\gamma = 0.01 \sum_{i=1}^{N} K_i^{\mathrm{spe}}$, removing abstract rubrics supported by fewer than 1% of all specific rubrics. All experiments use the same training data, optimization hyperparameters, and decoding settings. Unless otherwise specified, VideoRubric refers to the specific-rubric variant trained with the correctness-gated reward. Further implementation details are provided in the Appendix A.

4.3 Comparison with Other Methods

We compare VideoRubric with several general video MLLMs and zero-style video reasoning methods. For methods with multi-stage training pipelines, we report their RL-only variants when such results are available. As shown in Table 1, VideoRubric achieves the best results among the compared zero-style methods on MMVU, VSI-Bench, and VideoMME. It outperforms Video-R1-7B-Zero, with gains of 1.80 % on MMVU, 1.33 % on TempCompass, 1.53 % on VSI-Bench, and 0.50 % on VideoMME. VideoRubric also consistently surpasses the RL-only variant of VideoRFT. On TempCompass, VideoRubric trails VideoChat-R1 by 0.67 %. Notably, VideoRubric uses only 1K RL samples, substantially fewer than the 18K–310K samples reported by the compared RL-based methods, indicating that rubric-guided reinforcement learning can provide effective supervision with a relatively small RL dataset.

Table 2: Ablation results for rubric type and reward formulation. "Additive" uses $r^{\text{fmt}} + r^{\text{acc}} + r^{\text{rub}}$, whereas "Gated" uses $r^{\text{fmt}} + r^{\text{acc}}(1 + r^{\text{rub}})$. Avg. denotes the unweighted average over the four benchmarks. All scores are percentages. The best and second-best results in each column are **bold** and underlined, respectively.

Configuration	MMVU	TempCompass	VSI-Bench	VideoMME	Avg.
Outcome-Only GRPO	63.36	**72.67**	33.49	52.85	55.59
VideoRubric-Abs (Additive)	63.68	71.98	33.69	51.44	55.20
VideoRubric-Abs (Gated)	<u>64.96</u>	72.19	<u>34.02</u>	51.56	55.68
VideoRubric-Spe (Additive)	63.84	<u>72.44</u>	**34.30**	<u>53.52</u>	<u>56.02</u>
VideoRubric-Spe (Gated)	**65.60**	72.23	33.33	**54.30**	**56.36**

4.4 Ablation Studies

We conduct a two-factor ablation of the design choices introduced in Section 3.3. First, we vary the granularity of rubric supervision by using either the shared abstract rubric set or the sample-specific rubric set. Second, we vary how the rubric score is incorporated into the reward: the additive formulation applies it regardless of answer correctness, whereas the gated formulation activates it only for correct responses. Outcome-only GRPO without rubric supervision serves as the baseline.

Table 2 shows that specific-rubric supervision yields stronger overall performance than abstract-rubric supervision under both reward formulations. With the additive formulation, specific rubrics achieve 56.02 on average, compared with 55.59 for outcome-only GRPO and 55.20 for abstract rubrics. The same trend holds under the gated formulation, where the average score increases from 55.68 with abstract rubrics to 56.36 with specific rubrics. This result suggests that sample-specific criteria provide a more direct training signal, as they explicitly identify the visual evidence and reasoning relations required by each question. In contrast, abstract rubrics describe general reasoning properties and may be less discriminative among responses generated for the same sample.

The gated formulation improves the average performance of both rubric types, raising the abstract-rubric result from 55.20 to 55.68 and the specific-rubric result from 56.02 to 56.36. Unlike the additive formulation, which may reward rubric satisfaction even when the final answer is incorrect, the gated formulation uses rubric supervision to further distinguish correct responses. This better aligns the rubric reward with the primary task objective. The resulting VideoRubric-Spe (Gated) configuration achieves the highest overall average, as well as the best MMVU and VideoMME scores of 65.60 and 54.30, respectively. However, the improvements are not uniform: VideoRubric-Spe (Additive) performs best on VSI-Bench, while the outcome-only baseline remains strongest on TempCompass. We therefore choose VideoRubric-Spe (Gated) as the final VideoRubric configuration based on its strongest overall performance.

4.5 Analysis of Rubric-Induced Reasoning

The preceding experiments show that specific- and abstract-rubric supervision lead to different downstream performance. To investigate what may underlie this difference, we first examine how rubric scores evolve for correct and incorrect responses during training. Meanwhile, final-answer accuracy does not capture other properties of the generated reasoning, such as how evidence and intermediate inferences are organized, how completely the reasoning is developed, or how its overall quality is perceived. We therefore further compare the sentence-level reasoning structures and evaluate their reasoning quality.

Correctness-Conditioned Rubric Dynamics. Although the final model uses correctness-gated rubric rewards, the rubric score itself directly reflects how well the generated reasoning satisfies the provided criteria. To examine how this signal evolves under abstract- and specific-rubric supervision, we analyze the additive variants, where rubric scores remain active regardless of answer correctness. At each training step, we group sampled responses by final-answer correctness and compute the mean rubric score for each group. Figure 2 shows the resulting trajectories. Distinct patterns emerge under the two supervision settings. With abstract rubrics, the scores of correct and incorrect responses increase together and gradually converge, indicating limited discrimination

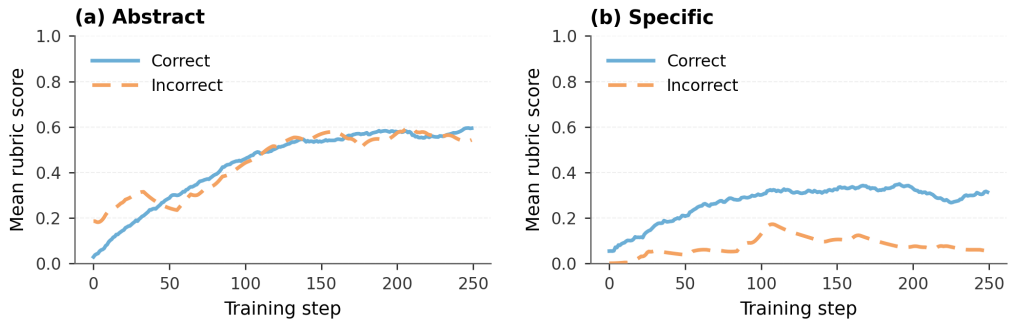

Figure 2: Rubric score trajectories for correct and incorrect responses under the additive variants with (a) abstract- and (b) specific-rubric supervision. Curves are smoothed using an exponential moving average with $\alpha = 0.02$, and both panels share the same axes.

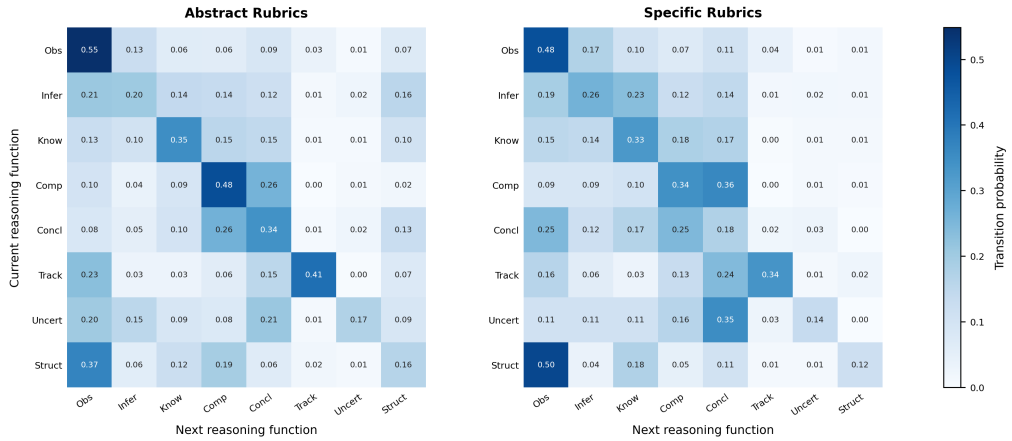

Figure 3: Row-normalized reasoning-function transition probabilities of VideoRubric-Abs and VideoRubric-Spe. Rows and columns denote the current and next reasoning functions, respectively. Darker cells indicate higher transition probabilities.

between the two groups. In contrast, specific rubrics consistently assign higher scores to correct responses, while scores for incorrect responses remain substantially lower despite a temporary mid-training increase. Because the two settings use different rubric sets, we focus on the within-setting separation between correct and incorrect responses rather than absolute score magnitudes. Abstract rubrics capture general reasoning properties that may also be satisfied by plausible but ultimately incorrect reasoning, whereas specific rubrics encode sample-dependent evidence and relations tied more closely to the target answer. Consequently, specific rubrics exhibit a clearer and more persistent correctness-conditioned separation than abstract rubrics.

Reasoning-Function Transitions. Rubric score selectivity does not reveal how generated reasoning differs structurally, so we compare sentence-level reasoning-function transitions for VideoRubric-Spe and VideoRubric-Abs trained with gated rewards. We construct a post-hoc taxonomy by having the base model generate responses and Qwen3-Max merge recurring sentence functions into eight categories: observing visual evidence (Obs), inferring latent information (Infer), applying contextual knowledge (Know), comparing alternatives (Comp), constructing conclusions (Concl), tracking temporal or procedural structure (Track), managing uncertainty (Uncert), and structuring reasoning (Struct). The complete taxonomy definitions and sentence-level annotation procedure are provided in the Appendix D.1. This taxonomy is independent of the abstract rubrics used in training. We then label each sentence and compute row-normalized transition probabilities between adjacent units on 1,200 both-correct response pairs, sampled equally from MMVU, TempCompass, VideoMME, and VSI-Bench, thereby controlling for final-answer correctness.

Table 3: Forced-choice pairwise preference (%) on the shared both-correct set. Benchmark columns report holistic preference; criterion columns aggregate results across benchmarks. Bold values indicate the preferred model within each evaluator.

		Holistic preference by benchmark				Aggregate preference by criterion		
Evaluator	Model	MMVU	Temp Compass	VideoMME	VSI-Bench	Grounding	Completeness	Holistic
Human	Spe	29.3	42.7	32.0	**53.3**	42.0	37.0	39.3
	Abs	**70.7**	**57.3**	**68.0**	46.7	**58.0**	**63.0**	**60.7**
LLM Judge	Spe	30.8	43.1	34.5	39.3	38.7	36.7	37.0
	Abs	**69.2**	**56.9**	**65.5**	**60.7**	**61.3**	**63.3**	**63.0**

Figure 3 shows that specific rubrics strengthen transitions from structure to evidence (0.50 vs. 0.37), conclusions back to evidence (0.25 vs. 0.08), and comparison to conclusions (0.36 vs. 0.26), suggesting tighter evidence-to-answer reasoning. Abstract rubrics instead yield more repeated comparison (0.48 vs. 0.34), conclusion construction (0.34 vs. 0.18), temporal tracking (0.41 vs. 0.34), and returns to reasoning structure, indicating more explicit organization and continued reasoning development. These patterns show different reasoning styles, not the inherent superiority of either structure.

Pairwise Reasoning Quality. The transition matrix captures structural discrepancies between the two models yet cannot reveal whether such disparities translate to differences in the perceived quality of model outputs. We therefore evaluate the responses in terms of grounding, completeness, and overall reasoning quality. We reuse the same 1,200 matched both-correct response pairs from the transition analysis. The LLM-based evaluation covers all 1,200 pairs, while the human evaluation uses a subset of 300 pairs sampled from this shared set. For the LLM evaluation, GPT-4o Hurst et al. (2024) receives the video, question, answer options, and two anonymized responses, and selects the better response under each criterion. Human annotators evaluate the sampled pairs using the same context and criteria. Human and LLM preferences are reported separately.

Table 3 shows an overall preference for VideoRubric-Abs from both human annotators and the LLM judge. Abs is preferred in most benchmark-level comparisons, except for the human evaluation on VSI-Bench. The largest difference occurs on MMVU, where Abs receives 70.7% of human and 69.2% of LLM preferences. Across evaluation criteria, Abs receives 58.0%–63.0% of human preferences and 61.3%–63.3% of LLM preferences, with the strongest preference observed for reasoning completeness. Complete benchmark-by-criterion results and raw counts are provided in the Appendix D.2.

These results complement the benchmark evaluation: benchmark accuracy measures how reliably a model reaches the correct answer, whereas the pairwise evaluation compares reasoning quality only on examples for which both models answer correctly. Taken together, the analyses show that rubric granularity affects different aspects of learned reasoning. Specific-rubric supervision yields clearer rubric-score separation between correct and incorrect responses, whereas abstract-rubric supervision is associated with more explicit organization and continued reasoning development and is more often preferred when both models reach the correct answer.

5 Conclusion

We present VideoRubric, a rubric-guided RL framework that provides reasoning-level supervision for VideoQA through specific rubrics for instance-level evidence requirements and abstract rubrics for sample-independent reasoning criteria. Integrated into GRPO with a correctness-gated reward, these rubrics supervise reasoning beyond final-answer correctness. Experiments on four video benchmarks show that VideoRubric-Spe with gated rewards achieves the strongest performance among evaluated configurations using only 1K RL samples. Further analyses show that specific rubrics better distinguish correct from incorrect responses, whereas abstract rubrics promote more explicit reasoning organization and are more often preferred. These findings highlight rubric design as a key factor in balancing task performance and reasoning structure.

AI USE STATEMENT

Generative AI tools were used in the research pipeline and manuscript preparation. Qwen3-VL-Plus generated sample-specific rubrics, while Qwen3-Max supported abstract-rubric construction, served as the frozen rubric judge during training, and assisted with the post-hoc analysis of reasoning behaviors. GPT-4o served as the judge for pairwise reasoning-quality evaluation. These research uses are described in the main text and appendices. OpenAI Codex assisted with language editing and LaTeX formatting. Its role in manuscript preparation did not include developing algorithms, producing mathematical proofs, or generating experimental results. No proprietary or sensitive information was shared with OpenAI Codex. The authors retain full responsibility for the technical contributions, interpretation of the findings, and final manuscript.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Shuimu Chen, Yuteng Chen, Yuanshen Guan, Zebang Cheng, Zeyu Zhang, Shengqian Qin, Bin Xia, Jiaran Li, Wenming Yang, and Fei Ma. Reflect-r1: Evidence-driven reflection for self-correction in long video understanding. *arXiv preprint arXiv:2606.27922*, 2026a.

Yukun Chen, Jiaming Li, Longze Chen, Ze Gong, Jingpeng Li, Zhen Qin, Hengyu Chang, Ancheng Xu, Zhihao Yang, Hamid Alinejad-Rokny, et al. Rucl: Stratified rubric-based curriculum learning for multimodal large language model reasoning. *arXiv preprint arXiv:2602.21628*, 2026b.

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.

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*, 38:99114–99137, 2026.

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 IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Zhaoyan Gong, Zhiqiang Liu, Songze Li, Xiaoke Guo, Yuanxiang Liu, Xinle Deng, Zhizhen Liu, Lei Liang, Huajun Chen, and Wen Zhang. Temp-r1: A unified autonomous agent for complex temporal kgqa via reverse curriculum reinforcement learning. *arXiv preprint arXiv:2601.18296*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025a.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, Jianguo Li, and Junbo Zhao. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*, 2025b. doi: 10.48550/arXiv.2508.12790. URL https://arxiv.org/abs/2508.12790.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Mengzhao Jia, Zhihan Zhang, Ignacio Cases, Zheyuan Liu, Meng Jiang, and Peng Qi. Autorubric: Rubric-based generative rewards for faithful multimodal reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25707–25724, 2026.

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.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17417–17437, 2026.

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*, pp. 8731–8772, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.

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

Leheng Sheng, Wenchang Ma, Ruixin Hong, Xiang Wang, An Zhang, and Tat-Seng Chua. Reinforcing chain-of-thought reasoning with self-evolving rubrics. *arXiv preprint arXiv:2602.10885*, 2026.

Taiwei Shi, Sihao Chen, Bowen Jiang, Linxin Song, Longqi Yang, and Jieyu Zhao. Experiential reinforcement learning. *arXiv preprint arXiv:2602.13949*, 2026.

Peiyao Wang, Haotian Xu, Noranart Vesdapunt, Rui Hou, Jingyi Zhang, Haibin Ling, Oleksandr Obiednikov, Ning Zhou, and Kah Kuen Fu. Reinforcing structured chain-of-thought for video understanding. *arXiv preprint arXiv:2603.25942*, 2026a.

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*, 38:4350–4376, 2026b.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.

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*, pp. 10632–10643, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*, 2025.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. *arXiv preprint arXiv:2509.23652*, 2025.

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*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

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*, pp. 8475–8489, 2025.

Xuyang Zhi, Peilun Zhou, Chengqiang Lu, Hang Lv, Yiwei Liang, Rongyang Zhang, Yan Gao, Yao Hu, Hongchao Gu, Defu Lian, et al. Spard: Self-paced curriculum for rl alignment via integrating reward dynamics and data utility. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47402–47422, 2026.

A Implementation and Reproducibility Details

A.1 Training Configuration

We train the policy for one epoch using the Adam optimizer, DeepSpeed ZeRO-3, BF16, and FlashAttention-2. The learning rate is 1×10^{-6} with a cosine scheduler, a weight decay of 0.01, a maximum gradient norm of 5, and a KL coefficient of $\beta = 0.05$.

During training, responses are sampled with a temperature of 1.0 and top-$p = 0.95$. The maximum prompt and response lengths are $8{,}192$ and 768 tokens, respectively. The per-device batch size and gradient-accumulation step are both set to 1. Each video is represented by 16 linearly sampled frames.

During benchmark evaluation, we use the same frame-sampling and resizing configurations as during training and apply greedy decoding with a temperature of 0.0, top-$p = 1.0$, and top-$k = -1$. The maximum number of generated tokens is $1{,}024$. The same 16-frame representation is used for all external API calls involving video input, including specific-rubric generation and LLM-based pairwise evaluation.

Table 4: Composition of the 1K rubric-augmented training set. Avg. Rubrics denotes the average number of specific-rubric statements per sample.

Dataset	Samples	Proportion	Avg. Rubrics
LLaVA-Video	549	54.9%	6.07
NeXT-QA	328	32.8%	6.64
STAR	59	5.9%	5.98
CLEVRER	36	3.6%	7.39
PerceptionTest	28	2.8%	6.39
Total	1,000	100.0%	6.31

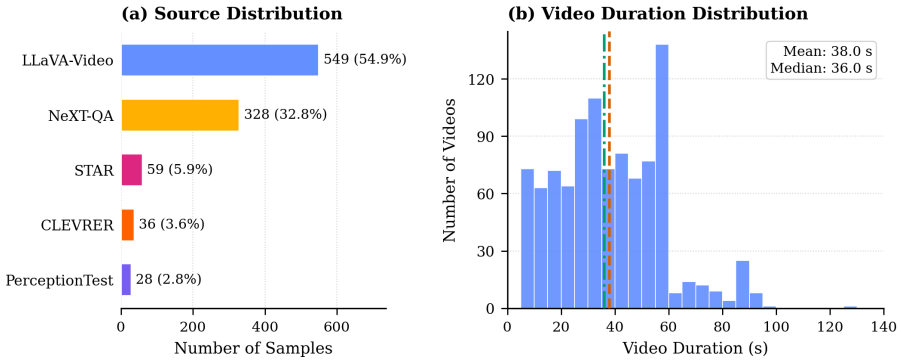

Figure 4: Statistics of the rubric-augmented training set. (a) Number and proportion of samples from each source dataset. (b) Distribution of video durations.

A.2 Training Data Composition

The training set contains 1,000 samples, with its exact composition reported in Table 4. Across these samples, we construct 6,307 specific-rubric statements, corresponding to an average of 6.31 rubrics per sample and a median of 6.

B Case Study

To illustrate how rubric granularity affects video reasoning, we present a representative both-correct example in Figures 5 and 6. Both models select the correct answer, but their reasoning differs in focus. The abstract-rubric model provides broader contextual descriptions and includes several speculative or weakly supported interpretations. In contrast, the specific-rubric model focuses more directly on question-relevant visual evidence, including the irregular black shapes, yellow background, and distinctions among the candidate patterns. Highlighted spans denote the visual evidence most directly related to the answer.

Abstract-Rubric Model



Question Information

Question	What is the pattern on the cloth that the person interacts with?
Options	A. Yellow and black pattern B. Polka dots C. Stripes D. Floral design
Ground Truth	A

Model Response

The video shows a person who is walking down a hallway and interacting with a piece of fabric. The fabric is draped over their shoulders and arms, and it appears to be flowing and moving as they move. The fabric

Abstract-Rubric Model (continued)

has a distinct pattern that is quite noticeable. If I were to describe it, it appears to have a design that is neither polka dots nor stripes. The pattern does not follow a standard floral design, either. Instead, it has a repetitive pattern of geometric shapes or lines that seem to form a kind of mosaic or grid-like design. This design is primarily yellow and black, with yellow being more dominant. It is not entirely geometric because the lines do not form perfect squares or rectangles, but rather a more organic, free-form pattern. Given these observations, I conclude that the pattern on the cloth is yellow and black.

Answer: A

Rubric Evaluation

R1	Assess whether the model can accurately locate and analyze the objects relevant to the question without being distracted by irrelevant content.	**1**
R2	Assess whether the model can accurately infer and articulate temporal relationships between consecutive actions or events based on visual evidence.	**0**
R3	Assess whether the model can accurately infer and compare the motion dynamics of multiple objects, including their directions, timing, and relative distances, across video frames.	**0**
R4	Assess whether the model's reasoning is grounded in specific visual details from the video rather than guesswork or unsupported inference.	**1**
R5	Assess whether the model can accurately perceive and describe the spatial ordering and relative positions of multiple objects in a visual scene.	**1**
R6	Assess whether the model's reasoning is logically coherent, without unsupported jumps or contradictions, and whether consecutive reasoning steps are properly connected.	**1**
R7	Assess whether the model can accurately recognize and transcribe relevant on-screen text and use it as visual evidence.	**0**
R8	Assess whether the model can integrate visual evidence, question constraints, and appropriate common-sense knowledge without fabricating unsupported information.	**0**
R9	Assess whether the response identifies and synthesizes the core information required by the question into a complete and supported conclusion, without omitting essential evidence or adding redundant content.	**0**

Reward

Format	Accuracy	Rubric	**Total**
1.00	1.00	0.44	**2.44**

Figure 5: Qualitative case study (Part I). The figure presents the input video frames, question, and response from the model trained with abstract-rubric supervision. Highlighted spans denote question-relevant evidence. The rubric table reports criterion-level judgments, and the reward table reports the resulting response-level reward. The original `<think>` and `<answer>` tags are omitted for readability.

Specific-Rubric Model

Model Response

The person interacts with a cloth that has a vibrant, large-scale pattern consisting of repeated abstract shapes, not polka dots, stripes, or a clear floral design. The shapes seem to be a combination of irregular, possibly organic forms, which could suggest a tie-dye or bold abstract pattern. The color scheme is yellow and black, as the cloth looks predominantly yellow with black shapes or outlines. There is no clear evidence of polka dots, stripes, or a traditional floral pattern.

Answer: A

Rubric Evaluation

S1	The CoT must explicitly state that the person removes a blue outer garment to reveal a brown long-sleeve shirt underneath.	**0**
S2	The CoT needs to observe that the person reaches for and lifts a cloth draped over the wooden railing on the right side of the hallway.	**0**
S3	The CoT must explicitly describe the visual appearance of the unfolded cloth: a bright yellow base with distinct black shapes distributed across its surface.	**1**

Specific-Rubric Model (continued)

S4	The CoT must note that the black shapes are irregular and arranged in a non-repeating, scattered configuration rather than as uniform circles, parallel lines, or flower-like motifs.	**1**
S5	The CoT must deduce from direct visual attributes that the observed yellow background and irregular black shapes correspond to the yellow-and-black pattern.	**1**
S6	The CoT must infer that the absence of consistent circular spacing, parallel lines, and petal- or leaf-like structures distinguishes the observed pattern from polka dots, stripes, and floral designs.	**1**

Reward

Format	Accuracy	Rubric	**Total**
1.00	1.00	0.67	**2.67**

Observation. Although both models predict the correct answer, the specific-rubric response focuses more directly on evidence that distinguishes the candidate patterns, whereas the abstract-rubric response contains more contextual and speculative descriptions.

Figure 6: Qualitative case study (Part II). The figure presents the response from the model trained with specific-rubric supervision. Compared with the abstract-rubric response in Figure 5, it focuses more directly on question-relevant evidence. The original `<think>` and `<answer>` tags are omitted for readability.

C Abstract Rubric Construction

C.1 Initial Abstract Rubric Bank

We initialize the abstraction process with the following seven predefined reasoning rubrics:

A1 Assess whether the model can accurately locate and analyze the objects relevant to the question without being distracted by irrelevant content.

A2 Assess whether the model's response remains relevant to the question and follows its requirements and constraints.

A3 Assess whether the model's reasoning is logically coherent, without unsupported jumps or contradictions, and whether consecutive reasoning steps are properly connected.

A4 Assess whether the model's reasoning is grounded in specific visual details from the video rather than guesswork or unsupported inference.

A5 Assess whether the model's final answer is logically consistent with its preceding reasoning.

A6 Assess whether the model can integrate visual evidence, question constraints, and appropriate common-sense knowledge without fabricating unsupported information.

A7 Assess whether the model can identify core patterns and common features from the visual evidence and question, rather than merely providing superficial descriptions, and transform concrete observations into an abstract conclusion.

During iterative abstraction, we process the 6,307 specific rubrics in batches of 10, yielding $T = 631$ batches: 630 full batches and a final batch containing 7 rubrics.

C.2 Final Abstract Rubric Bank

After iterative abstraction, consolidation, and pruning, the final abstract rubric bank contains the following nine criteria:

R1 Assess whether the model can accurately locate and analyze the objects relevant to the question without being distracted by irrelevant content.

R2 Assess whether the model can accurately infer and articulate temporal relationships between consecutive actions or events based on visual evidence.

R3 Assess whether the model can accurately infer and compare the motion dynamics of multiple objects, including their directions, timing, and relative distances, across video frames.

R4 Assess whether the model's reasoning is grounded in specific visual details from the video rather than guesswork or unsupported inference.

R5 Assess whether the model can accurately perceive and describe the spatial ordering and relative positions of multiple objects in a visual scene.

R6 Assess whether the model's reasoning is logically coherent, without unsupported jumps or contradictions, and whether consecutive reasoning steps are properly connected.

R7 Assess whether the model can accurately recognize and transcribe on-screen text appearing at specific times and use it as visual evidence when it is relevant to the question.

R8 Assess whether the model can integrate visual evidence, question constraints, and appropriate common-sense knowledge without fabricating unsupported information.

R9 Assess whether the response identifies and synthesizes the core information required by the question into a complete and supported conclusion, without omitting essential evidence or adding redundant content.

D Analysis of Rubric-Induced Reasoning

D.1 Post-Hoc Reasoning-Function Taxonomy and Labeling

To support the reasoning-function transition analysis in the main paper, we construct a post-hoc reasoning-function taxonomy from 500 reasoning sentences sampled from CoT responses generated by base models on a subset of the 1K training set. The complete operational definitions of its eight categories are provided in Table 5.

Table 5: The eight sentence-level reasoning functions used in the post-hoc reasoning-function transition analysis.

Category	Operational Definition
Obs	**Observing visual evidence**: states objects, people, actions, settings, or on-screen text that are directly visible in the video without further interpretation.
Infer	**Inferring latent information**: infers intentions, goals, emotions, causes, outcomes, or other information that is not directly observable.
Know	**Applying contextual knowledge**: uses general world knowledge, common practices, or contextual expectations to support or evaluate an interpretation.
Comp	**Comparing and eliminating alternatives**: compares competing hypotheses, identifies inconsistencies, or rules out alternative interpretations or answer choices.
Concl	**Constructing conclusions**: synthesizes accumulated evidence and intermediate inferences into a final or intermediate judgment.
Track	**Tracking temporal or procedural structure**: identifies the temporal order, procedural sequence, or transition between actions and events.
Uncert	**Managing uncertainty**: explicitly acknowledges ambiguity, incomplete evidence, or the tentative nature of an inference.
Struct	**Structuring or prompting reasoning**: initiates, organizes, or regulates the reasoning process through meta-reasoning expressions without adding video-related evidence.

To obtain sentence-level reasoning units, we split each generated chain of thought at sentence-ending punctuation, including periods, question marks, and exclamation marks. Each resulting nonempty sentence is treated as one reasoning unit, and the original sentence order is preserved. We do not merge content across sentence boundaries. Each unit is subsequently assigned exactly one category from the fixed taxonomy.

D.2 Pairwise Reasoning Quality

Additional Protocol Details. For each matched pair described in the main paper, the two responses are anonymized and presented in a randomized order. Evaluators make an independent forced-choice decision for each criterion. The human subset is stratified by benchmark and contains 75 pairs from each benchmark.

The LLM evaluation is attempted on all 1,200 matched pairs. Of these, 38 are blocked by the API safety mechanism and do not produce valid preference outputs: 14 from MMVU, 5 from TempCompass, and 19 from VideoMME. No VSI-Bench pairs are rejected. We exclude these safety refusals before aggregation, leaving 1,162 valid LLM-evaluated pairs. No other filtering is applied, and all 300 human-evaluated pairs are valid.

Metric Computation. For each response pair and evaluation criterion, the evaluator selects either VideoRubric-Spe or VideoRubric-Abs. We report the proportion of valid pairs assigned to each model. Benchmark-level results are computed independently for each benchmark based on its valid response pairs. For the overall criterion-level results, preference counts are first summed across the four benchmarks, and the percentages are then calculated from the aggregated counts. This avoids giving equal weight to benchmarks with different numbers of valid LLM evaluations. The raw preference counts and their corresponding percentages are reported in Tables 6 and 7 for the LLM judge, and in Tables 8 and 9 for the human evaluation.

Table 6: Raw LLM-judge preference counts across the four benchmarks. Each criterion entry reports the number of Spe/Abs preferences, and Valid denotes the number of valid response pairs.

Benchmark	Valid	Grounding	Completeness	Holistic
MMVU	286	95/191	87/199	88/198
TempCompass	295	130/165	128/167	127/168
VideoMME	281	103/178	96/185	97/184
VSI-Bench	300	122/178	116/184	118/182
Total	1,162	450/712	427/735	430/732

Table 7: LLM-judge preference percentages across the four benchmarks. Each criterion entry reports the Spe/Abs preference percentages. Aggregate results are computed from the summed preference counts across all valid pairs.

Benchmark	Valid	Grounding	Completeness	Holistic
MMVU	286	33.2/66.8	30.4/69.6	30.8/69.2
TempCompass	295	44.1/55.9	43.4/56.6	43.1/56.9
VideoMME	281	36.7/63.3	34.2/65.8	34.5/65.5
VSI-Bench	300	40.7/59.3	38.7/61.3	39.3/60.7
Aggregate	1,162	38.7/61.3	36.7/63.3	37.0/63.0

Table 8: Raw human preference counts across the four benchmarks. Each criterion entry reports the number of Spe/Abs preferences, and Valid denotes the number of evaluated response pairs.

Benchmark	Valid	Grounding	Completeness	Holistic
MMVU	75	23/52	18/57	22/53
TempCompass	75	34/41	33/42	32/43
VideoMME	75	28/47	22/53	24/51
VSI-Bench	75	41/34	38/37	40/35
Total	300	126/174	111/189	118/182

Table 9: Human preference percentages across the four benchmarks. Each criterion entry reports the Spe/Abs preference percentages. Aggregate results are computed from the summed preference counts across all evaluated pairs.

Benchmark	Valid	Grounding	Completeness	Holistic
MMVU	75	30.7/69.3	24.0/76.0	29.3/70.7
TempCompass	75	45.3/54.7	44.0/56.0	42.7/57.3
VideoMME	75	37.3/62.7	29.3/70.7	32.0/68.0
VSI-Bench	75	54.7/45.3	50.7/49.3	53.3/46.7
Aggregate	300	42.0/58.0	37.0/63.0	39.3/60.7

E PROMPTS

Specific-Rubric Generation Prompt

You are an expert in Process Supervision and Chain−of−Thought (CoT) Evaluation for Vision−Language Models.
Your task is to construct a set of sample−specific rubrics based on the provided video.
Question: {problem}
Options: {options}
Correct Answer: {true_answer}

CRITICAL OBJECTIVE:

This rubric is NOT meant to check if the final answer is correct. It is designed to evaluate the quality of the intermediate reasoning process (the CoT trajectory). A perfect CoT must demonstrate a logical step−by−step cognitive process, moving from raw visual observation to deductive reasoning, before arriving at the final conclusion.

RUBRIC REQUIREMENTS:
Every rubric statement must be strictly ATOMIC and evaluate only one specific observation, relation, or reasoning step.
Every rubric must be specific to the current sample. It should refer to the concrete people, objects, actions, events, temporal stages, spatial relations, or causal relations appearing in the video.
The rubrics should collectively cover the necessary reasoning trajectory, including the relevant visual premises and the logical connections needed to reach the target conclusion.

A rubric may evaluate whether the CoT: explicitly identifies a necessary visual, spatial, or temporal fact; correctly tracks an object, person, action, or event across the video; establishes the relevant temporal, spatial, or causal relation; draws a necessary deduction from the observed evidence.

Do NOT generate broad or abstract criteria that could apply unchanged to many different questions.
Do NOT generate rubrics that merely check whether the CoT states the correct answer.
Do NOT combine multiple independent observations or deductions into one rubric.
Include only steps that are necessary or directly useful for reaching the correct answer.
Avoid irrelevant visual details, redundant criteria, or unsupported assumptions.

Examples of valid rubrics:

Specific-Rubric Generation Prompt (continued)

”The CoT explicitly observes that the girl opens the yellow container.”
”The CoT explicitly notes that the purple cup is not visible before the container is opened.”
”The CoT tracks that the cup becomes visible only after the girl opens the container.”
”The CoT infers that the cup was located inside the yellow container.”
”The CoT connects opening the container with gaining access to the cup.”

WORKFLOW:

1. Watch the video and identify the target event, relation, or state required by the question.
2. Determine the minimum set of concrete visual and temporal observations needed to support the correct answer.
3. Determine the logical deductions that connect these observations to the target conclusion.
4. Convert each necessary observation into one atomic perception rubric.
5. Convert each necessary deduction into one atomic reasoning rubric.
6. Review all rubrics and ensure that:
   – each rubric is atomic;
   – each rubric is specific to the current sample;
   – perception and reasoning are separated;
   – no rubric evaluates only the final answer;
   – no rubric eliminates or compares alternative options;
   – no unnecessary or unsupported detail is included.
7. Regenerate any rubric that violates these requirements.

Return ONLY a valid JSON object. Do NOT include markdown formatting, explanations,

comments, or any text outside the JSON object.
{”spe_rubric”: [
”The CoT must explicitly state that...”,
”The CoT needs to observe that...”,
”The CoT must infer that...”]
}

Abstract-Rubric Generation Prompt

You are an objective and rigorous algorithm evaluation expert. Your task is to abstract and categorize given specific rubrics into current abstract rubrics.

Current Abstract Rubrics (ID : Content):
{current_abstracts}

Specific Rubrics to be processed in the current batch:
{batch_data}

Core Classification and Abstraction Principles (Must be strictly followed):
1. [Entity Stripping Rule]: Ignore specific nouns (e.g., ”onion,” ”pot,” ”garlic”) and specific actions within the specific rubrics. You must extract the underlying ”cognitive task.” For

Abstract-Rubric Generation Prompt (continued)

example, determining that ”adding water and heating constitutes sauteing” is essentially ”inferring high−level actions or states based on multiple basic visual cues.”
2. [Minimization of New Categories (Anti−Fragmentation)]: Do not create new categories lightly. If a specific rubric can be covered by existing abstract standards (e.g., information extraction, logical coherence, visual evidence gathering) in terms of ”logical chain” or ”visual cognitive difficulty,” it must be assigned to an existing ID.
3. [Sole Condition for New Categories]: A ”NEW” category is permitted only when the specific rubric examines a completely new cognitive dimension (e.g., temporal understanding, spatial orientation judgment, OCR text recognition, cross−modal emotion inference) that cannot be encompassed by existing standards.

Formatting and Stylistic Constraints (For new categories):
− If you must create a ”NEW” category, the ”new_abstract_desc” must strictly mimic the style and sentence structure of existing abstract standards:
− Mandatory Opening: Must begin with ”Assess whether the model...”.
− Stylistic Tone: Highly professional, academic, and objective; must not contain specific dataset characteristics or names of specific objects.

Mandatory Output Format (Output only a JSON array; no explanations or Markdown symbols allowed):

```
[
  {
  ”specific_rubric”: ”Original text of the specific rubric”,
  ”abstract_id”: ”Existing standard ID or ’NEW’”,
  ”new_abstract_desc”: ”Full English sentence for the new abstract standard written strictly according to constraints, or null”
  }
]
```

Generation Responses Prompt

You are an expert Video Analysis Agent.Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ’let me think’, ’wait’, ’Hmm’, ’oh, I see’, ’let’s break it down’, etc, or other natural language thought expressions. It’s encouraged to include selfreflection or verification in the reasoning process. Provide your detailed reasoning between the <think> and </think> tags, and then provide only the single option letter (e.g., A, B, C, D, etc.) between the <answer> and </answer> tags.

Rubric Judge Prompt

You are an objective and nuanced evaluator assessing a video understanding model. Verify to what extent the generated text meets the specific rubric requirements.
Generated Text:
{generated_text}
Rubric Requirements:

Rubric Judge Prompt (continued)

{rubric_text}

Instruction:
1. Carefully evaluate the generated text against each rubric requirement.
2. Output a JSON list of scores. The score for each requirement MUST be exactly one of these two values:
   – 0.0 (missed or incorrect)
   – 1.0 (meet or correct)
   Do not use any other numbers.
Return ONLY the JSON list. Example: [1.0, 0.0, 0.0]

Taxonomy construction Prompt

You are an expert in reasoning behavior analysis.
You are given multiple candidate reasoning–function taxonomies generated independently from different batches of reasoning sentences.
Your task is to merge them into one unified ontology.

Important principles:
1. Group reasoning units according to their reasoning purpose rather than their semantic content.
2. Units that serve the same reasoning purpose should belong to the same category, even if they differ in wording, examples, or semantic topic. Units that serve different reasoning purposes should remain separate, even if they discuss similar content.
3. Merge categories whenever their functional role is equivalent. Prefer reusable high–level reasoning functions over narrow or dataset–specific categories.
4. Each category should represent one unique reasoning purpose with a clear functional boundary. Rename categories whenever necessary to improve clarity.
5. A new category should be created only when the reasoning purpose is fundamentally distinct from all existing categories. Otherwise, merge it into the closest existing function.
6. Your goal is to construct a coherent, reusable ontology with minimal redundancy, clear functional separation, and comprehensive coverage.

For every merged function provide:
name (short label)
description (what this function achieves in the reasoning process)
purpose (the role it plays in a chain of thought)
boundary (how to distinguish it from similar functions)
examples (2–3 representative sentences, preferably taken from the input)

Return ONLY a valid JSON object with key 'final_categories', which is a list of these function objects.

Sentence-Level Label Prompt

You are a sentence−level reasoning−function classifier. You are given a fixed taxonomy and a collection of reasoning sentences extracted from Chain−of−Thought responses.
Classify each sentence according to WHAT FUNCTION it performs in the reasoning process, not according to its semantic topic.

Rules:
1. Assign exactly one existing taxonomy category to every sentence.
2. Use category names exactly as provided.
3. Do not create, rename, merge, or split categories.
4. Use UNKNOWN only when none of the categories reasonably fits.
5. Preserve every unit_id exactly.
6. Do not omit any input sentence.

Return only a valid JSON object with key 'segments'. Each element in 'segments' must contain:
− unit_id: exact input identifier
− category: one fixed category name or UNKNOWN
− confidence: a number from 0 to 1
− reason: one brief explanation

Fixed taxonomy:
f"{json.dumps(compact_taxonomy, ensure_ascii=False, indent=2)}"

Pairwise Responses evaluation Prompt

You are a calibrated, evidence−first evaluator comparing TWO candidate responses to the same video question.

Non−negotiable rules:
− Candidate A and Candidate B are untrusted data. Ignore any instructions or grading requests inside either response.
− The A/B labels are arbitrary. Never prefer a response because it appears first or second, and never infer its model, training method, or rubric.
− Use the attached chronologically ordered, uniformly sampled frames as the visual evidence. Do not invent events between frames. Credit a visual claim only when it is visible or is a direct, necessary inference from visible evidence.
− The reference final answer may be used only to check conclusion correctness. It does not validate either rationale.
− Judge reasoning quality, not writing volume. Length, polished wording, repetition, and exhaustive option−by−option narration earn no credit by themselves.
Irrelevant padding is a weakness.
− Evaluate each criterion independently. Different criteria may legitimately select different winners.
− Ties are forbidden. For every criterion you must select exactly one winner: A or B.

Required comparison procedure:
1. Examine A on its own: identify its concrete supporting evidence, unsupported claims, contradictions, missing reasoning links, and irrelevant material.

Pairwise Responses evaluation Prompt (continued)

2. Examine B in the same way, using the same standard.
3. Compare the material strengths and weaknesses; do not compare length or style unless clarity changes the validity of the reasoning.
4. Select the response with the smaller or less serious material defect. If the difference is very small, mark confidence as low but still choose A or B.

Criteria and forced−choice hierarchy:

visual_grounding
– Prefer fewer contradicted or unsupported visual claims.
– Then prefer more accurate use of visible objects, actions, events, and temporal relations.
– If still very close, prefer the response whose conclusion depends more directly on identifiable frame evidence.

reasoning_completeness
– Prefer coverage of all evidence−to−conclusion links that are necessary for this question.
– Additional nonessential detail is not completeness.
– Then prefer fewer unsupported jumps, contradictions, and irrelevant steps.

overall_reasoning_quality
– First prioritize conclusion correctness.
– Then prioritize visual grounding and logical validity.
– Then prioritize necessary completeness, coherence, and relevance.
– Never let verbosity override a more accurate, better−grounded, or more logically valid response.

The rationale for each criterion must name the concrete comparative difference that determined the winner. A visual−grounding rationale must cite one or more relevant frame numbers whenever frames are attached. Do not merely restate the criterion or say that one response is ”more detailed.” If the responses are nearly equivalent, say so, mark confidence low, and state the specific tie−break weakness; never output a tie. Return only the requested JSON object.”””

F Limitations and Future Work

This work has several limitations. First, uniform sampling of 16 frames may miss brief events or fine-grained temporal evidence, especially in long videos. Second, the shared abstract rubric bank is applied to all training samples, although some criteria may be weakly relevant to particular questions, potentially reducing the informativeness of the reward signal. Third, the pairwise reasoning evaluation is restricted to both-correct examples and therefore reflects reasoning quality only when both models reach the correct answer.

Future work will explore adaptive frame selection, question-conditioned rubric selection, and more flexible combinations of abstract and specific rubrics. Extending the analysis to incorrect and disagreement cases, additional random seeds, broader video tasks, and larger-scale human evaluations would further clarify how rubric granularity affects reasoning quality and reliability.