BEYOND ENTROPY: SELF-DIAGNOSTIC MULTI-ROLE TOKEN OPTIMIZATION FOR VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards has substantially advanced multimodal reasoning, yet it remains fundamentally limited by ambiguous token-level credit assignment. While high-entropy token heuristics encourage possibility exploration, naively extending them to video reasoning tends to induce lengthy reasoning, as the model becomes overly reliant on high-entropy visual activations. Alternative approaches that rely on counterfactual-based visual token localization for credit assignment also tend to over-prioritize visual exploration at the expense of decisive reasoning cues for answer derivation, thereby exacerbating the interference from spurious visual nuances. Moreover, these methods employ static counterfactual strategies that fail to co-evolve with the policy during training. In this paper, we introduce DyCPO, a co-evolutionary framework that jointly optimizes reliable token selection and adaptive counterfactual intervention. It constructs a multi-role dependence metric to balance visual exploration and answer-relevance mining in token-wise contrastive learning, while suppressing exploration-only filler tokens and spurious visual noise. Rather than relying on static counterfactual priors, DyCPO dynamically derives counterfactual signals from the model's own successful and failed rollouts, enabling self-diagnostic analysis and co-evolution of the optimization objective with the policy. Extensive experiments on complex video reasoning and general video understanding benchmarks demonstrate consistent performance improvements, establishing DyCPO as a robust token-level credit assignment paradigm for multimodal reinforcement learning.

1 INTRODUCTION

Reinforcement learning (RL) with verifiable rewards (Yu et al., 2025; DeepSeek-AI, 2025), exemplified by Group Relative Policy Optimization (DeepSeek-AI, 2025) (GRPO), has emerged as a pivotal paradigm for enhancing reasoning in large language models. However, GRPO uniformly distributes trajectory-level rewards across all tokens, leaving the critical challenge of token-level credit assignment unresolved. While recent work (Cui et al., 2025; Wang et al., 2025b) identifies high-entropy tokens (approximately 20%) as key decision points for exploration in text-only reasoning, the generalizability of this heuristic to video reasoning remains uncertain. Although recent multimodal RL studies (Ye et al., 2026; Rezaei et al., 2026; Yang et al., 2026; Jin et al., 2026) emphasize visual perception for finer-grained credit assignment, blindly amplifying visual signals risks introducing noise rather than informative signals. Meanwhile, on-policy distillation (OPD) methods (Wang et al., 2026a; Yuan et al., 2026; Liu et al., 2026) offer fine-grained supervision but are hindered by their reliance on privileged annotations and limited scalability.

To investigate these issues, **(1)** we first demonstrate that naively rewarding token entropy deteriorates video reasoning performance. As shown in Figure 2(d), the instruct model without RL training exhibits a rise-then-fall entropy trend during decoding, with reliance on low-entropy visual activations decaying toward the end of generation. This dynamic reflects two distinct functional stages: in the early perception stage, reliance on high-entropy visual cues peaks as the model resolves fine-grained visual details; in the later logic stage, once most visual facts are incorporated into the CoT prefix, the model shifts to low-entropy activations corresponding to stable visual scaffolds that provide object representation and spatio-temporal consistency. Far from being redundant, these activations serve as structural anchors that enforce global consistency and mitigate over-reasoning and hallucination. This

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.

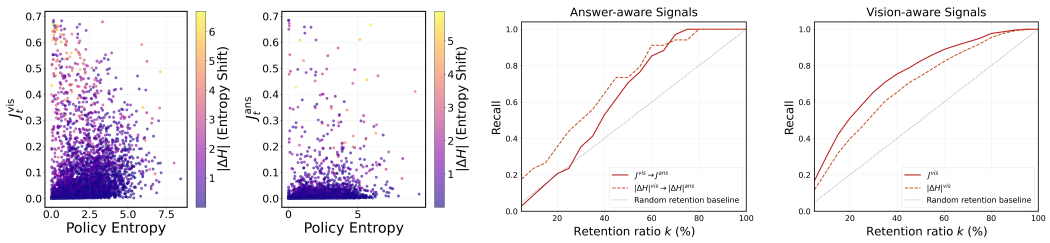


Figure 1: **Left to right:** (a, b) Numerous tokens with high vision-dependent and answer-dependent ΔH_t / J_t reside in the low-entropy token set. (c, d) Token omission induced by single-selection metrics. Specifically, the recall in (c) is defined as $|A_{\text{ent}} \cap A_{\text{vis}}|/|A_{\text{vis}}|$, quantifying the coverage of visual-dependency tokens within the high-entropy selection. Similarly, the recall in (d) is calculated as $|A_{\text{vis}} \cap A_{\text{ans}}|/|A_{\text{ans}}|$, measuring the alignment between vision-sensitive and answer-critical tokens.

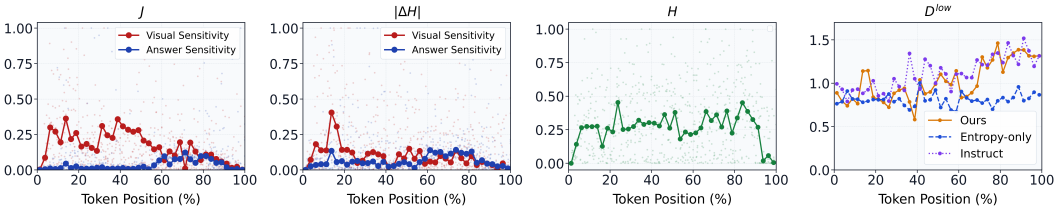


Figure 2: **Left to right:** (a, b) Vision- and answer-aware sensitivity across token positions, (c) entropy dynamics, and (d) reliance on towards low-entropy visual activations. Token position denotes the normalized decoding length.

finding extends the observation in work (Xing et al., 2025), providing further empirical substantiation for this mechanism. In contrast, the model optimized with high-entropy tokens consistently relies on uncertain visual cues throughout the decoding process. Without converged visual grounding, entropy-based reweighting misdirects gradient signals toward exploratory tokens at the expense of decisive ones, resulting in noisy updates and the neglect of decision-critical content. These observations prompt us to move beyond naive entropy heuristics.

(2) As Figure 1(a)(b) shows, relying on just one rule to select tokens inevitably creates blind spots. On one hand, if we only pick the top-20% highest-entropy tokens, we miss nearly half of the tokens that are actually grounded in the video (J_{vis} and $|\Delta H|_{vis}$ in the right plot). On the other hand, focusing only on vision-dependent tokens fails to capture how the reasoning actually unfolds. As the left plot ($J_{vis} \rightarrow J_{ans}$) shows, picking the top-20% most vision-sensitive tokens only catches 20% of the signals needed for the answer, which performs no better than random guessing. This two-way mismatch proves that "seeing" (perceptual grounding) and "thinking" (logical deduction) follow different paths. Therefore, this prompts us to evaluate tokens separately using *vision-dependent* and *answer-dependent* metrics.

(3) To further probe the divergence in token selection, we apply visual and answer counterfactual interventions to measure token sensitivity to input perturbations. As shown in Figure 1(a)(b), tokens with high ΔH_t or J_t under visual counterfactuals concentrate in early decoding steps (0–40), with magnitudes monotonically decaying across the sequence. In contrast, answer-sensitive tokens localize to middle and late reasoning stages (60–80). The token cloud distributions in Figure 5 further corroborate their semantic complementarity: the former anchor on visual entities or attributes, while the latter mark logical pivots and conclusions, filtering out vision-irrelevant noise. These results collectively demonstrate that any single-role criterion is inherently insufficient for token-level credit assignment in video reasoning.

We therefore advocate for *multi-role tokens* as the most informative and least noise-prone candidates for credit assignment: they are sensitive to both visual and answer counterfactuals while retaining high entropy. In this paper, we design a multi-role dependency score to quantify token importance and steer policy-gradient updates toward more robust optimization directions. A further critical observation is that random or static counterfactual strategies can induce model collapse when applied to dynamic video reasoning scenarios. We conjecture that fixed perturbation strategies are suboptimal because they fail to adapt to the evolving model state during training. To address this, we propose a

state-conditioned mechanism that dynamically adjusts counterfactual generation. Specifically, we extract informative signals from rollout trajectories to guide this adaptive process by thoroughly mining fine-grained cues. These rollouts are generated by a proposal network that is updated in lockstep with the policy model at each training step.

Extensive experiments on knowledge reasoning, general understanding, spatial perception, long-video benchmarks, and video hallucination demonstrate that DyCPO achieves state-of-the-art or highly competitive performance across diverse tasks. Notably, on Video-Holmes, DyCPO attains an accuracy of 44.8%, outperforming GPT-4o (42.0%) by a significant margin. Comprehensive analyses and ablations confirm that jointly optimizing multi-role credit assignment and state-conditioned counterfactuals provides an effective, interpretable, and robust RL foundation for multimodal video reasoning.

2 METHODS

2.1 PRELIMINARIES

RL for video reasoning. Given a multimodal query $x = (\mathcal{V}, q)$ with video $\mathcal{V}$ and question q, the policy π_θ generates a reasoning trajectory $\mathbf{o} = (o_1, \ldots, o_M)$. Group Relative Policy Optimization (GRPO) samples G candidate responses $\{\mathbf{o}_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ and computes group-normalized relative advantages $A_i = (R(\mathbf{o}_i) - \mu_R)/(\sigma_R + \delta)$, where $R(\mathbf{o}_i)$ is a scalar verifier reward, μ_R and σ_R are the reward mean and standard deviation over the group, and $\delta > 0$ prevents numerical division by zero. The policy is updated by maximizing:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{o}_i|}\sum_{m=1}^{|\mathbf{o}_i|}\min\left(r_{i,m}A_i, \text{clip}(r_{i,m}, 1-\epsilon, 1+\epsilon)A_i\right) - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})\right], \tag{1}$$

where $r_{i,m} = \pi_\theta(o_{i,m} \mid x, \mathbf{o}_{i,<m})/\pi_{\theta_{\text{old}}}(o_{i,m} \mid x, \mathbf{o}_{i,<m})$ is the importance sampling ratio, ϵ is the clipping threshold, and $\beta \geq 0$ penalizes divergence from the frozen reference policy π_{ref}. While GRPO steers exploration via trajectory-level advantages A_i, its uniform weighting across decoding steps obscures fine-grained credit assignment, motivating our token-level analysis.

To diagnose token behavior, we introduce the following definitions.

Token entropy. We define the Shannon entropy of a distribution P as $H(P) = -\sum_w P(w)\log P(w)$, and the Jensen–Shannon Divergence between distributions P and Q as:

$$J(P, Q) = \tfrac{1}{2}D_{\text{KL}}(P \parallel M) + \tfrac{1}{2}D_{\text{KL}}(Q \parallel M), \quad \text{where } M = \tfrac{P+Q}{2}. \tag{2}$$

Reliance on low-entropy visual activations. To quantify token grounding in stable visual cues, we define a reliance score on low-entropy visual tokens. We project layer-l visual states $\mathbf{h}_v^{(l)} \in \mathbb{R}^d$ into the vocabulary space via the unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{|\mathcal{V}|\times d}$, obtaining $P_v^{(l)} = \text{softmax}(\mathbf{W}_U\mathbf{h}_v^{(l)})$. The deterministic visual anchor set $\mathcal{V}_{\text{low}}^l$ and the visual reliance score D_t^{low} of token o_t are defined as:

$$D_m^{\text{low}} = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\sum_{v\in\mathcal{V}_{\text{low}}^l} a_{m,v}^{(l)}, \quad \mathcal{V}_{\text{low}}^l = \left\{v \in \mathcal{V} \mid H(P_v^{(l)}) \leq \tau_{\text{low}}^{(l)}\right\} \tag{3}$$

where $\tau_{\text{low}}^{(l)}$ is the bottom q_{low}-th quantile threshold of visual entropies at layer l, and $a_{m,v}^{(l)} \in [0, 1]$ is the normalized cross-attention weight from token o_m to visual position v at layer l.

Vision- and answer-aware sensitivity. We probe token dependencies on visual and answer cues via counterfactual perturbations. For visual sensitivity, replacing video $\mathcal{V}$ with a perturbed counterfactual stream $\mathcal{V}'$ yields distribution $Q_m = \pi_\theta(\cdot \mid \mathcal{V}', q, \mathbf{o}_{<m})$, with shifts:

$$J_m^{\text{vis}} = J(P_m, Q_m), \quad \Delta H_m^{\text{vis}} = H(Q_m) - H(P_m). \tag{4}$$

For answer sensitivity, fixing $\mathcal{V}$ and contrasting verified answer a^+ against counterfactual alternative $a^- \in \mathcal{A} \setminus \{a^+\}$ yields conditional distributions $P_m^{\pm} = \pi_\theta(\cdot \mid \mathcal{V}, q, a^{\pm}, \mathbf{o}_{<m})$, inducing shifts:

$$J_m^{\text{ans}} = J(P_m^+, P_m^-), \quad \Delta H_m^{\text{ans}} = H(P_m^-) - H(P_m^+). \tag{5}$$

These metrics form the foundation for diagnosing token behavior, and are leveraged to optimize multi-role credit assignment in Section 2.2.

2.2 MULTI-ROLE CONTRASTIVE CREDIT ASSIGNMENT

Multi-role token selection. As motivated above, we select *multi-role tokens* as the optimal units for policy optimization, as they jointly capture sensitivity to visual grounding and answer derivation while filtering out noise-prone tokens with low informativeness. Building on the sensitivity metrics defined earlier, we construct a composite score to identify these tokens. Given a rollout $\mathbf{o}_i = \{o_{i,1}, \ldots, o_{i,M}\}$ generated by policy π_θ, we first normalize the two sensitivity signals per rollout, then fuse them into a single joint score via a soft probabilistic OR:

$$c_{i,m} = \prod_{c \in \{\text{vis},\text{ans}\}} \left[\left(1 - \text{Norm}\left(J^c_{i,m}\right)\right)\left(1 - \text{Norm}\left(|\Delta H^c_{i,m}|\right)\right)\right]^{\alpha_c} \cdot H_{i,m}, \tag{6}$$

where $c \in \{\text{vis}, \text{ans}\}$ indexes the visual and answer modalities, $J^c_{i,m}$ is the Jensen–Shannon divergence between the original and perturbed token distributions (i.e., $J^{\text{vis}}_{i,m}$ for visual perturbation and $J^{\text{ans}}_{i,m}$ for answer perturbation), $\Delta H^c_{i,m}$ denotes the entropy shift induced by the corresponding counterfactual intervention, $\text{Norm}(\cdot)$ denotes min-max normalization within the rollout, $\alpha_c \in [0, 1]$ with $\sum_c \alpha_c = 1$ balances the contribution of visual exploration and reliable grounding, and $H_{i,m}$ denotes the entropy of the token distribution at step m, which reweights the fused sensitivity score by the token's informativeness. Finally, the top-k response tokens by $c_{i,m}$ form the multi-role set:

$$\mathcal{C} = \left\{ m \;\middle|\; \frac{\text{Rank}(c_{i,m})}{|\mathbf{o}_i|} \leq k \right\}, \tag{7}$$

where $\text{Rank}(\cdot)$ denotes the descending rank and $k \in (0, 1]$ the retention ratio. The resulting binary mask $M_{i,m} = \mathbb{I}\left[m \in \mathcal{C}\right]$ dynamically identifies the most informative and reliable multi-role tokens.

Overall objective and policy update. Building on the multi-role token set $\mathcal{C}$ defined in Eqn. 7, we introduce a contrastive regularization term to reinforce informative and reliable token during policy update. Unlike prior work that directly reweights tokens (Wang et al., 2026b), our approach separates selected tokens from unselected ones by aligning their rollout prediction distributions with an information-preserving counterfactual view π^+_θ (with slight jitter $\mathcal{V}^+$) and distancing them from an information-removing view π^-_θ (with crucial visual evidence $\mathcal{V}^-$ removed). Aligning with π^+_θ stabilizes training and makes the model robust to small input changes, while distancing from π^-_θ prevents the model from relying on irrelevant or misleading features. For each selected token, the contrastive loss is defined as:

$$\mathcal{L}^{\text{MRC}}_{i,m} = M_{i,m} \cdot \text{softplus}\left(\frac{D_{\text{KL}}\left(\pi_\theta(o_{i,m}) \,\|\, \pi^+_\theta(o_{i,m})\right) - D_{\text{KL}}\left(\pi_\theta(o_{i,m}) \,\|\, \pi^-_\theta(o_{i,m})\right)}{\tau} \right), \tag{8}$$

where τ is a temperature parameter and $\text{softplus}(x) = \log(1 + e^x)$. The token-level loss $\mathcal{L}^{\text{MRC}}_{i,m}$ smoothly vanishes once the model achieves the desired separation between the positive and negative counterfactual views on the selected multi-role tokens. To avoid reinforcing spurious patterns from failed rollouts, we gate the loss exclusively on successful samples with positive advantages ($A_i > 0$) and directly incorporate it into the standard GRPO loss as the final training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}}(\theta) + \frac{\lambda_{\text{MRC}}}{|\mathcal{C}|} \sum_{m \in \mathcal{C}} \mathbb{I}(A_i > 0) \cdot \mathcal{L}^{\text{MRC}}_{i,m}, \tag{9}$$

where λ_{MRC} balances the weight between outcome-driven policy optimization and token-wise self-distillation.

2.3 SELF-DIAGNOSTIC EVOLVING OPTIMIZATION

Model optimization is inherently dynamic, and counterfactual interventions play a non-trivial role in identifying salient tokens. This motivates examining whether static counterfactual strategies can align with such dynamic training processes. As shown in Figure 3, the fixed random masking baseline exhibits relatively low and volatile token-to-answer correlation, while our state-aware strategy achieves consistently higher correlation, particularly in later training stages, suggesting its capacity to co-evolve with the policy and prioritize answer-relevant visual clues. This indicates the value of adaptive counterfactual strategies for effective long-video reasoning.

Rollout-contrastive modeling. Specifcially, we mine informative signals from rollout trajectories, observing that visual attention divergence between correct and incorrect rollouts highlights reasoning-relevant cues, while answer segments contain the visual evidence the model converges on. Motivated by this, we model dynamics from two perspectives: (1) *Dynamic visual counterfactuals*: We introduce a rollout-contrastive keyframe localization method, leveraging frame-level attention discrepancies between correct and incorrect rollouts as reasoning signals. Given G rollouts from a proposal model π_ϕ, we categorize them into $\mathcal{R}_r = \{\mathbf{o}_i \mid A(\mathbf{o}_i) \geq 0\}$ and $\mathcal{R}_w = \{\mathbf{o}_i \mid A(\mathbf{o}_i) < 0\}$ based on advantage values.

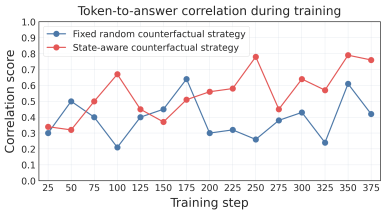


Figure 3: Training dynamics of the correlation between activated tokens and answer-relevant reasoning cues.

For each rollout trajectory $\mathbf{o}_i = \{o_{i,1}, \ldots, o_{i,M}\}$, we track the attention weights from response tokens to visual frames. Motivated by prior findings (Zhang et al., 2025) demonstrating that attention converges to semantically relevant regions in intermediate layers but diverges in deeper ones, we compute the aggregated attention from the m-th token to the t-th frame $\mathcal{V}_t$ by averaging over a selected subset of model layers:

$$\mathcal{A}_{m,t}^{(i)} = \frac{1}{H \cdot |\mathcal{V}_t|} \sum_{h=1}^{H} \sum_{v \in \mathcal{V}_t} a_{m,v}^{(h)}, \tag{10}$$

where H is the number of attention heads, $|\mathcal{V}_t|$ denotes the number of visual tokens in frame $\mathcal{V}_t$, and $a_{m,v}^{(h)}$ is the raw attention weight from token m to visual patch v at head h. We first select the top-K_1 response tokens with the strongest visual grounding to frame $\mathcal{V}_t$, and average their attention scores to obtain the representative response-to-frame attention score:

$$\bar{\mathcal{A}}_t^{(i)} = \frac{1}{K_1} \sum_{m \in \Phi_t^{(i)}} \mathcal{A}_{m,t}^{(i)}, \quad \text{where } \Phi_t^{(i)} = \mathrm{TopK}_{m \in \{1,\ldots,M\}} \left(\mathcal{A}_{m,t}^{(i)},\ K_1 \right). \tag{11}$$

Subsequently, to pinpoint the video frames predominantly attended to by the entire response $\mathbf{o}_i$, we apply a second Top-K_2 selection across the video temporal dimension:

$$\Psi^{(i)} = \mathrm{TopK}_{t \in \{1,\ldots,T\}} \left(\bar{\mathcal{A}}_t^{(i)},\ K_2 \right), \tag{12}$$

where $\Psi^{(i)}$ represents the indices of the K_2 most visually salient frames for rollout $\mathbf{o}_i$.

To construct a contrastive set of keyframes, we extract the union of identified focal frames within each respective group and define the critical region set $\mathcal{U}^*$ via set difference:

$$\mathcal{U}^* = \Psi^+ \setminus \Psi^-, \quad \text{where } \Psi^+ = \bigcup_{\mathbf{o}_i \in \mathcal{R}_r} \Psi^{(i)}, \ \ \Psi^- = \bigcup_{\mathbf{o}_i \in \mathcal{R}_w} \Psi^{(i)}. \tag{13}$$

This formulation interprets the gap between the visual focus of correct and incorrect rollouts as dense, region-level supervision derived from sparse outcome rewards. However, fixing a static keyframe set throughout training may compromise stability and generalization, as the policy's visual attention often shifts as its reasoning capabilities evolve. We therefore linearly anneal the keyframe count K_2 over the course of training, a curriculum that allows the model to first master fine-grained visual grounding before tracking broader temporal spans, smoothly extending contrastive self-distillation from static perception to temporal reasoning. (2) *Answer contrastive counterfactuals*: We extract the correct and incorrect answer texts from the current rollouts as privileged information. Unlike directly sampling random answers from incorrect options, our strategy aligns with the model's own learning state, enabling it to better identify its current confusion or weaknesses, and thus further update gradients along this direction.

EMA Self-Teacher. We construct the proposal model π_ϕ via an Exponential Moving Average (EMA) of the parameters of policy model π_θ, eliminating the need for a separately pretrained model. After each optimization step, the proposal parameters are updated according to:

$$\pi_\phi \leftarrow (1 - \alpha)\pi_\phi + \alpha \pi_\theta, \tag{14}$$

Table 1: **Performance Comparison. Bold** indicates the best results. † marks results from original papers under different evaluation protocols. "Data Size" refers to the RL training data scale. The default video input consists of 32 frames.

		Video Reasoning					Video General Benchmark	
Model	**Data Size**	**VSI.**	**VideoMMMU**	**MMVU**	**Video-Holmes**	**MV.**	**TempC.**	**VideoMME**
Proprietary Models								
GPT-4o Hurst et al. (2024)	–	34.0	61.2	75.4	-	-	-	71.9
Open-Source Models								
VideoLLaMA2 Cheng et al. (2024)	–	-	-	44.8	-	54.6	-	47.9
LongVA-7B Zhang et al. (2024)	–	29.2	23.9	-	-	-	56.9	52.6
VILA-1.5-8B Lin et al. (2024)	–	28.9	20.8	-	-	-	58.8	-
LLaVA-OneVision-7B Li et al. (2024)	–	32.4	33.8	49.2	-	56.7	-	58.2
mPLUG-Owl3-8B Ye et al. (2024)	–	-	-	-	-	54.5	-	53.5
Qwen2.5-VL-7B Bai et al. (2025)	–	31.8	47.4	61.3	-	59.4	69.2	52.8
Concurrent R1-based Models								
Video-R1 NeurIPS' 25 Feng et al. (2025)	260K	35.8	52.3	63.8	36.5	63.9	72.6	59.3
VideoChat-R1 NeurIPS' 25 Feng et al. (2025)	18K	-	-	64.2	-	63.1	72.9	-
TW-GRPO Li et al. (2025b)	1K	-	-	65.8	-	63.3	73.3	-
Video-RTS ICLR' 26 Wang et al. (2025d)	6K	-	52.7	66.4	40.7	-	-	55.1
Video-RFT NeurIPS' 25 Wang et al. (2025a)	102K	36.8	52.3	68.5	-	62.1	73.2	59.8
APPO CVPR' 26 Du et al. (2026)	260K	32.7	-	-	40.8	64.6	71.9	59.8
Video-KTR ICLR' 26 Wang et al. (2026b)	16.5K	36.8	52.6	65.9	41.6	-	73.4	60.3
Ours (7B)	3K	**39.2**	**53.6**	**68.5**	**44.8**	**65.8**	**74.2**	**62.6**

where $\alpha \in (0, 1)$ is the EMA momentum coefficient. The proposal model π_ϕ is updated exclusively via exponential moving average (EMA) without backpropagation, serving as a slowly evolving rollout-contrastive anchor for self-distillation of the policy model. This setup prevents overly aggressive target drifts, enabling the policy model π_θ to fully adapt within a relatively stable counterfactual phase, thereby preserving rich temporal context. The complete training procedure is summarized in Algorithm 1 in the Appendix.

3 Experiments

3.1 Experimental Setup

Training datasets and benchmarks. For Supervised Fine-Tuning (SFT), we leverage VideoRFT-CoT-102K to establish fundamental reasoning capabilities. For RL training, we curate two 3K subsets from the raw Video-R1 dataset (Feng et al., 2025) via difficulty-based filtering to train the Qwen-2.5-VL and Qwen-3-VL models, respectively. This strategy retains only samples with verifiable answers while excluding trivially easy or excessively hard instances, ensuring that the RL data difficulty aligns with model capabilities. We evaluate our approach on seven challenging benchmarks: (1) **Video-Holmes** (Cheng et al., 2025), **VideoMMMU** (Hu et al., 2025), and **MMVU** (Zhao et al., 2025) for temporal, causal, and knowledge-intensive reasoning; (2) **TempCompass** (Liu et al., 2024) and **VideoMME** (Fu et al., 2025) for general capability; (3) **LVBench** (Wang et al., 2025c) for long-video comprehension; (4) **VSIBench** (Yang et al., 2025) for spatial perception; and (5) video hallucination diagnosis **VidHalluc** (Li et al., 2025a) .

Implementation details. We use Qwen2.5-VL-7B-Instruct as the backbone for public benchmark evaluation and train all models on 16× NVIDIA H20 GPUs (96GB each). For fair comparison, we follow the Video-R1 training protocol, and all ablations adopt identical configurations. During training, each video is uniformly sampled into 16 frames under a visual token budget of $128 \times 28 \times 28$ by default. We optimize the model with a learning rate of $1e{-}6$ and a global batch size of 64. At inference, we increase the temporal input to 32 frames with resolution $256 \times 28 \times 28$. The RL pipeline is built on `verl`. More details are provided in the Appendix.

Baselines. We evaluate our method against a comprehensive set of state-of-the-art baselines with similar scale, including LLaVA-OneVision (Li et al., 2024), VILA-1.5 (Lin et al., 2024), Qwen2.5-VL (Bai et al., 2025), Video-R1 (Feng et al., 2025), Video-RTS (Wang et al., 2025d), VideoRFT (Wang et al., 2025a), TW-GRPO (Dang et al., 2025), Video-KTR (Wang et al., 2026b), APPO (Du et al., 2026) and on-policy distillation method Video-OPSD (Wang et al., 2026a).

Table 2: Ablation study of different scoring metrics across MMVU-MC, VSI-MC, VideoMMMU, Video-Holmes, and VideoMME (short/medium/long/overall) over three backbone architectures.

Model	Method	MMVU	VSI	VideoMMMU				Video-Holmes	VideoMME			
				Perc.	Compre.	Adapt.	Avg.		Short	Medium	Long	Avg.
Qwen2.5-VL 7B	GRPO	65.3	35.9	61.8	47.3	41.5	50.2	40.9	70.0	58.4	49.5	59.3
	wo ΔH_{ans}/ J_{ans}	64.3	36.5	61.0	47.6	41.8	50.1	41.2	71.3	59.7	50.4	60.5
	wo ΔH_{vis}/ J_{vis}	67.2	37.9	62.3	47.0	37.0	48.8	42.1	70.7	57.8	49.5	59.4
	wo H_t	69.0	39.1	64.0	51.6	42.7	52.8	44.2	73.3	60.5	50.5	61.5
	Our Full Model	68.5	39.2	63.8	52.3	43.2	53.1	44.8	74.3	61.5	52.0	62.6
Qwen3-VL 2B	GRPO	60.1	47.4	60.0	42.3	29.0	43.8	35.5	69.7	52.8	46.2	56.3
	wo ΔH_{ans}/ J_{ans}	60.2	48.1	58.6	46.5	35.1	46.7	36.1	68.4	51.9	47.2	55.8
	wo ΔH_{vis}/ J_{vis}	59.5	46.8	57.3	47.2	34.2	46.2	37.4	68.2	51.3	46.5	55.3
	wo H_t	61.1	48.1	58.3	48.8	35.4	47.5	37.1	70.2	53.6	47.9	57.2
	Our Full Model	61.6	49.4	60.0	49.3	36.3	48.6	37.7	70.5	54.3	48.4	57.7
Qwen3-VL 4B	GRPO	66.9	46.2	73.0	61.0	41.0	58.3	41.5	72.9	60.2	49.0	60.2
	wo ΔH_{ans}/ J_{ans}	66.2	47.8	73.3	61.2	42.8	59.1	39.1	72.4	60.9	51.2	61.5
	wo ΔH_{vis}/ J_{vis}	65.1	47.1	71.0	62.8	42.3	58.7	40.4	73.2	59.3	49.5	60.7
	wo H_t	65.6	47.8	74.3	62.3	44.5	60.4	41.2	74.5	60.2	49.9	61.5
	Our Full Model	67.0	48.3	74.7	63.0	45.3	61.0	42.2	74.4	61.9	53.0	63.1

3.2 Main Results

Performance comparison with state-of-the-art models. As shown in Table 1, our method consistently sets new state-of-the-art results across three reasoning and two general video understanding benchmarks. On **Video-Holmes**, which diagnoses high-level temporal and social reasoning in short films, our model hits **44.8%**, beating the strongest baseline (Video-KTR) by 3.2% and sitting just behind proprietary models like Gemini-2.5-Pro (45.0%). For knowledge-intensive tasks, we scores **68.5%** on **MMVU (MC)** using noticeably less RL training data, showing our effective design towards complex multimodal reasoning. Crucially, these reasoning gains don't come at the cost of general comprehension: we hit **62.6%** on general video benchmarks ahead of the previous best at 60.3%.

Performance comparison with existing RL methods. As reported in Table 3, our method consistently outperforms all RL-based approaches, including the adapted PAPO and Video-OPSD, across both backbone settings. The largest gains appear on knowledge-intensive benchmarks: VideoMMMU (+6.5 over PAPO on Qwen3-VL-4B) and MMVU (+4.2 over PAPO on Qwen2.5-VL-7B). These results confirm that answer-relevant token selection and temporal dynamics modeling effectively incorporate useful knowledge while filtering noise.

Table 3: Performance comparison with other RL methods. We use the same backbone for fairness, and † denotes our reimplementation results using the same data configurations.

Method	VideoMME	VideoMMMU	MMVU	Video-Holmes
Qwen2.5-VL-7B				
+PAPO†	61.0	45.7	64.3	41.5
+Video-OPSD	58.6	49.9	-	36.2
+Ours	**62.6**	**53.6**	**68.5**	**44.8**
Qwen3-VL-4B				
+PAPO†	60.2	54.5	64.8	41.5
+Video-OPSD	60.7	52.9	-	41.8
+Ours	**63.1**	**61.0**	**67.0**	**42.7**

Effect of multi-role token selection. To validate the efficacy of our multi-role token selection strategy, we construct several ablation variants by systematically ablating individual scoring signals. Specifically, we exclude either the answer-aware components (ΔH_t^c and J_t^c) or the visual-aware counterparts, while also introducing an entropy-only baseline that allocates credit solely based on decoding uncertainty. As reported in Table 2, the integration of these complementary roles yields consistent performance gains across diverse datasets and evaluation capabilities. These findings indicate that no single metric is sufficient to fully capture the critical factors underlying effective reasoning. Notably, while the entropy-based credit allocation narrows the performance gap with our method on temporally sensitive benchmarks in VideoMME, it exhibits substantial degradation on reasoning-oriented benchmarks. This discrepancy may arise because enhancing visual exploration capability facilitates better cross-frame semantic association, but may also introduce visual noise in reasoning tasks, leading to visual ambiguity and performance deterioration.

3.3 Ablation Study

Component effects of self-diagnostic evolving modeling. As shown in Figure 4, **(1)** removing A_i degrades performance across all benchmarks (largest drops: 4.0%

on VideoMME, 3.7% on Video-Holmes), confirming that it stabilizes policy optimization by filtering low-quality rollouts and preventing suboptimal reasoning reinforcement. **(2)** Reverse curriculum variants (Narrow → Broad) underperform the Broad → Narrow setting by 3.5% on VideoMME, indicating that global-to-local learning misaligns with the model's evolving capacity. Our progressive curriculum, by contrast, balances visual exploration and reasoning reliability by mastering local frame-level reasoning before expanding to longer temporal horizons. **(3)** The re-weighting (softmax) baseline underperforms our contrastive self-distillation by 2.1% on VideoMME and 3.6% on VideoMMMU, showing that token weight amplification along training dynamics is more susceptible to visual noise. **(4)** $\mathcal{R}_r$-focus, $\mathcal{R}_w$-focus, and their set difference yield comparable results, suggesting that video information attended during incorrect reasoning is not entirely noise and highlighting the inherent redundancy of video.

Table 4: Ablation study on Qwen2.5-VL-7B.

Method	VideoMME	VideoMMMU	MMVU	Video-Holmes
Full Model	62.6	53.6	68.5	44.8
Effect of Advantage Gating				
w.o. A_i Gating	$58.6_{\downarrow 4.0}$	$51.0_{\downarrow 2.6}$	$64.8_{\downarrow 3.7}$	$41.1_{\downarrow 3.7}$
Effect of Curriculum				
Broad → Narrow	$59.1_{\downarrow 3.5}$	$49.7_{\downarrow 3.9}$	$65.4_{\downarrow 3.1}$	$42.3_{\downarrow 2.5}$
Fixed	$60.6_{\downarrow 2.0}$	$51.2_{\downarrow 2.4}$	$65.6_{\downarrow 2.9}$	$42.1_{\downarrow 2.7}$
Optimization Strategy				
Re-weighting (Softmax)	$60.5_{\downarrow 2.1}$	$50.0_{\downarrow 3.6}$	$66.5_{\downarrow 2.0}$	$42.1_{\downarrow 2.7}$
Key Frame Selection Strategy				
$\mathcal{R}_r$-$focus$	$61.8_{\downarrow 0.8}$	$52.8_{\downarrow 0.8}$	$67.6_{\downarrow 0.9}$	$44.6_{\downarrow 0.2}$
$\mathcal{R}_w$-$focus$	$61.0_{\downarrow 1.6}$	$52.5_{\downarrow 1.1}$	$67.2_{\downarrow 1.3}$	$41.5_{\downarrow 3.3}$

3.4 Analysis

Effect on video hallucination. As shown in Table 5, DyCPO consistently improves all hallucination metrics on VidHalluc, validating its effectiveness in suppressing spurious visual noise. It outperforms the w/o $J^{ans}/\Delta H_{ans}$ variant by +1.1 on ACH MCQ (90.6 vs. 89.5), demonstrating that multi-role dependence steers contrastive learning toward answer-relevance mining. Compared to the "Only Entropy" ablation, DyCPO achieves a +5.4 gain on STH (93.5 vs. 88.1), reflecting its ability to anchor spatio-temporal grounding and prevent hallucinations of non-existent objects or temporal relationships.

Table 5: Performance comparison on video hallucination benchmark *VidHalluc*.

Method	ACH		TSH	STH
	Binary QA	MCQ		
w/o $J^{ans}/\Delta H_{ans}$	**90.4**	89.5	**94.8**	91.9
Only Entropy	89.8	89.3	93.6	88.1
Ours	**91.4**	90.6	**94.8**	**93.5**

Effect of curriculum strategy. As shown in Figure 5(a), where TCI stands for *Temporal Causal Inference* and TA denotes *Timeline Analysis*, we observe that the introduction of curriculum learning progressively enhances the model's temporal reasoning capabilities (TA from 28.5% → 31.6%), while perceptual abilities remain stable. This suggests that our strategy of gradually expanding the visual receptive field from local keyframes enables the model to simultaneously learn local cues and global dependencies.

Visual- and answer-sensitive token distributions. To intuitively illustrate the distinct preferences of visual-aware and answer-aware tokens, we visualize their respective distributions via word clouds in Figure 5(b). Tokens with high vision-aware sensitivity concentrate on semantic and color attributes, such as "blue", "heated", and "person", whereas answer-aware tokens are dominated not only by nouns and attribute words but also by conclusive function words like "aligns", "suggests", and "followed", which predominantly reveal conclusions or final answers.

Visualization for rollout-contrastive visual anchors. To intuitively interpret the behavior of the proposal model, we visualize the keyframes selected via our rollout-contrastive mechanism. As illustrated in Figure 5(c), for the question "What does the disappearance of wall reflection when pressing the switch at 0:13 imply?", the correct rollout predominantly attends to frames 0–4, while the incorrect rollout notably neglects frames 14, and 15. This indicates that the critical visual cues for resolving the query are the switch-related elements in frames 9, 14, and 15, whereas the early frames merely provide contextual background without decisive evidential value. Our self-diagnostic counterfactual strategy effectively guides the model to attend to these answer-critical visual cues, thereby enhancing reasoning performance.

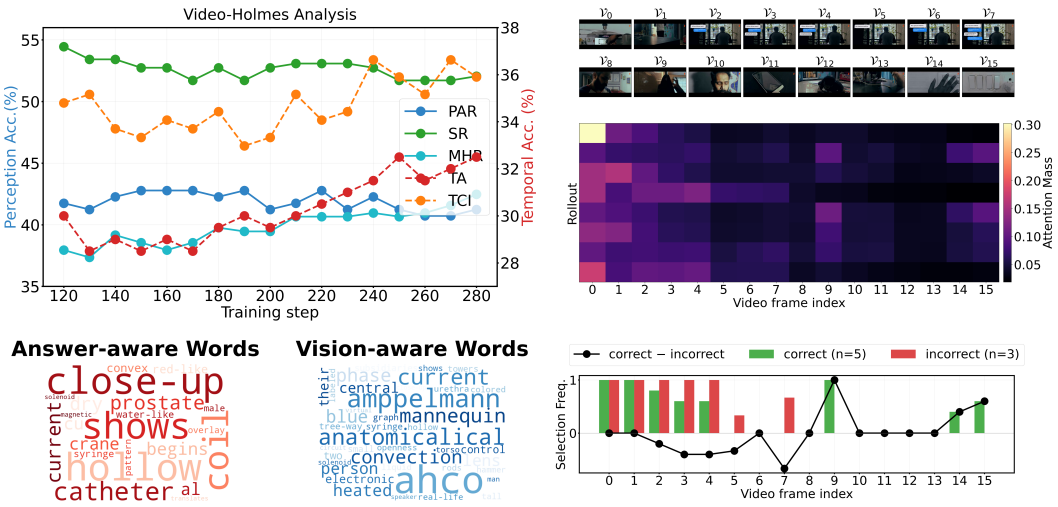


Figure 5: **Left to right:** (a) Steady rise in temporal metrics (TA, TCI) with stable perception (MHR). (b) Word distribution comparison between answer-aware and vision-aware tokens. (c) Detailed qualitative rollout and fine-grained visual diagnosis on Video-Holmes. The attention maps visualize the frame-level focus of each rollout, where lighter colors indicate higher attention weights. The bar chart at the bottom compares the frequency with which correct and erroneous rollouts attend to specific frames, while the black curve depicts the divergence between them.

Relationship between reward dynamics and token activation ratio. We investigate the coupling between reward magnitude and token activation ratio—the fraction of tokens contributing to $\mathcal{L}_{\text{MRC}}$ gradients. Since $\mathcal{L}_{\text{MRC}}$ activates only on rollouts with positive advantages ($A_i > 0$, Eq. 9), the mean reward directly reflects the number of informative rollouts per group. As shown in Figure 4, the activation ratio exhibits an M-shaped (bimodal) pattern as reward increases. This non-monotonicity stems from two failure modes at opposite difficulty extremes: (1) overly simple questions yield deterministic "all-correct" rollouts, where zero intra-group variance forces all advantages toward zero under GRPO's advantage normalization, eliminating gradients and collapsing token-wise credit assignment. (2) Conversely, prohibitively hard prompts produce near-zero success rates, leaving $\mathcal{L}_{\text{MRC}}$ without valid activation signals from positive-advantage rollouts and similarly suppressing token engagement. Therefore, maintaining rewards within an intermediate band of balanced difficulty is essential to prevent advantage vanishing and signal sparsity, thereby maximizing token activation and the optimization efficacy of $\mathcal{L}_{\text{MRC}}$.

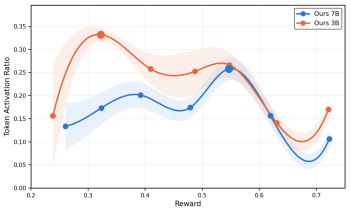


Figure 4: Correlation between reward and token activation fraction overall training steps. Our method based on 7B and 3B share the similar trends.

4 Conclusion

In this paper, we empirically investigate the distinct roles that individual tokens play during video reasoning and propose a multi-role token selection mechanism coupled with a self-diagnostic rollout-contrastive intervention paradigm. Specifically, we introduce a proposal-policy collaborative learning framework operating at the token level. We first quantify the inherent limitations of high-entropy tokens in video understanding and reveal the functional complementarity between visual-aware and answer-aware activations through a systematic token distribution analysis. We further demonstrate that our self-diagnostic curriculum masking strategy can better and more reliably stimulate the model's ability to capture fine-grained visual concepts and temporal dynamics, yielding robust gains under challenging conditions. Extensive experiments across multiple backbones, including ablations on scoring metrics, masking strategies, and curriculum directionality, as well as thorough comparisons with on-policy distillation methods, consistently demonstrate the effectiveness and superiority of our approach, establishing a new state of the art on key video reasoning benchmarks.

AI USE STATEMENT

Large language models, such as GPT-5.6 or Claude, are used exclusively for grammar checking during the writing process. They are not used for research ideation. In addition, LLM-based coding assistants, such as Copilot, are employed to support code implementation partially.

ETHICS STATEMENT

This work enhances the training robustness and scalability of video reasoning models through rollout-contrastive self-distillation. The reported gains on video understanding and reasoning benchmarks do not establish safety or reliability beyond the evaluated settings. Applications of the method should therefore include task-specific safety evaluation and appropriate review of generated outputs.

REPRODUCIBILITY STATEMENT

To support reproducibility, Sec. 2 specifies the keyframe-driven contrastive perception objective, rollout-contrastive keyframe localization, advantage gating, and progressive curriculum strategy, while Algorithm 1 in Appendix presents the training procedure. Experimental settings, model configurations, and data sources are described in Appendix A, which also documents hardware, training hyperparameters (including K_1/K_2, λ, α_{ans} and α_{vis}, η, and layer selection L as detailed in Table 6), prompt templates, baseline configurations (GRPO with symmetric clipping $\epsilon_l = \epsilon_h = 0.2$ and no reference KL penalty), decoding settings, evaluation metrics, and answer verification procedures. For the theoretical result, Additional training-dynamics analyses and the procedures used to construct the corresponding smoothed curves are provided in Appendix C.

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. *arXiv preprint arXiv:2502.13923*, 2025. 6

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can MLLM think like holmes for complex video reasoning? *CoRR*, abs/2505.21374, 2025. 6

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. 6

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=inccdtfx8x. 14

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, abs/2505.22617, 2025. 1

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. 2025. URL https://arxiv.org/abs/2505.24718. 6

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948. 1

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. APPO: attention-guided perception policy optimization for video reasoning. *CoRR*, abs/2602.23823, 2026. 6

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. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. 6, 14

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 *CVPR*, 2025. 6

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 14

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *CoRR*, abs/2501.13826, 2025. 6

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. 6

Senjie Jin, Peixin Wang, Boyang Liu, Xiaoran Fan, Shuo Li, Zhiheng Xi, Jiazheng Zhang, Yuhao Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. Entropy is not enough: Unlocking effective reinforcement learning for visual reasoning via vision-anchored token selection. *CoRR*, abs/2606.03937, 2026. 1

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2024. 6

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13723–13733. Computer Vision Foundation / IEEE, 2025a. 6

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*, 2025b. 6

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 6

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *CoRR*, abs/2605.21924, 2026. 1

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *ACL*, 2024. 6

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. In *NeurIPS*, 2025. 14

Ahmad Rezaei, Mohsen Gholami, Saeed Ranjbar Alvar, Kevin Cannons, Mohammad Asiful Hossain, Zhou Weimin, Yong Zhang, and Mohammad Akbari. CPPO: Contrastive perception policy optimization for VLM agents. In *ICML*, 2026. 1

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 14

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. 2025a. 6

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. 1

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. Lvbench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 22958–22967. IEEE, 2025c. 6, 15

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. 2025d. URL https://arxiv.org/abs/2507.06485. 6, 14

Ziyue Wang, Shiqi Huang, Weiwen Xu, Bihan Wen, and Xudong Jiang. Video-opsd: Exploiting privileged visual evidence for on-policy self-distillation in video large language models. 2026a. URL https://arxiv.org/abs/2608.27065. 1, 6, 14

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-ktr: Reinforcing video reasoning via key token attribution. *CoRR*, abs/2601.19686, 2026b. 4, 6, 14

Yun Xing, Xiaobin Hu, Qingdong He, Jiangning Zhang, Shuicheng Yan, Shijian Lu, and Yu-Gang Jiang. Boosting reasoning in large multimodal models via activation replay. *CoRR*, abs/2511.19972, 2025. 2

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 *CVPR*, 2025. 6

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in MLLM reasoning. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 11694–11702. AAAI Press, 2026. 1

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 6

Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models. *CoRR*, abs/2604.01840, 2026. 1

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. 1

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026. 1

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026. 14

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Trans. Mach. Learn. Res.*, 2024. 6

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information flow in lvlms across reasoning tasks. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 2289–2299. Association for Computational Linguistics, 2025. 5, 17

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *CVPR*, 2025. 6

A Related Work

While reinforcement learning (RL) algorithms such as GRPO (Guo et al., 2025) or minimalist policy gradient baselines like GPG (Chu et al., 2026) have have revolutionized textual reasoning or image-based reasoning , extending them to multimodal large language models (MLLMs) often neglects temporal dynamics. To bridge this gap, recent video-centric RL methods (Feng et al., 2025; Wang et al., 2025d) incorporate temporally perturbed contrastive learning, introduce tool-based agent pipelines (Zeng et al., 2026) that decouple grounding from reasoning for long-form video understanding, or design scalable reward functions such as difficulty-aware regression (Park et al., 2025). Nevertheless, these approaches predominantly rely on coarse sequence-level rewards or simple heuristics, failing to model fine-grained and reliable reasoning processes. Consequently, recent efforts have shifted toward improving token-level credit assignment via attribution modeling (Wang et al., 2026b) or leveraging the teacher's privileged information to provide fine-grained supervision within on-policy distillation (OPD) frameworks (Wang et al., 2026a).

B More Implementation Details

Models. We employ Qwen2.5-VL-7B, Qwen3-VL-2B, and Qwen3-VL-4B as our base models for detailed ablative analysis. The Qwen2.5-VL models are initialized from a VideoRFT SFT checkpoint, while the Qwen3-VL models start from their official weights. By default, the base model for benchmark comparison refers to Qwen2.5-VL-7B to ensure fairness. All variants share the same GRPO configuration and differ only in the perception objective; we consider the following main model variants and default configurations:

- **GRPO**: GRPO (Shao et al., 2024) baseline with symmetric clipping factors $\epsilon_l = \epsilon_h = 0.2$, no reference KL penalty (β removed, i.e. `use_kl_loss=False` and `use_kl_in_reward=False`), GRPO advantage without per-group standard-deviation normalization, and zero entropy bonus.
- **Ours**: Given the on-policy rollout, the student's own attention over layers L selects K keyframes, on which a keyframe-driven contrastive perception objective $\mathcal{L}_{MRC}$ is computed. The perception score fuses four our defined metrics—vision and answer Jensen–Shannon divergence together with the absolute entropy shift $|\Delta H|$ (equal 0.25 weights)—with fusion weight α, entropy-gate strength η, temperature $\tau = 1$, and hard occlusion negatives. PAPO introduces *no* EMA self-distillation teacher; the keyframes only drive the CPL perception term, which is added to the policy loss with coefficient λ under advantage gating.

Model	Frames	K_2	λ	α	η	Layers L	β	$\varepsilon_l, \varepsilon_h$
Qwen2.5-VL GRPO-3B	16	-	0	-	-	-	0.0	0.2, 0.2
Qwen2.5-VL GRPO-7B	16	-	0	-	-	-	0.0	0.2, 0.2
Qwen2.5-VL GRPO-7B	24	-	0	-	-	-	0.0	0.2, 0.2
Qwen2.5-VL GRPO-7B	32	-	0	-	-	-	0.0	0.2, 0.2
Qwen2.5-VL Ours-3B	16	4→8	0.05	0.5	0.01	Removing last five layers	0.0	0.2, 0.2
Qwen2.5-VL Ours-7B	16	4→8	0.05	0.5	0.01	Removing last five layers	0.0	0.2, 0.2
Qwen2.5-VL Ours-7B	24	6→12	0.05	0.5	0.01	Removing last five layers	-	0.2, 0.2
Qwen2.5-VL Ours-7B	32	8→16	0.05	0.5	0.01	Removing last five layers	-	0.2, 0.2
Qwen3-VL GRPO-2B	16	-	0	-	-	-	0.0	0.2, 0.2
Qwen3-VL GRPO-4B	16	-	0	-	-	-	0.0	0.2, 0.2
Qwen3-VL Ours-2B	16	4→8	0.05	0.5	0.01	wo injection layer	0.0	0.2, 0.2
Qwen3-VL Ours-4B	16	4→8	0.05	0.5	0.01	wo injection layer	0.0	0.2, 0.2

Table 6: **Hyperparameter configurations** for different models. K_1 and K_2 respectively refers to the ratio of selected response token for keyframes localization and the number of attention-selected keyframes, annealed linearly over training by a curriculum; λ is the MRC perception coefficient (we set to 0 in GRPO setting), α_1 and α_2 the balacedd weight of scoring fusion weight, η the entropy-gate strength, and L the transformer layers whose response→frame attention is pooled for keyframe selection. All models remove the reference KL penalty (β) for better exploration.

Table 6 summarizes the hyperparameter configurations for the different model variants. For our method, we employ hard occlusion negatives alongside an answer-conditioned hard negative (defined as the rollout with the lowest accuracy within the group) and option-text answers serving as privileged information. The keyframe count K_1 follows a linear curriculum ($4 \rightarrow 8$ for 16-frame inputs and $8 \rightarrow 16$ for 32-frame inputs), which scales the temporal receptive field of the perception module according to the input frame budget. Regarding the attention layer set $\mathcal{L}$, we utilize a fixed subset for Qwen2.5-VL, whereas for Qwen3-VL—which employs DeepStack to inject visual features across all layers—we select all layers except those directly receiving visual injections.

Training. We conduct RLVR on all model variants using the following typical response format, where reasoning steps are enclosed in `<think></think>` and the final answer in `\boxed{}`. Each model is trained for 1 epoch on our 3K video-QA training set with a learning rate of $1e{-6}$, frames sampled uniformly from each video at 100352 pixels (128 visual tokens per frame), a maximum prompt length of 4096 and response length of 512. During RL policy optimization, we employ a rollout group size of $n = 8$ generated responses per prompt, maintaining a batch size of 8.

C Additional Experiments

Generalization to extended temporal horizons. Long-video understanding poses severe challenges for multimodal models due to overwhelming frame redundancy and high computational costs. To assess whether our method maintains effectiveness over extended temporal contexts, we evaluate on LVBench (Wang et al., 2025c). Table 7 presents the comparative results. Our method establishes a new state-of-the-art with 40.67% average accuracy, outperforming GRPO (40.03%) and entropy-only (39.70%) baselines. The advantage is particularly pronounced in dynamic content categories: we achieve substantial improvements in Sport (+0.83% over GRPO), Self-Media (+2.84%), and Live (+1.15%). Although GRPO slightly leads in Documentary and TV categories, our method exhibits more balanced performance across the spectrum.

Table 7: Performance comparison on different subtasks in LVBench.

Model	Avg.	Cartoon	Documentary	Live	Self-Media	Sport	TV
GRPO	40.03	34.28	**45.18**	39.69	40.06	42.50	**40.40**
Entropy-only	39.70	34.63	43.15	38.93	41.01	42.08	39.60
Ours	**40.67**	**35.34**	43.15	**40.84**	**42.90**	**43.33**	39.20

Table 8: Performance comparison of models trained with different input video frames across five benchmarks.

Frames	Method	MMVU	VSI	VideoMME	Video-Holmes	VideoMMMU	Avg.
$f = 24$	GRPO	66.7	36.6	61.9	43.8	50.0	51.80
	Ours	**68.8**$_{\uparrow 2.1}$	**39.7**$_{\uparrow 3.1}$	**63.9**$_{\uparrow 2.0}$	**45.2**$_{\uparrow 1.4}$	**51.3**$_{\uparrow 1.3}$	**53.78**$_{\uparrow 1.98}$
$f = 32$	GRPO	66.2	36.7	61.5	42.4	51.3	51.62
	Ours	**69.0**$_{\uparrow 2.8}$	**40.2**$_{\uparrow 3.5}$	**64.8**$_{\uparrow 3.3}$	**45.2**$_{\uparrow 2.8}$	**52.8**$_{\uparrow 1.5}$	**54.40**$_{\uparrow 2.78}$

Training with more frames. To investigate the impact of input frame capacity during RL training, we evaluate models trained under different video frames ($f = 24, 32$). As shown in Table 8, we have two key observations:

(1) Scaling gains from denser video frame inputs: Increasing the training frame count from $f = 24$ to $f = 32$ yields modest but consistent gains. Specifically, our method improves from 63.9% to 64.8% on VideoMME and from 51.3% to 52.8% on VideoMMMU, indicating that exposure to richer temporal contexts during RL training enhances the model's capacity to reason over complex visual semantics.

(2) Robustness of designed components: Under both frame configurations, our method consistently outperforms the standard GRPO baseline across diverse benchmarks. The performance advantage is

particularly pronounced on reasoning-oriented benchmarks like **Video-Holmes** ($42.4\% \rightarrow 45.2\%$) and spatial perception benchmark **VSI** ($36.7\% \rightarrow 40.2\%$) with *f = 32* video input. This indicates that our multi-role selection effectively activates valuable visual clues inherent in denser frames, ensuring that RL optimization focuses on discriminative, answer-relevant visual tokens. Meanwhile, the self-diagnostic curriculum facilitates the model's ability to search, integrate, and analyze multiple clues before drawing an accurate conclusion.

Table 9: **Feedback source ablation.** Comparative performance on three video benchmarks (Qwen2.5-VL-7B, 3K QA training). The final row indicates our default setting.

Positive r^+	**Negative** r^-	**MMVU**	**VSI.**	**VideoMME**	**Avg.**
Peer rollout (right full)	Peer rollout (wrong full)	64.9	34.0	60.3	53.1
Peer rollout (right full)	Peer rollout (answer only)	65.5	34.8	60.4	53.6
Ground truth answer	Peer rollout (wrong full)	66.2	35.8	61.4	54.5
Ground truth answer	Peer rollout (random wrong answer)	67.3	37.0	61.8	55.4
Ground truth answer	Peer rollout (answer only)	**68.5**	**39.2**	**62.6**	**56.8**

Performance with different answer-aware feedback. We investigate the impact of incorporating the answer as privileged information to guide token-wise credit assignment. Table 9 reveals three main findings: First, utilizing the ground-truth answer as the positive reference (r^+) substantially outperforms relying solely on peer full-length rollouts (average 55.6% vs. 53.4%). This performance gap likely stems from noisy or erroneous tokens present in seemingly correct reasoning chains, which affects the quality of positive anchor. Second, regarding the negative reference (r^-), conditioning on privileged full-length rollout information underperforms directly providing the wrong answer. This indicates that concise error signals can refine the multi-role token set selection and yield more robust contrastive signals for training. Third, sampling negative answers from random incorrect choices rather than actual erroneous rollouts degrades performance. This suggests that effective continual learning requires information derived from the model's intrinsic states and feedback, further validating our design philosophy.

Parametric sensitivity to $\boldsymbol{\alpha}$. We investigate the impact of the loss weighting vector $\boldsymbol{\alpha} = [\alpha_{vis}, \alpha_{ans}]$, which balances the trade-off between the visual-perception and answer-aware components. As shown in Table 10, a balanced configuration ($\alpha_{vis} = \alpha_{ans} = 0.25$) consistently achieves superior performance across all benchmarks, whereas any skewed weight distribution leads to noticeable degradation. Specifically, when the answer-aware component heavily dominates visual perception (e.g., shifting the ratio to $0.35 : 0.15$), overall accuracy drops universally, with the most pronounced decline occurring on reasoning-intensive tasks. This behavior stems from the intrinsic bias of answer-dominated tokens: they tend to concentrate on definitive conclusions or logical transition cues, thereby circumventing the diverse visual exploration within intermediate chain-of-thought trajectories.

Table 10: **Ablation on hyperparameters α_{vis} and α_{ans}.** Performance comparison on VideoMMU and Video-Holmes benchmarks.

α_{vis}	α_{ans}	**VideoMMU**	**Video-Holmes**	**Avg.**
0.25	0.25	**61.0**	**42.7**	**51.9**
0.35	0.15	57.9	42.2	50.1
0.15	0.35	58.4	42.2	50.3

Sensitivity analysis on hyperparameters K_1 and K_2. We evaluate the impact of hyperparameters K_1 (window size) and K_2 (number of selected windows) on overall performance, as depicted in Table 11. As shown in Table 11(a), increasing the window size from 2 to 6 gradually improves performance from 59.8%

Table 11: (a) Effect of window size (K_1) comparing fixed sizes with adaptive selection. (b) Effect of the number of selected tokens for video frame localization (K_2).

(a) K_1: Window Size

Fix 2	**Fix 4**	**Fix 6**	**Fix 8**	**Adaptive**
59.8	61.8	62.1	61.4	**62.6**

(b) K_2: Number of Windows

15	**20**	**25**	**30**	**40**
61.9	**62.6**	62.1	61.5	59.3

to 62.1%, while the adaptive selection strategy achieves the best result of 62.6%, demonstrating the effectiveness of dynamically adjusting window sizes. Regarding K_2 in Table 11(b), performance follows an inverted U-shape with respect to the number of selected windows, reaching an optimal peak at $K_2 = 20$ with 62.6%. A smaller K_2 (e.g., 15) risks omitting essential visual context, whereas an excessively large K_2 (e.g., 40) introduces redundant frame noise that degrades performance to 59.3%. These findings confirm that our method achieves optimal performance with adaptive window sizing and moderate window counts, ensuring robustness across different configurations.

Effect of high-attention frames from different layers on training. Motivated by the findings in (Zhang et al., 2025), which demonstrate that attention tend to converge on task-relevant semantic regions in earlier-to-intermediate layers but progressively diverge in the deepest layers, we hypothesized that the terminal layers may introduce representational noise detrimental to precise visual grounding. To verify this, we conducted additional ablation studies in which we systematically excluded the terminal layers (specifically, the last five layers) when computing the attention flow. Our empirical results strongly corroborate this hypothesis: pruning the final few layers consistently yields the optimal model performance, whereas relying exclusively on these terminal layers generally results in different degree of degradation. This provides a stronger premise for localizing key frame videos from rollouts.

Table 12: Ablation study on selecting high-attention tokens/frames driven by different layer focus across five benchmarks. The base model is Qwen2.5-VL-7B.

Attention Focus (Layer Group)	MMVU	VSI	VideoMME	Video-Holmes	VideoMMMU
Low Layers *(e.g., L1–L8)*	66.8	36.6	61.6	43.6	51.1
Mid Layers *(e.g., L9–L20)*	67.0	38.8	62.1	44.2	52.8
Last Few Layers *(e.g., L21–L32)*	67.2	37.8	61.9	44.8	51.1
Ours (Removing Last Five Layers)	**68.5**	**39.2**	**62.6**	**44.8**	**53.6**

Perturbation strategies. Unlike image-based contrastive perception methods such as PAPO, we adopt diverse perturbations within videos, which require dynamic temporal understanding. Given an on-policy rollout $\mathbf{o} = (o_1, \ldots, o_M)$, the student's own response-to-frame attention selects a keyframe set $\mathcal{K} \subseteq \{1, \ldots, F\}$, where F denotes the total number of input frames. All negative-view perturbations act exclusively on the target frames $\mathcal{T}$ (by default $\mathcal{T} = \mathcal{K}$, optionally extended to all F frames), ensuring that corruption aligns with the temporal support actually utilized by the policy. The positive view $\mathcal{V}^+$ applies mild photometric jittering (brightness, contrast, saturation), while the negative view $\mathcal{V}^-$ removes visual information from $\mathcal{T}$. We consider three perturbation variants, all applied in pixel space with patch size $s \times s$ ($s = 14$).

(1) Random masking. We design two stochastic occlusion strategies that preserve temporal order while degrading appearance:

- **Random region occlusion**: Scatters n random rectangles per target frame $t \in \mathcal{T}$. Each rectangle's side lengths are sampled independently as

$$f_h, f_w \sim \mathcal{U}(\rho_{\min}, \rho_{\max}), \quad r_h = \lfloor f_h H \rfloor, \quad r_w = \lfloor f_w W \rfloor, \tag{15}$$

  where $\mathcal{U}$ denotes the uniform distribution. Rectangles are placed at uniformly random locations and filled with per-pixel RGB noise sampled from $\mathcal{U}(\min_t, \max_t)$ (the frame's dynamic range). We use $n = 50$, $\rho_{\min} = 0.05$, and $\rho_{\max} = 0.15$.
- **Single-block noise** (`region_noise`): Masks one contiguous rectangle covering a fraction $\rho = 0.5$ of each frame, filled with uniform noise.

(2) Attention-aware masking. These strategies implement "semantic-aware" corruptions targeting regions where the model focuses its attention. To identify critical regions, we compute an attention map $\mathbf{a}_t \in \mathbb{R}^{H_b \times W_b}$ by summing the cross-attention weights $a^{(l)}_{m,v}$ (defined in Eq. 3) over visual tokens in frame t, where $H_b = H/s$ and $W_b = W/s$ after merging. We select the top-ϕ tokens with highest attention scores, denoted $\mathcal{S}_t = \text{Top-}\lfloor \phi H_b W_b \rfloor(\mathbf{a}_t)$, and define their bounding box $\mathcal{B}_t$ as the *attention hotspot*.

We propose two hotspot-based corruption strategies:

- **Attention-focused occlusion**: Places all n noise rectangles strictly inside the hotspot $\mathcal{B}_t$ ($\phi = 0.2$ in our experiments), attacking regions the model relies on most heavily.
- **Curriculum occlusion**: Gradually expands the corruption scope. Each rectangle is placed randomly in the full frame with probability p (where p increases from 0 to 1 during training), and inside the hotspot $\mathcal{B}_t$ otherwise:

$$\text{region}(t) = \begin{cases} \text{full frame} & \text{with probability } p, \\ \mathcal{B}_t & \text{with probability } 1-p. \end{cases} \tag{16}$$

  This anneals from purely attention-focused corruption ($p \approx 0$, hard but localized) to purely random corruption ($p = 1$, easier but global), enabling progressive adaptation to diverse noise patterns.

(3) Temporal order destruction. To disrupt sequential dependencies, we shuffle the keyframe slots using a derangement permutation π, where $\mathcal{V}_t^- = \mathcal{V}_{\pi(t)}$ for $t \in \mathcal{K}$. This breaks the temporal order while keeping the spatial content of each frame intact, which is different from permuting tokens within a single frame.

In our training, we apply temporal shuffling and masking-based corruptions independently with a probability of 0.5. Table 13 shows the results of different perturbation strategies. We find that **random region occlusion** works best overall. This aligns with previous studies on masked visual modeling, which show that random masking is often better than attention-based masking. We explain this with two reasons:

- **Attention-focused masking removes a single salient block.** This leaves too much context in the negative view, allowing the model to guess the answer from the surroundings. As a result, the contrastive signal is weaker on appearance-heavy tasks like MMVU (67.04 vs. 68.16) and VideoMME (61.91 vs. 62.37). It only helps slightly on VSI (38.55 vs. 37.67), where forcing the model to look away from the most attended region is beneficial.
- **Scattering many small random rectangles degrades information more evenly.** Instead of removing an entire object, this approach creates a harder and better-calibrated negative sample. The curriculum variant starts with attention-focused masking and inherits its weakness, so it does not outperform plain random occlusion.

Negative-view perturbation strategy	MMVU	VSI	VideoMME
Random region occlusion	**68.16**	**38.55**	**62.37**
Attention-focused occlusion	67.04	37.67	61.91
Curriculum (attention→full region)	66.88	37.55	62.03

Table 13: **Impact of the negative-view (masking) strategy** on our method. All rows share the same base model, data, and hyperparameters and differ only in how the negative view is corrupted.

D TRAINING DYNAMICS

Reward, response length, and loss analysis. We analyze the optimization behavior of our method in Figure 2, highlighting three key observations regarding the interplay between perception supervision and policy learning.

(1) The perception objective exhibits stable convergence with an initial adaptation phase. As shown in Figure 2(a), the MRC loss experiences a transient spike within the first ~ 50 steps as the model adapts to the contrastive views. Following this brief adjustment, the loss decreases monotonically and stabilizes near 0.03. This trajectory confirms that the auxiliary perception objective converges smoothly without causing destructive interference with the primary policy-gradient updates.

(2) The contrastive objective effectively enforces representation separation. Figure 2(b) decomposes the embedding space dynamics by tracking distances from the anchor to positive (d_{pos}) and negative (d_{neg}) views. Throughout training, d_{pos} remains consistently and significantly lower than d_{neg} (maintaining a gap of ~ 0.015). This clear separation demonstrates that the model reliably pulls

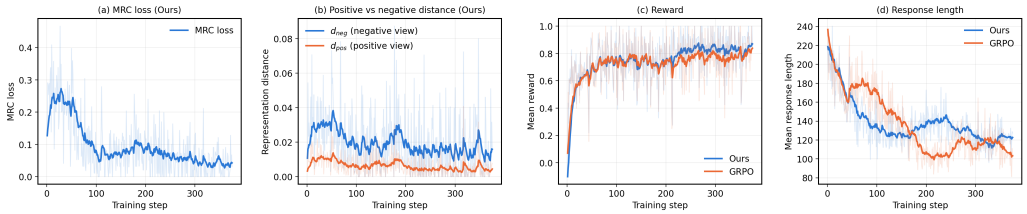


Figure 6: **Left to right**. (a) Convergence curve of $\mathcal{L}_{MRC}$ loss. (b) Distance metrics of positive vs. negative samples relative to the anchor. (c) Reward trajectory. (d) Variations in reasoning chain length.

Table 14: **Per-step training efficiency and computational overhead breakdown across different model scales.** All time metrics are measured in seconds (s) per training step. The extra computational overhead introduced by our method is modest ($+12\% \sim +40\%$) and is almost entirely dominated by the forward pass of the teacher model during policy updates.

Base Model	Method	Per-Step Time Breakdown (s)				Overhead vs. GRPO	
		Rollout Gen.	Actor Update	Extra Fwd.	Avg. / Step	Absolute (Δt)	Relative (Δ)
Qwen3-VL-2B	GRPO	16.0	14.9	0.0	38.1	–	–
	Ours	16.5	22.0	7.5	46.6	+8.5s	+22.3%
Qwen2.5-VL-3B	GRPO	19.2	20.8	0.0	49.3	–	–
	Ours	17.7	29.0	10.4	55.2	+5.9s	+12.0%
Qwen3-VL-4B	GRPO	15.6	20.5	0.0	45.9	–	–
	Ours	18.2	34.3	11.9	64.2	+18.3s	+39.9%
Qwen2.5-VL-7B	GRPO	19.8	30.2	0.0	62.6	–	–
	Ours	20.1	48.3	17.1	81.6	+19.0s	+30.4%

answer-consistent views closer while pushing perturbed negatives away, validating the efficacy of the contrastive term in structuring the visual-textual representation space.

(3) Perception supervision is reward-aligned yet promotes concise reasoning. Crucially, adding the perception objective does not compromise reinforcement learning performance. In Figure 2(c), our method's reward trajectory closely aligns with the GRPO baseline, indicating that the auxiliary loss is effectively reward-neutral. However, the response-length dynamics in Figure 2(d) reveal a distinct advantage: while GRPO struggles to compress outputs (stabilizing around 140–160 tokens), our method drives a more aggressive reduction in length, converging to a more concise $\sim$120 tokens. We attribute this to the perception objective encouraging the model to discard redundant reasoning steps while retaining essential evidence-grounded content, leading to more efficient inference without sacrificing accuracy.

Training efficiency and computational overhead. To evaluate the practical overhead introduced by our method, we report a step-wise time breakdown across various model architectures (2B $\sim$ 7B) in Table 14. Compared to the standard GRPO baseline, PAPO incurs a modest increase in total step time, ranging from $+12.0\%$ ($+5.9$s on Qwen2.5-VL-3B) to $+39.9\%$ ($+18.3$s on Qwen3-VL-4B). Importantly, rollout generation times remain virtually unchanged across all scales, indicating that our method introduces no latency bottleneck during context sampling. Crucially, a closer inspection reveals that the additional step latency is almost entirely concentrated in the actor update phase, which aligns closely with the forward inference time of the teacher model (e.g., 18.1s update increment vs. 17.1s teacher forward time on 7B). This confirms that our mechanism maintains modest computational efficiency without imposing unmanageable optimization burdens.

E Additional Analysis on Related Work

We present a fine-grained comparison with Video-KTR to underscore our distinct contributions. Our method differs from Video-KTR in three pivotal aspects: **(1) Token selection vs. reward reweighting.** Unlike Video-KTR, which directly integrates score-based reweighting into the GRPO objective, we eschew this strategy. Our empirical analysis reveals that such reweighting yields suboptimal performance on specific datasets, as it tends to amplify visual noise and bias the model

towards the base policy. Instead, we leverage scores exclusively to identify salient tokens without explicitly modulating their relative importance, thereby enhancing robustness against visual noise. **(2) Multi-dimensional token analysis.** We perform a distributional quantification of diverse token types with positional preferences, introducing answer-aware perturbation as a critical metric. Furthermore, we trace the origins of high entropy to low-entropy visual activations and characterize the decoding patterns associated with these tokens. Additionally, our self-diagnostic curriculum masking mechanism stabilizes the training process, delivering consistent performance gains across multiple benchmarks. **(3) Superior performance.** Our method demonstrates superior efficacy, consistently outperforming Video-KTR across a comprehensive range of evaluation metrics.

Algorithm 1 DyCPO: Self-Diagnostic Multi-Role Policy Optimization

Input: π_θ, π_ϕ (EMA), $\mathcal{V}$, q, G, $p = 0.2$, $\alpha_{\text{vis}} = \alpha_{\text{ans}} = 0.5$, $\lambda_{\text{MRC}} = 0.05$, $\tau = 1$, $\alpha = 0.99$
Output: Updated π_θ

for each iteration **do**
  Sample $\{\mathbf{o}^{(i)}\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid \mathcal{V}, q)$; compute A_i; partition into $\mathcal{R}_r$, $\mathcal{R}_w$
  // Visual counterfactuals
  Compute attention $\mathcal{A}_{m,t}^{(i)}$; select top-K_1 tokens $\Phi_t^{(i)}$; compute $\bar{\mathcal{A}}_t^{(i)}$; select top-$K_2$ frames $\Psi^{(i)}$
  Construct $\mathcal{U}^* = \Psi^+ \setminus \Psi^-$; generate $\mathcal{V}^+$ (jitter), $\mathcal{V}^-$ (occlusion on $\mathcal{U}^*$)
  // Answer counterfactuals
  Extract a^+ from $\mathcal{R}_r$, a^- from $\mathcal{R}_w$
  // Forward passes
  for each $\mathbf{o}^{(i)}$, m **do**
    $P_m^{(i)} \leftarrow \pi_\theta(\cdot \mid \mathcal{V}, q, \mathbf{o}_{<m}^{(i)})$; $\hat{P}_m^{(i)} \leftarrow \pi_\theta(\cdot \mid \mathcal{V}^+, q, \mathbf{o}_{<m}^{(i)})$; $Q_m^{(i)} \leftarrow \pi_\theta(\cdot \mid \mathcal{V}^-, q, \mathbf{o}_{<m}^{(i)})$
    $P_m^{+,(i)} \leftarrow \pi_\theta(\cdot \mid \mathcal{V}, q, a^+, \mathbf{o}_{<m}^{(i)})$; $P_m^{-,(i)} \leftarrow \pi_\theta(\cdot \mid \mathcal{V}, q, a^-, \mathbf{o}_{<m}^{(i)})$
  end for
  // Multi-role selection
  for each m **do**
    Compute J_m^{vis}, ΔH_m^{vis}, J_m^{ans}, ΔH_m^{ans}, H_m; normalize
    $g_m \leftarrow 1 - \prod_{c\in\{\text{vis,ans}\}}(1 - \hat{J}_m^c)^{\alpha_c}(1 - |\widehat{\Delta H^c}_m|)^{\alpha_c}$; $c_{i,m} \leftarrow g_m \cdot \hat{H}_m$
  end for
  $\mathcal{C} \leftarrow \{m : c_{i,m} \in \text{Top-}p\} \cap \{m : A_i > 0\}$
  // Contrastive optimization
  if $\mathcal{C} \neq \varnothing$ **then**
    for each $m \in \mathcal{C}$ **do**
      $d_{i,m}^+ \leftarrow D_{\text{KL}}(P_m^{(i)} \| \hat{P}_m^{(i)})$; $d_{i,m}^- \leftarrow D_{\text{KL}}(P_m^{(i)} \| Q_m^{(i)})$
      $\ell_{i,m}^{\text{MRC}} \leftarrow \text{softplus}((d_{i,m}^+ - d_{i,m}^-)/\tau)$
    end for
    $\mathcal{L}^{\text{MRC}} \leftarrow \frac{1}{|\mathcal{C}|}\sum_{m\in\mathcal{C}} \ell_{i,m}^{\text{MRC}}$
  else
    $\mathcal{L}^{\text{MRC}} \leftarrow 0$
  end if
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{GRPO}} + \lambda_{\text{MRC}}\mathcal{L}^{\text{MRC}}$; $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{total}}$; $\phi \leftarrow (1-\alpha)\phi + \alpha\theta$
end for