

Let VLMs Grade Their Own Thoughts: A Self-Quantification Approach to Reasoning-Aware Reward Modeling

Xing Xi^{1,2} Yu Qiu¹ Ronghua Luo^{1*} Peixian Chen^{2*} peilin tong²
¹South China University of Technology ²Ant Group

Abstract

Chain-of-Thought (CoT) is a key technique for enhancing the reasoning capabilities of Vision Language Models (VLMs). Existing methods often employ Reinforcement Learning (RL) with external constraints to align the model’s reasoning process with human cognitive patterns. However, we argue that the model’s intrinsic reasoning paths may differ from human cognition, and that forcing such alignment can constrain the model’s potential and even degrade its performance. To address this, we propose leveraging the model’s intrinsic self-evaluation to guide its optimization. We hypothesize that a model’s self-generated confidence scores are effective indicators of its reasoning quality. Based on this evaluation metric, we design two novel reward functions: (1) Sequential Confidence Rigorous Evaluation (SCRE) for challenging problems that demand strict logical reasoning, and (2) Intra-group Score Re-ranking (IGSR) for general-purpose, open-ended scenarios. We name our method Video-RAISE (Reasoning Alignment through Intrinsic Self-Evaluation). Comprehensive experiments on six video understanding benchmarks demonstrate that Video-RAISE achieves state-of-the-art (SOTA) performance, significantly outperforming previous methods and even proprietary models, e.g. GPT-4o. For instance, on the VideoMMM benchmark, our Video-RAISE achieves a new SOTA accuracy of 52.8%, outperforming the previous best model by a significant 3.0%. In addition, our method achieves a reasoning path consistency of 90%, which is double that of the Qwen2.5-VL-Instruct and even surpasses the performance of supervised fine-tuning.

1. Introduction

Video understanding seeks to enable machines to comprehend complex video streams by analyzing not only static objects and scenes but also their temporal dynamics, such as actions, interactions, and causal relationships. This capacity for in-depth spatio-temporal reasoning makes video under-

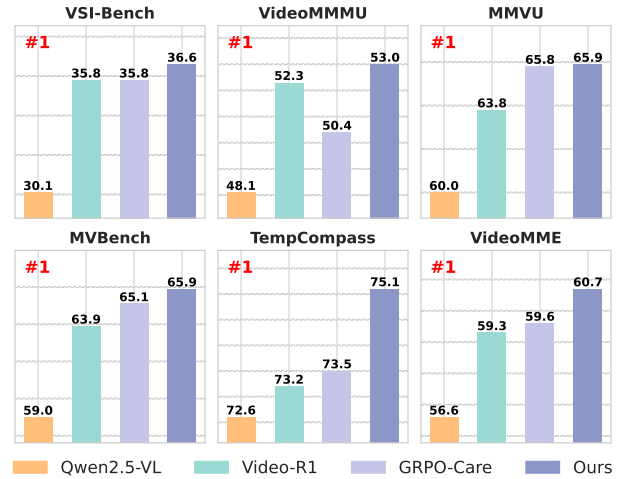


Figure 1. **Performance comparison on six video understanding benchmarks.** Our method, Video-RAISE, achieves SOTA performance across all benchmarks. All methods are evaluated at 32 FPS. More results and analysis are provided in experiment section.

standing fundamental to a wide array of applications, including surveillance [20, 31, 32], recommendation systems [5, 15, 22], and human-computer interaction [14, 25, 26].

The paradigm of post-training with Reinforcement Learning (RL) has recently emerged as a powerful technique to unlock the reasoning capabilities of foundation models. Pioneered in Large Language Models (LLMs) with methods like Group Relative Policy Optimization (GRPO) [9], this approach was rapidly and successfully extended to multi-modal domains [24, 28]. In video understanding, this trend continued with methods like Video-R1 [11], which enforces temporal logic, and GRPO-Care [7], which aligns reasoning with high-quality rationales. However, a common thread unites these advancements: they all rely on imposing external, human-defined constraints to guide the model. Whether enforcing a correct temporal order or rewarding alignment with prescribed reasoning steps, these methods are fundamentally anchored to the assumption that optimal performance is achieved by mimicking human cognitive patterns.

*Corresponding author. Ronghua Luo: rhluo@scut.edu.cn; Peixian Chen: pchenac@connect.ust.hk

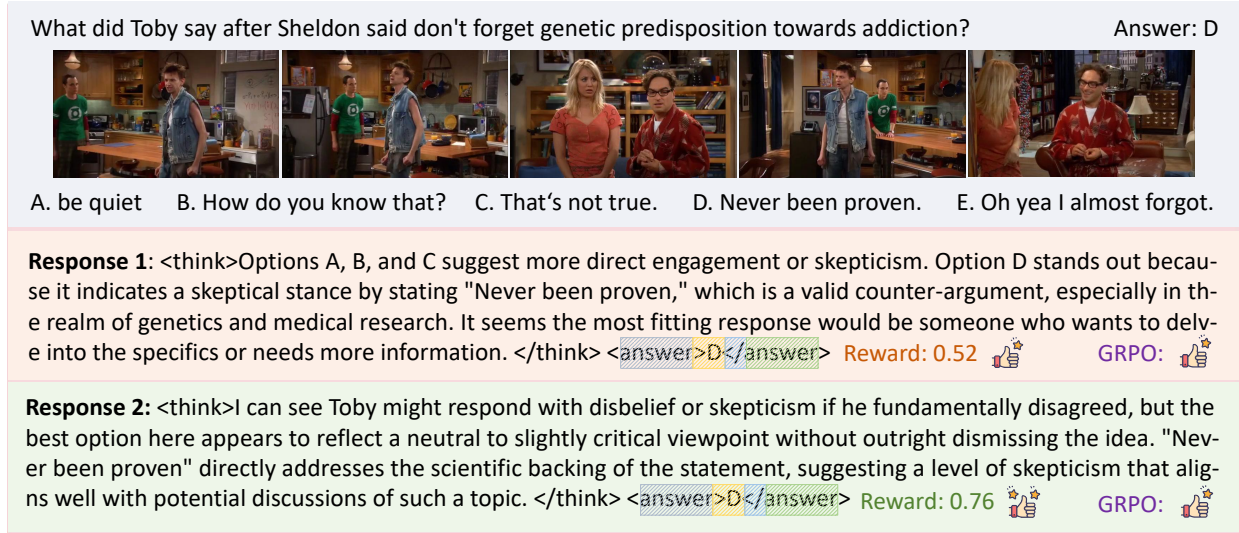


Figure 2. **Illustration of response quality evaluation.** GRPO’s evaluation based on textual accuracy yields a sparse 0/1 reward, offering a weak learning signal. Our method, Video-RAISE, instead constructs a continuous reward from answer token probabilities, which allows for a more fine-grained distinction between high- and low-quality reasoning for more effective training.

We contend that this forced alignment is a critical bottleneck. By constraining the model to human-like reasoning paths, such approaches may prevent the discovery of more effective, non-human-like reasoning strategies and can even degrade performance. In this work, we challenge this prevailing paradigm. We introduce an approach that shifts the focus from external alignment to intrinsic self-evaluation, empowering the model to discover its own optimal reasoning paths. As evidenced in Figure 1, our method, Video-RAISE, establishes a new state-of-the-art across six video understanding benchmarks.

To enable model self-evaluation, we hypothesize that a high-quality reasoning path allows the model to exhibit higher confidence when generating the final answer. We quantify this confidence using the probabilities of the answer tokens and convert it into a reward signal for RL. Recognizing that different problems require different evaluation criteria, we categorize tasks into two types: strict logical reasoning and general-purpose scenarios. For the former, which involves problems with unique answers requiring exact matches, we propose Sequential Confidence Rigorous Evaluation (SCRE). SCRE operates on two dimensions: it penalizes uncertainty by considering the lowest token probability in the sequence, and it assesses overall fluency and confidence through the entire probability distribution. For the latter, which focuses more on semantic quality, we introduce Intragroup Score Re-ranking (IGSR). IGSR first performs an initial grouping of multiple generated candidates using text-based metrics to ensure that samples within each group have similar semantic quality. Subsequently, within each semantically similar group, IGSR allocates a higher

reward to the candidate for which the model shows higher intrinsic confidence (i.e., a higher answer probability). As illustrated in Figure 2, in contrast to the standard GRPO which relies solely on the binary correctness of the final answer, our approach transforms model’s intrinsic confidence into a continuous learning signal. This enables a more fine-grained optimization of the VLM’s reasoning process.

To evaluate our proposed Video-RAISE, we conduct experiments on six mainstream video understanding benchmarks following previous methods [7, 11]. Video-RAISE demonstrates leading performance across all benchmarks. Notably, its advantage is more pronounced under the more demanding 16 Frames setting, which requires a stronger temporal understanding from the model due to fewer sampled frames. For instance, on the VideoMMU benchmark, Video-RAISE surpasses the previous SOTA by a margin of 3.0 % in accuracy. The main contributions of this paper are summarized as follows:

- We propose an optimization paradigm for VLMs that leverages intrinsic evaluation of its reasoning paths, instead of external constraints, to guide learning process.
- We propose two reward strategies tailored for different problem types, namely, SCRE and IGSR, which convert the model’s intrinsic confidence into a continuous and fine-grained learning signal to guide the optimization.
- Validated on six benchmarks, our method Video-RAISE sets a new state-of-the-art in video understanding. It not only excels in accuracy but also achieves a remarkable 90% reasoning consistency, a result surpassing even SFT.

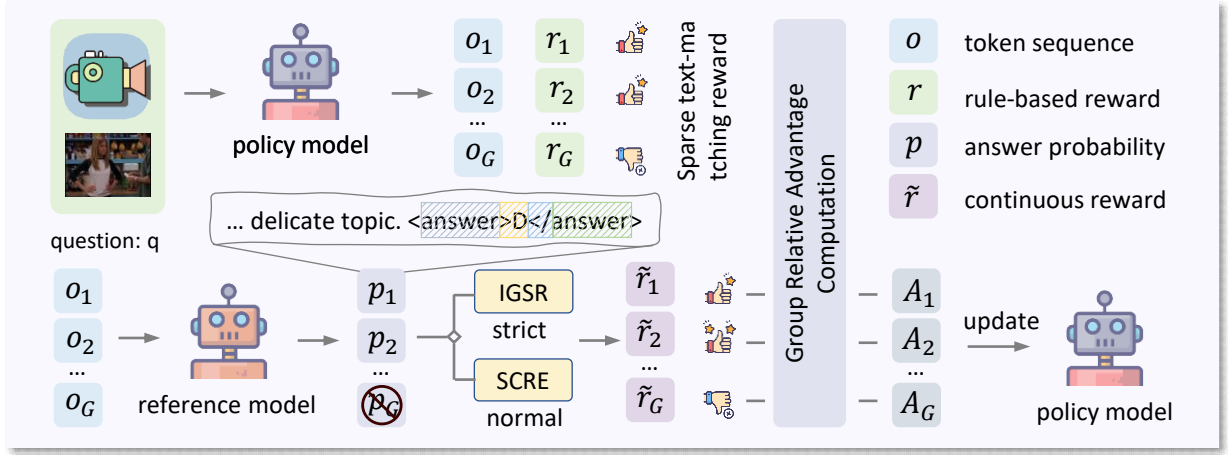


Figure 3. **An illustration of the training pipeline for our Video-RAISE.** For a given video-question input, the model first samples a set of candidate responses. Subsequently, it computes an initial reward via a rule-based function and concurrently extracts the corresponding answer probabilities from a reference model. Finally, these rewards are selectively refined using either the IGSR or SCRE strategy.

2. Related Works

2.1. Video Understanding with VLMs

The technical approach to video understanding has rapidly evolved from early CNN [10, 13] and Transformer-based models [1, 3] to the current paradigm dominated by Large Language Models (LLMs) [16, 36]. To overcome the information bottleneck inherent in traditional cascaded video-to-text pipelines, the field has converged on end-to-end Vision-Language Models (VLMs) such as Qwen2.5-VL [2] and GPT-4o [12]. These models directly project rich visual features into the LLM’s semantic space, enabling unified cross-modal reasoning and setting the state-of-the-art across major benchmarks. Our work, Video-RAISE, builds upon this modern framework, employing reinforcement learning to further enhance the reasoning capabilities.

2.2. Reinforcement Learning for Post-Training

Reinforcement Learning (RL) has become a cornerstone for enhancing the reasoning abilities of Vision-Language Models (VLMs) in post-training [6, 33]. Prevailing methods adapt the GRPO [9] by engineering task-specific rewards. For instance, Video-R1 [11] introduces a reward for temporal logic, while GRPO-CARE [7] adds a consistency-aware bonus to align reasoning with answers. However, these approaches are fundamentally tethered to a common paradigm: dependency on external, handcrafted rules. This reliance on manually specified rewards not only makes the system brittle and difficult to scale but also presupposes that human-defined heuristics are the optimal path to improving model reasoning. In sharp contrast, Video-RAISE departs from this paradigm of manual rule specification. Our work introduces a method that leverages the model’s own intrinsic

confidence during inference as a more authentic and scalable alignment signal. This design circumvents need for task-specific reward engineering and instead empowers model to optimize based on its own emergent cognitive

3. Method

3.1. Overview and Motivation

Overview. Figure 3 illustrates the training pipeline of our Video-RAISE. First, for each input video-question pair, we use the policy model to sample a set of token sequences. Subsequently, similar to GRPO, we employ a rule-based matching method to conduct a preliminary evaluation of the generated text and assign initial rewards. For responses that fail this match, the final reward is directly set to zero. For all other responses, we utilize a reference model to calculate the generation probability of their corresponding token sequences and then extract the probability value of the answer portion. This probability is then processed by either the IGSR or SCRE, depending on the question type, to generate a continuous, fine-grained reward. Finally, we use these rewards to calculate advantages, which are then used to update the policy model. The pseudocode for SCRE and IGSR is provided in Algorithm 1 and 2, respectively (see Appendix A). The appendix also includes theoretical proofs, as well as background details on reinforcement learning.

Motivation. Our core hypothesis is that a model’s intrinsic confidence (i.e., answer probability) serves as a proxy for its reasoning quality. To validate this, our analysis (Figure 4) reveals that the intuitive average sequence confidence (right plot) is an ambiguous and insufficient signal due to significant overlap between correct (blue) and incorrect (orange) answers. This ambiguity compels us to examine the

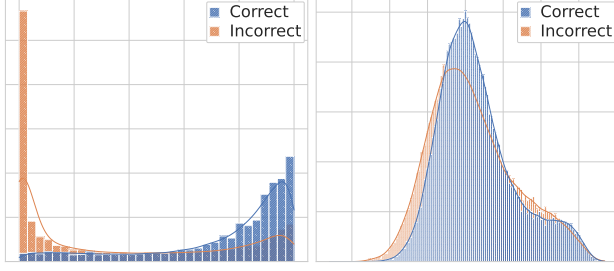


Figure 4. **Model confidence distributions for correct (blue) vs. incorrect (orange) answers.** (Left) Confidence distribution at the answer tokens, showing clear polarization for incorrect answers. (Right) Confidence distribution at the sequence token, showing significant overlap. Further details are provided in the Appendix.

more granular answer-token level (left plot). Here, we find a clearer, more complex pattern: correct answers are sharply concentrated at 1.0 confidence, while incorrect answers exhibit a distinct polarization: (1) A tall spike near 0.0 (guessing due to uncertainty). (2) A significant peak near 1.0 (onfidently wrong due to flawed, yet internally consistent, reasoning). This discovery shows the true signal lies at the token level, motivating our core design: applying specialized strategies tailored to different problem types based on these observed phenomena. For strict-logic tasks, which are highly sensitive to guessing (the 0.0-peak) and require perfect consistency, we apply SCORE. It is designed as a rigorous filter, targeting this extreme lack of confidence by penalizing both the lowest probability and the confidence variance. For general-purpose scenarios, the primary challenge is to refine a set of high-quality, high-confidence answers. For this, we apply IGSR, which addresses the challenge of ranking semantically similar, high-quality candidates by rewarding those with the highest, most stable confidence (targeting the 1.0-peak of the correct distribution).

3.2. Sequential Confidence Rigorous Evaluation

(Figure 5, Top) For tasks demanding strict correctness, where a single incorrect token can invalidate entire answer, a reward signal is required that identifies the confidence bottleneck of the answer sequence. We term this strategy Sequential Confidence Rigorous Evaluation (SCORE). Given a video-question pair, we first use policy model to sample G candidate sequences. To obtain a stable signal, we utilize reference model π_{ref} ¹ to compute token-wise probabilities:

$$\bar{p}_{i,j} = \pi_{\text{ref}}(o_{i,j} \mid q, o_{i,<j}), \quad (1)$$

where q represents the input video-question query and $o_{i,j}$ is the j -th token of the i -th token sequence. $o_{i,<j}$ refers to

¹ Both the policy model and the reference model achieve strong performance. The reference model is the base Qwen2.5-VL without SFT, updated via exponential moving average (EMA) during training. Further details are provided in the experiment section and Appendix.

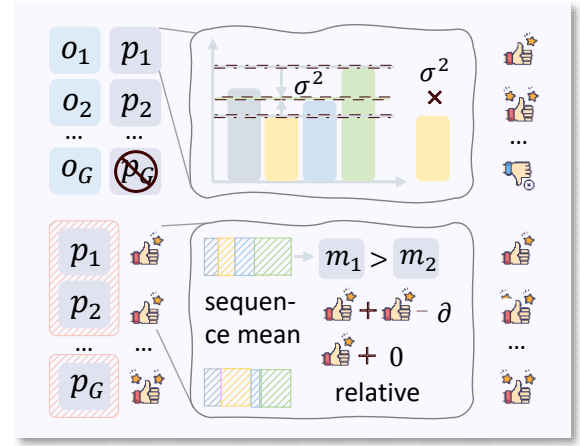


Figure 5. **An illustration of the SCORE (Top) and IGSR (Bottom).** SCORE targets exact-match tasks, considering both the minimum probability and variance. IGSR targets general questions, prioritizing accuracy while rewarding greater confidence.

the prefix of the i -th sequence before the $o_{i,j}$. Furthermore, $\bar{p}_{i,j}$ is the probability of generating $o_{i,j}$, $\bar{p}_{i,j} \in [0, 1]$. Then, we extract the probability of the answer portion by locating its corresponding token span within the generated sequence:

$$p_i = \text{Match}(\bar{p}_i, \langle \text{answer} \rangle \dots \langle / \text{answer} \rangle), \quad (2)$$

where, $\text{Match}(\cdot, \cdot)$ denotes the operation of matching and extracting the corresponding probabilities. To transform each answer sequence into a reward value, a simple averaging of these probabilities (mean) would be insufficient, as it could mask the impact of a single, low-confidence (and likely incorrect) token. Thus, we propose a reward function based on the weighted product of the answer probabilities. This formulation is highly sensitive to any single low probability, effectively penalizing sequences with a weak token:

$$\bar{r}_i = \prod_{j=1}^{|p_i|} p_{i,j}^{w_j}, \quad w_j = e^{-\eta j}, \quad (3)$$

where $|\cdot|$ denotes the length of the token sequence. w_j represents a position-based decay factor for the j -th token. The weighting scheme, controlled by hyperparameter $\eta > 0$, serves a distinct purpose: it assigns greater importance to tokens at earlier positions. This promotes the generation of well-formed outputs from the start, as the confidence of initial tokens has a larger impact on the final reward. Finally, to encourage a stable and consistent level of confidence throughout the answer, we give a variance penalty. This discourages outputs with erratic confidence scores:

$$\tilde{r}_i = \bar{r}_i \cdot e^{-\lambda \sigma^2}, \quad \sigma^2 = \frac{1}{|p_i|} \sum_{j=1}^{|p_i|} (p_{i,j} - \text{mean}(p_i))^2, \quad (4)$$

where the hyperparameter λ serves as a scaling factor.

Table 1. Comparison of our method with previous state-of-the-art (SOTA) methods on benchmarks for video reasoning and general understanding. Frame refers to the number of sampled frames per video. Note that for this evaluation, to ensure a fair comparison, we use only the multiple-choice subset of MMVU and do not use subtitles for VideoMME. The implementation details are given in Appendix.

Method	Pub	Frame	Video Reasoning Benchmark			Video General Benchmark		
			VSI-Bench	VideoMMMU	MMVU	MVBench	TempCompass	VideoMME
• Proprietary Models								
GPT-4o	-	-	34.0	61.2	75.4	-	-	71.9
Gemini 1.5 Pro	-	-	-	53.4	71.2	-	-	75.0
• SFT & Agent Models								
LLaMA-VID [18]	ECCV 24	-	-	-	-	41.9	45.6	-
VideoLLaMA2 [8]	arXiv 24	-	-	-	44.8	54.6	-	47.9
LongVA-7B [37]	arXiv 24	-	29.2	23.9	-	-	56.9	52.6
LLaVA-OneVision [17]	TMLR 24	-	32.4	33.8	49.2	56.7	-	58.2
Kangaroo-8B [21]	Arxiv 24	-	-	-	-	61.1	62.5	56.0
ShareGPT4Video [4]	NeurIPS 24	-	-	-	-	51.2	-	39.9
VILA-1.5-8B [19]	CVPR 24	-	28.9	20.8	-	-	58.8	-
VILA-1.5-40B [19]	CVPR 24	-	31.2	34.0	-	-	-	60.1
Video-UTR-7B [35]	ICLR 25	-	-	-	-	58.8	59.7	52.6
VideoTree (GPT 4) [30]	CVPR 25	64	-	47.8	54.2	-	-	56.1
mPLUG-Owl3-8B [34]	ICLR 25	-	-	-	-	54.5	-	53.5
• R1 Models								
RRPO-7B [27]	NeruIPS 25	16	-	-	-	59.0	44.3	-
ReAgent-V [38]	NeruIPS 25	16	33.1	47.9	63.0	61.4	70.3	54.2
Video-R1-7B [11]	NeruIPS 25	16	34.6	49.8	64.2	62.7	72.6	57.4
Video-RAISE-7B	-	16	34.8	52.8	66.4	64.3	74.7	57.4
Video-R1-7B [11]	NeruIPS 25	32	35.8	52.3	63.8	63.9	73.2	59.3
CARE-7B [7]	Arxiv 25	32	35.8	50.4	65.8	65.1	73.5	59.6
Video-RAISE-7B	-	32	36.6	53.0	65.9	65.9	75.1	60.7
Video-R1-7B [11]	NeruIPS 25	64	37.1	52.4	63.8	64.8	73.2	61.4
Video-RTS-7B [29]	EMNLP 25	64	-	52.7	66.3	-	-	-
Video-RAISE-7B	-	64	38.7	53.6	66.9	65.9	75.0	62.6

3.3. Intragroup Score Re-ranking

(Figure 5, Bottom) For tasks that permit diversity in answer, such as open-ended question answering, a strict token-by-token matching like SCORE is unsuitable. Such strategies penalizes correct responses that are semantically consistent with the reference answer but simply use different wording. To address this issue, we propose Intra-group Score Re-ranking (IGSR). The core design of IGSR is based on two principles. First, it provides a reward bonus to higher-confidence responses within a group of semantically equivalent candidates. Second, it introduces a novel cross-group reward constraint that sets a reward upper bound by utilizing the relationships between groups at different accuracy levels. This design aims to balance semantic accuracy, generation confidence, and content diversity in model’s responses.

To implement this, we first assess the semantic match between each candidate response and the reference answer

by calculating an accuracy score (e.g., using ROUGE-L). Based on these scores, candidates are clustered into discrete groups using a fixed interval size τ . Candidates in the same group are thus considered semantically similar:

$$g = \text{Group}(r(t), \tau), t = \text{Decode}(o), \quad (5)$$

where $\text{Group}(\cdot, \cdot)$ and $\text{Decode}(\cdot)$ respectively denote the operations of grouping into fixed intervals and decoding tokens into text. $r(t)$ represents the text-based accuracy reward and τ is the grouping interval. Since the differences in average confidence among semantically similar candidates within the same group are perhaps small and difficult to distinguish, we amplify these differences to enable a robust, fine-grained re-ranking of these candidates by calculating average negative log-probability for each candidate:

$$e_i = \frac{1}{|p_i|} \sum_{j=1}^{|p_i|} -\log(p_{i,j} + \delta), \quad (6)$$

Table 2. **Analysis of Reasoning Paths.** The prompt templates and examples used for this analysis are shown in the Appendix. Match indicates the consistency (exact match) between the LLM-generated answer and the original VLMs answer, while Answer measures the correctness of the LLM-generated answer. Our Video-RAISE demonstrates a significant lead on both metrics.

Method	Video Reasoning Benchmark						Video General Benchmark					
	VSI-Bench		VideoMMMU		MMVU		MVBench		TempCompass		VideoMME	
	Answer	Match	Answer	Match	Answer	Match	Answer	Match	Answer	Match	Answer	Match
Qwen2.5-VL	10.0	17.2	46.9	46.7	54.2	55.5	24.0	35.9	63.4	87.8	36.8	25.0
Qwen2.5-VL-SFT	16.3	41.2	47.4	87.8	61.9	52.3	52.2	86.6	65.4	93.5	49.6	90.5
Qwen2.5-VL-GRPO	18.5	41.4	40.4	34.0	45.7	47.2	29.1	26.9	47.3	43.7	32.9	37.3
Ours: Video-RAISE	30.3	84.9	55.3	87.9	66.7	94.9	63.5	95.5	73.5	95.9	56.1	93.4

where δ represents a small value used to avoid the undefined $\log(0)$ operation. Within each group g , we aim to reward candidates with higher confidence (i.e., lower average negative log-probability e_i). To achieve this, we first establish a representative reward for the group, $r_{(m)}$, defined as the median of all candidate rewards within that group. We then compute an adjustment bonus for each candidate $i \in g^{\{k\}}$ based on its relative confidence compared to its peers:

$$a_i = \left((r_{(m)}^{\{k+1\}} - r_{(m)}^{\{k\}}) - \partial \left(1 - 1/(|g^{\{k\}}| + 1) \right) \right) \cdot \left((\max(e^{\{k\}}) - e_i^{\{k\}}) / (\max(e^{\{k\}}) - \min(e^{\{k\}})) \right) \quad (7)$$

where $r_{(m)}^{\{k+1\}} - r_{(m)}^{\{k\}}$ is the reward gap between group k and the subsequent group $k + 1$ (for the last group, we set the next stage to 1.0). $e^{\{k\}}$ represents the negative log-probability within group. To enhance stability, we subtract a penalty, $\partial (1 - 1/(|g^{\{k\}}| + 1))$, which dynamically reduces the bonus range for larger, potentially noisier groups. We use the hyperparameter τ to serve a dual role. It not only defines the grouping interval but also acts as the minimum threshold for the adjustment bonus. The rationale is that τ represents the minimum resolution of semantic accuracy in our scheme; therefore, any reward adjustment a_i with a magnitude less than τ is considered negligible noise and is filtered out. The final re-ranked reward is computed by:

$$\tilde{r}_i = \sqrt{r_{(m)}^{\{k\}} \cdot (r_{(m)}^{\{k\}} + a_i \cdot \mathbf{1}(a_i \geq \tau))}, \quad (8)$$

where $\mathbf{1}(\cdot)$ is an indicator function that equals 1 if the condition is met, and 0 otherwise. This geometric mean ensures a smooth adjustment proportional to the group’s baseline reward. Ultimately, IGSR thus effectively re-ranks candidates within each semantic group while, at the same time, maintaining stable inter-group reward separation.

4. Experiments

The Appendix provides detailed settings, further particulars, and a description of additional experiments.

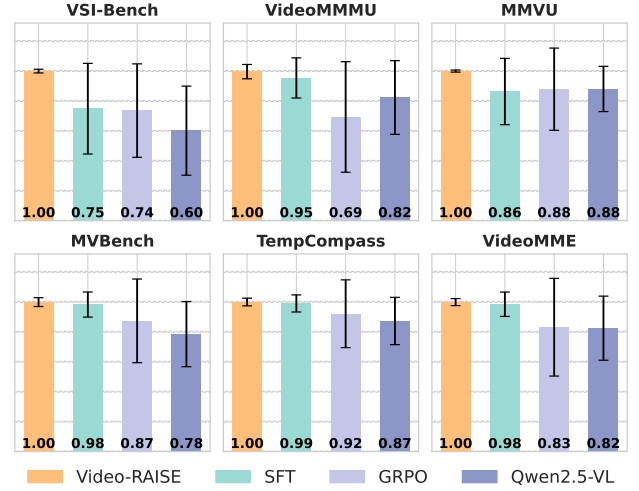


Figure 6. **Comparison of Probability.** Bars indicate the mean, while error bars represent the standard deviation. On high-consistency datasets, SFT performs similarly to Video-RAISE.

4.1. Main Results

Table 1 presents a detailed comparison of our Video-RAISE against current SOTA methods on all benchmarks.

Comparison with SFT/Agent Methods. As shown in Table 1, our approach significantly outperforms previous SFT and agent-based SOTA models. On VideoMMMU (16-frame), Video-RAISE (52.8) surpasses the agent-based VideoTree [30] (47.8) and other SFT methods like Kangaroo-8B [21] by a **19.0-point** margin. This robust lead extends to general benchmarks, such as outperforming Video-UTR-7B [35] on TempCompass by **15.0** points.

Comparison with RL-based Methods. More importantly, Video-RAISE shows comprehensive leadership over other RL-based SOTA methods. In the 16-frame setting, it outperforms Video-R1 [11] by **+3.0** points on VideoMMMU and **+2.1** points on TempCompass. In the 32-frame setting, it surpasses CARE [7] by **+2.6** points on VideoMMMU (53.0 vs. 50.4) and **+1.6** points on TempCompass.

Table 3. **Ablation Study on Training Configurations.** We compare reward sources (Policy vs. Reference), the effect of SFT (the model fine-tuned on Video-R1-260k [11]), and a KL divergence penalty (coefficient 0.04). Results show strong performance from both policy and reference models as reward sources, each with distinct advantages. For more detailed analysis, please refer to the Appendix.

Method	Frame	Video Reasoning Benchmark			Video General Benchmark		
		VSI-Bench	VideoMMMU	MMVU	MVBench	TempCompass	VideoMME
Qwen2.5-VL-SFT-GRPO	16	34.4	50.1	63.2	64.0	73.2	56.9
Qwen2.5-VL-GRPO	16	33.6	47.7	65.3	63.8	73.9	57.0
Video-RAISE-Policy-Instruct	16	35.6	52.4	67.2	64.4	74.8	58.4
Video-RAISE-Policy-SFT	16	36.5	52.1	64.5	63.5	72.6	56.5
Video-RAISE-Reference-Instruct	16	34.8	52.8	66.4	64.3	74.7	57.4
Video-RAISE-Reference-SFT	16	36.0	49.9	61.4	63.5	72.1	57.3
Video-RAISE-Policy-KL	16	33.5	52.3	66.4	64.1	73.7	55.9
Video-RAISE-Reference-KL	16	31.9	52.2	65.3	63.1	73.8	55.3
Video-RAISE-Policy-Instruct	32	37.0	52.3	65.9	66.2	75.3	60.4
Video-RAISE-Reference-Instruct	32	36.6	53.0	65.9	65.9	75.1	60.7
Video-RAISE-Policy-Instruct	64	39.2	53.0	67.0	66.3	75.2	61.0
Video-RAISE-Reference-Instruct	64	38.7	53.6	66.9	65.9	75.0	62.6

This consistent leadership across varying frame rates establishes Video-RAISE as the new SOTA performance.

4.2. Comparison of Reasoning Path Consistency

We extract path (the content within `<think>...</think>` tags) and feed it with the question to a text-only LLM. We evaluate two metrics: **Answer** (correctness of the new answer) and **Match** (consistency, i.e., if the new answer matches the original). The details are given in Appendix.

Metric. Answer: As shown in Table 2, Video-RAISE achieves the highest correctness scores, demonstrating superior logical quality by substantially outperforming SFT and GRPO on VSI-Bench. Notably, the GRPO method shows clear reasoning degradation, with its scores significantly lower than SFT and even the baseline on benchmarks like TempCompass. **Match:** Video-RAISE also exhibits a significant and more robust advantage. As noted in prior studies [23], the SFT method performs well on in-distribution datasets like VideoMMMU, but these improvements do not generalize well to out-of-distribution benchmarks. On VSI-Bench, its consistency drops sharply to 41.2%. The GRPO method performs poorly on consistency across the board, achieving only 34.0% on VideoMMMU and 43.7% on TempCompass. Video-RAISE achieves high consistency (nearly 90%) across all benchmarks.

Probability. We use the minimum probability observed during answer generation as a proxy for the model’s confidence. As shown in Figure 6, we find a positive correlation between this answer confidence and the consistency of the model’s reasoning path. On the TempCompass and VideoMME benchmarks, the SFT model exhibits reasoning consistency comparable to our Video-RAISE; correspond-

ingly, its minimum answer probability also reaches a high level of 0.98.

4.3. Training Configuration

To identify the optimal training configuration, we present a detailed comparison of different settings in Table 3. Our analysis addresses two key questions: (1) whether a preliminary Supervised Fine-Tuning (SFT) stage is beneficial, and (2) whether it is better to calculate rewards using the current policy model or the more stable reference model. Based on this analysis (detailed in the Appendix), we adopt the instruction-tuned version (without SFT) as the base model and utilize the reference model for reward calculation.

4.4. Case Study

Figure 7 presents a case study of a counting task. Qwen2.5-VL-Instruct fails because it only attends to the most salient table and omits the others, leading to an incorrect result. In contrast, Video-RAISE demonstrates a superior reasoning process: it first identifies all relevant tables, then lists their specific locations, and performs a verification step before providing the final, correct answer. This reasoning is both relevant and factually correct. This robust reasoning process stems from our core mechanism, which encourages model to seek high-confidence conclusions over hasty judgments.

4.5. Ablation Study

The ablation study in Table 4 validates our components, SCRE and IGSR, and their synergy. We first observe that the GRPO baseline is more robust than the SFT variant, which shows mixed results and can even degrade performance (e.g., on VideoMMMU), confirming the viability of RL

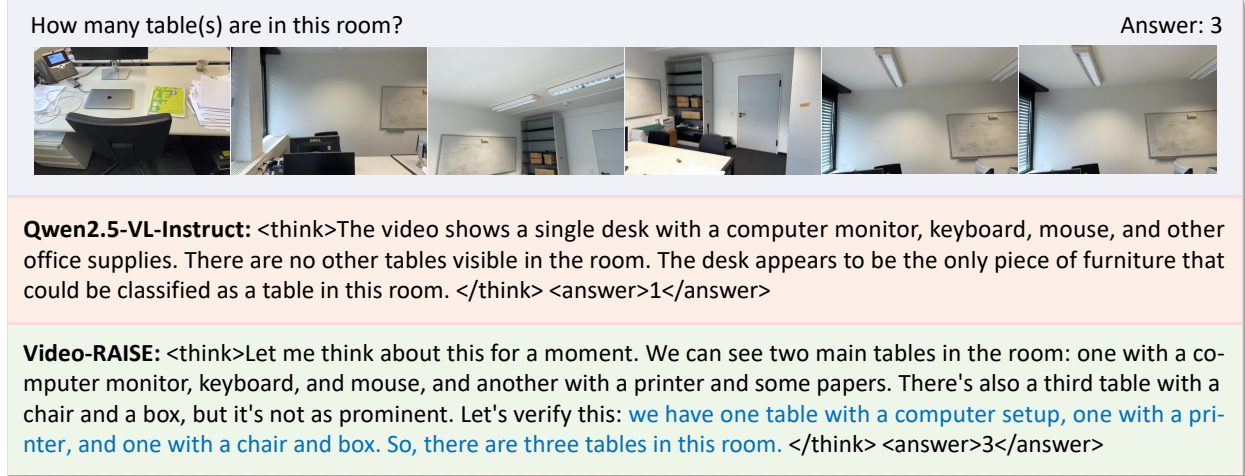


Figure 7. **Case Study of Counting Task (VSI-Bench).** While the Qwen2.5-VL-Instruct provides an incorrect answer, **in contrast**, our Video-RAISE gives the correct answer, supported by a more logical and coherent reasoning path. See the Appendix for further examples.

Table 4. **Ablation study for the components of our proposed method.** Qwen2.5-VL is the official baseline model. Qwen2.5-VL-SFT is obtained by applying supervised fine-tuning on the Video-R1-CoT-165k dataset. Qwen2.5-VL-GRPO applies a text-rule-based reward for optimization on Video-R1-260k dataset. Video-RAISE-SCRE and Video-RAISE-IGSR represent the two key components we propose.

Method	Frame	Video Reasoning Benchmark			Video General Benchmark		
		VSI-Bench	VideoMMMU	MMVU	MVBench	TempCompass	VideoMME
Qwen2.5-VL [11]	16	27.7	47.8	59.2	57.4	72.2	53.1
Qwen2.5-VL-SFT [11]	16	31.8	47.4	61.3	59.4	69.2	52.8
Qwen2.5-VL-GRPO	16	33.6	47.7	65.3	63.8	73.9	57.0
Video-RAISE-SCRE	16	35.7	48.8	65.9	63.6	74.3	57.6
Video-RAISE-IGSR	16	34.9	52.1	65.1	64.3	74.2	57.0
Ours: Video-RAISE	16	34.8	52.8	66.4	64.3	74.7	57.4

post-training. Isolating our components reveals their complementary strengths: **SCRE**, for strict-correctness tasks, yields consistent gains, notably boosting VSI-Bench by **+2.1** points over GRPO. **IGSR**, for open-ended questions, proves highly effective on complex reasoning, outperforming GRPO on VideoMMMU (52.1%) and securing the top score on MVBench (64.3%). The **Video-RAISE** model, integrating both, outperforms either component in isolation, showing powerful synergy on challenging benchmarks. This provides strong empirical validation for our core hypothesis: specialized, type-specific reward strategies are crucial for the optimal performance.

5. Conclusion

In this paper, we proposed Video-RAISE, a novel self-quantification paradigm for VLM post-training. Instead of relying on external human-defined constraints, Video-RAISE empowers the model to optimize its reasoning paths using its own intrinsic confidence. Our core motivation

stems from the finding that sequence-level confidence is an ambiguous signal, while token-level analysis reveals a complex polarization of guessing versus confidently wrong states. We address this with two specialized, complementary strategies: Sequential Confidence Rigorous Evaluation, which targets strict-logic tasks by penalizing low-probability guesses and confidence variance; and Intra-group Score Re-ranking, which refines general-purpose answers by re-ranking semantically similar, high-quality candidates. Video-RAISE achieves state-of-the-art performance by significant margins across six diverse video understanding benchmarks. Furthermore, it demonstrates superior reasoning consistency, significantly outperforming prior RL methods and even surpassing SFT.

Acknowledgements

This work was supported by Ant Group Research Intern Program.

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE, 2021. 3
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Ming-Hsuan 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. *CoRR*, abs/2502.13923, 2025. 3
- [3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 813–824. PMLR, 2021. 3
- [4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 5
- [5] Siran Chen, Boyu Chen, Chenyun Yu, Yuxiao Luo, Ouyang Yi, Lei Cheng, Chengxiang Zhuo, Zang Li, and Yali Wang. Vragent-r1: Boosting video recommendation with mllm-based agents via reinforcement learning, 2025. 1
- [6] Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Bohan Zeng, Yang Shi, Sihang Yang, Pengfei Wan, Qiang Liu, Liang Wang, and Tieniu Tan. Versavid-r1: A versatile video understanding and reasoning model from question answering to captioning tasks. *CoRR*, abs/2506.09079, 2025. 3
- [7] Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning, 2025. 1, 2, 3, 5, 6
- [8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476, 2024. 5
- [9] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. 1, 3
- [10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1933–1941. IEEE Computer Society, 2016. 3
- [11] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *CoRR*, abs/2503.21776, 2025. 1, 2, 3, 5, 6, 7, 8
- [12] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisposi, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pélisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogó Gierltler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. 3
- [13] Zhuolin Jiang, Viktor Rozgic, and Sancar Adali. Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 309–317. IEEE Computer Society, 2017. 3
- [14] Siwen Jiao, Yangyi Fang, Baoyun Peng, Wangqun Chen, and Bharadwaj Veeravalli. Lavidrive: Vision-text interaction VLM for autonomous driving with token selection, recovery and enhancement. *CoRR*, abs/2411.12980, 2024. 1
- [15] Jipeng Jin, Zhaoxiang Zhang, Zhiheng Li, Xiaofeng Gao, Xiongwen Yang, Lei Xiao, and Jie Jiang. Pareto-based multi-objective recommender system with forgetting curve. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 4603–4611. ACM, 2024. 1
- [16] Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, and Sumit Shekhar. Seeing the unseen: Visual metaphor captioning for videos. *CoRR*, abs/2406.04886, 2024. 3
- [17] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025. 5
- [18] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, pages 323–340. Springer, 2024. 5

- [19] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE, 2024. 5
- [20] Bo Liu, Pengfei Qiao, Minhan Ma, Xuange Zhang, Yinan Tang, Peng Xu, Kun Liu, and Tongtong Yuan. Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models. *CoRR*, abs/2505.12589, 2025. 1
- [21] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoli Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *CoRR*, abs/2408.15542, 2024. 5, 6
- [22] Saeid Aghasoleymani Najafabadi. Enhancing live broadcast engagement: A multi-modal approach to short video recommendations using MMGCN and user preferences. *CoRR*, abs/2506.23085, 2025. 1
- [23] Graham Neubig, Seungone Kim, Minxin Du, Xiang Yue, Tuney Zheng, Radha Poovendran, Yuetai Li, Xiaoyu Xu, and Maggie Huan. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. 2025. 7
- [24] Yi Peng, Chris, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork R1V: pioneering multimodal reasoning with chain-of-thought. *CoRR*, abs/2504.05599, 2025. 1
- [25] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispidder: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. *CoRR*, abs/2501.03218, 2025. 1
- [26] Tanqiu Qiao, Ruochen Li, Frederick W. B. Li, Yoshiki Kubotani, Shigeo Morishima, and Hubert P. H. Shum. Geometric visual fusion graph neural networks for multi-person human-object interaction recognition in videos. *Expert Syst. Appl.*, 290:128344, 2025. 1
- [27] Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *CoRR*, abs/2504.12083, 2025. 5
- [28] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Weixin Xu, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms, 2025. 1
- [29] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Re-thinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. *CoRR*, abs/2507.06485, 2025. 5
- [30] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for LLM reasoning on long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 3272–3283. Computer Vision Foundation / IEEE, 2025. 5, 6
- [31] Peng Wu, Wanshun Su, Guansong Pang, Yujia Sun, Qingsen Yan, Peng Wang, and Yanning Zhang. Avadclip: Audio-visual collaboration for robust video anomaly detection. *CoRR*, abs/2504.04495, 2025. 1
- [32] Zhenghao Xi, Xiang Liu, Yaqi Liu, Yitong Cai, and Yangyu Zheng. Integrating generative adversarial networks and convolutional neural networks for enhanced traffic accidents detection and analysis. *CoRR*, abs/2506.16186, 2025. 1
- [33] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *CoRR*, abs/2506.03340, 2025. 3
- [34] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 5
- [35] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, and Wenbing Tao. Unhackable temporal rewarding for scalable video mllms. *CoRR*, abs/2502.12081, 2025. 5, 6
- [36] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 21715–21737. Association for Computational Linguistics, 2024. 3
- [37] 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. *CoRR*, abs/2406.16852, 2024. 5
- [38] Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding. *CoRR*, abs/2506.01300, 2025. 5