REINFORCING VIDEO REASONING VIA FUSION-AWARE CHAIN-OF-THOUGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has recently been extended to video reasoning, where perceptual grounding must be integrated with reasoning across space and time. However, Group Relative Policy Optimization propagates a single outcome-based advantage across all tokens, assigning similar credit to decision-promoting and decision-undirected information within Chain-of-Thought. Our analysis reveals a fundamental contrast: whereas image–text RLVR concentrates probability mass and progressively reduces trajectory entropy, video sustains correct trajectories across high probability–entropy bands. This observation suggests that video reasoning benefits from preserving uncertainty during temporal evidence accumulation and reducing it during reasoning. We therefore propose **P**erception–**R**easoning **F**usion-**A**ware **P**olicy **O**ptimization (**PR-FAPO**), which derives a perception–reasoning fusion gap from probability–entropy statistics over perceptual-prefix and reasoning-suffix segments. PR-FAPO maps this signal to token-wise discount coefficients, enabling fusion-aware credit assignment that promotes informative early perception and selectively fuses articulated visual information into subsequent reasoning. Experiments on reasoning-centric and temporal-grounding benchmarks validate the effectiveness of PR-FAPO, demonstrating performance improvements together with improved training stability.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) has powerful paradigm for improving reasoning in LLMs OpenAI (2025a); Anthropic (2025); Guo et al. (2025) and MLLMs Yang et al. (2025a); Wang et al. (2025b); Zheng et al. (2025b). In particular, Group Relative Policy Optimization (GRPO) advances mathematical reasoning, decision making, and multimodal understanding by optimizing Chain-of-Thought (CoT) trajectories with outcome-level rewards, and has been extended to Video-Language Models. Unlike text or image-centric reasoning, video must locate, track, integrate reasoning, and cumulative evidence dispersed across space and time, thereby tightly coupling perceptual grounding with the reasoning that builds upon it.

This coupling exposes a limitation of outcome-level GRPO. Within a successful trajectory, decision-promoting and decision-undirected visual information receive similar positive credit, reinforcing both useful reasoning patterns and spurious correlations. Recent studies Zhu et al. (2026a); Tian et al. (2026); Wu et al. (2026) suggest that visual CoT benefits reasoning primarily through structured predictive supervision that promotes reasoning-relevant visual representations. Motivated by this insight, rather than improving perceptual recognition itself, we target on *reasoning policy*, encouraging selective fusion of decision-promoting based on articulated visual information into subsequent reasoning while suppressing decision-undirected information, seeking to achieve this without additional perceptual annotations or auxiliary latent objectives Yao et al. (2026a). From this perspective, improving reasoning quality can indirectly guide perception policy by encouraging utilization of decision-promoting articulated visual information, without requiring costly perceptual supervision. The perception policy first articulates visual information as a language prefix, upon which the reasoning policy autoregressively generates a reasoning suffix. Effective training should therefore favor next-token predictions that selectively fuse decision-promoting articulated visual information from the prefix into subsequent reasoning. However, outcome reward alone cannot determine whether a correct answer is derived from articulated visual information or from spurious correlations. Consequently, trajectory-level GRPO advantages provide limited insight into how articulated visual information contributes to successful reasoning. To better understand this process, we investigate *where* articulated visual information influences reasoning suffix and *how confidently* it supports the corresponding next-token predictions. This leads to our central question: *Does the GRPO advantage alone naturally promote such reasoning behavior in video domain?*

Entropy has recently emerged as a critical signal for policy uncertainty and advantage shaping in

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.

GRPO Deng et al. (2025b), yet remains underexplored for video. Empirically, under *think→answer* paradigm of Video-R1 Feng et al. (2025) with shuffled temporal anchors, it does *not* reach low-entropy solutions common in image–text RLVR Wang et al. (2025h); Huang et al. (2025b); Hu et al. (2026); trajectories instead drift into a high-probability, high-entropy regime that still attains high rewards (Fig. 1), suggesting that video reasoning requires *selectively preserving* entropy at perceptual tokens rather than collapsing it uniformly. VideoAuto-R1 Liu et al. (2026b) introduces an *answer→think→answer* paradigm that separates direct- and revised-answer rewards and achieves strong performance on reasoning tasks. Nevertheless, these outcome-level rewards still rescale a single trajectory-level advantage within each group, may lead to over-suppressing reasoning ability. Moreover, we observe that it tends to converge to low-entropy solutions with substantially shorter CoTs, accompanied by policy shifts that extend beyond trust region Schulman et al. (2015), weakening stabilization and potentially leading to unfaithful policy updates. Through an object-level analysis, we further show that both perception and reasoning components can, in principle, be governed by a synergistic, step-wise advantage signal. Applying GRPO directly to video reasoning, however, obscures this mechanism by replacing it with a nearly uniform advantage. This limitation extends across token positions as well. Tokens responsible for visual-information acquisition, intermediate reasoning, and final-answer commitment are subject to nearly uniform optimization pressure, even though their functions evolve throughout CoT. Moreover, because clipped importance ratios typically remain close to one, clipping is largely inactive and provides limited capacity to redistribute this pressure or reshape token-wise entropy Luo et al. (2026); Cheng et al. (2026); Zhou et al. (2026). GRPO therefore offers insufficient resolution to distinguish the evolving roles and contributions of trajectory fragments.

Motivated by this analysis, we propose **P**erception–**R**easoning **F**usion-**A**ware **P**olicy **O**ptimization (**PR-FAPO**), an advantage-shaping framework that encourages informative early perception and articulated-visual-information-fused subsequent reasoning. Based on analysed, PR-FAPO aggregates probability–entropy statistics over prefix and suffix segments of each trajectory into a compact *perception–reasoning fusion gap*: a scalar that quantifies *how strongly* perceptual-prefix confidence fuses to suffix reasoning certainty. A token-wise mapping then determines *where* along trajectory this signal rescales GRPO advantage, yielding fusion-aware credit assignment that progressively reduces reasoning uncertainty through perception at early positions. Extensive experiments on nine video benchmarks show that PR-FAPO improves reasoning reviewed answer baseline and improves GRPO applying for video reasoning, yielding more structured probability–entropy dynamics and stable optimization. Our contributions are as follows: ①We identify a broad high-probability, high-entropy regime in video RLVR, contrasting with the entropy collapse commonly observed in image–text RLVR. This finding suggests that video reasoning may require role- and position-dependent entropy regulation rather than uniform credit assign. ②We propose PR-FAPO, which derives token-wise discount coefficients from probability–entropy statistics over trajectory segments to enable fusion-aware credit assignment. ③PR-FAPO improves reasoning reviewed answer baselines across seven benchmarks spanning video perception, reasoning, and temporal grounding.

2 PRELIMINARY

For a query q ground truth answers a^* on a textual and video C from dataset D. The GRPO Shao et al. (2024) the gradient with respect as follows:

$$\nabla_\theta \mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\{q,C,a^*\}\sim D,\{y_i\}_{i=1}^{G}\sim\pi_{\text{old}}} \frac{1}{\sum_{i=1}^{G}|y_i|}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|} \Bigg[\underbrace{\left(r_{i,t}(\theta)\hat{A}_i\cdot\Big(clip(r_{i,t}(\theta),1-\epsilon_l,1+\epsilon_h)\hat{A}_i\Big)+\beta\frac{\pi_{\text{ref}}(y_{i,t})}{\pi_\theta(y_{i,t})}-\beta\right)}_{w_{i,t}}\Bigg]\cdot\nabla_\theta\log\pi_\theta(y_{i,t}\mid q,C,y_{i,<t}) \tag{1}$$

where $r_{i,t}(\theta)=\frac{\pi_\theta(y_{i,t}|q,C,y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|q,C,y_{i,<t})}$ is importance ratio, and $\{y_i\}_{i=1}^{G}$ denotes a rollout group of G responses. The clipping bounds ϵ_l,ϵ_h constrain the policy update, while the KL regularization coefficient is set to $\beta=0.3$. Each response receives a binary outcome reward $R_i\in\{0,1\}$ based on agreement with the ground-truth answer a^*, from which group-normalized advantage is

$$\hat{A}_i^{GRPO}=\frac{R_i-\text{mean}\big(\{R_k\}_{k=1}^{G}\big)}{\text{std}\big(\{R_k\}_{k=1}^{G}\big)} \tag{2}$$

Entropy quantifies policy uncertainty at both the token and trajectory levels. Excessively high entropy

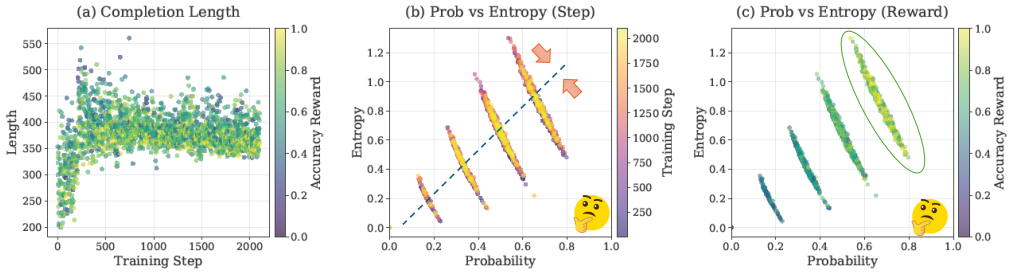


Figure 1: **Video-R1 training dynamics of trajectory propensities.** **(a)** Shorter-than-average responses receive disproportionately higher rewards. **(b)** Training increasingly favors **correct trajectories in which groups mean-rewards exceed** 0.5 that maintain both high-probability-entropy. **(c)** Correct trajectories are colored by their group mean reward, **rescaled from** $[0.5, 1]$ **to** $[0, 1]$.

can lead to unstable outputs Jin et al. (2025); Jiang et al. (2025); Hao et al. (2026), whereas excessively low entropy may result in premature convergence Li et al. (2025b); Zawistowski (2024). Motivated by its simplicity and effectiveness, we investigate entropy in the video modality (further motivation for applying Entropy as Appendix J). Following prior studies in the text domain Wang et al. (2025e); Sharma & Chopra (2025), we define

$$\mathcal{H}_t = -\sum_{y_t \in \mathcal{V}} p(y_t \mid y_{<t}) \log p(y_t \mid y_{<t}), \quad \mathcal{H}_i = \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \mathcal{H}_t. \tag{3}$$

$\mathcal{H}_{i,t}$ measures predictive uncertainty at step t, while $\mathcal{H}_i$ denotes mean uncertainty over trajectory y_i.

3 Analyzing Chain-of-Thought Reasoning

3.1 Empirical Training Dynamics

Setup. To characterize the reasoning behaviors favored within successful rollouts, we compare Video-R1 Feng et al. (2025) and VideoAuto-R1 Liu et al. (2026b) under the same Qwen2.5-VL-7B and GRPO framework. We adopt Video-R1's temporal-relevance reward (excluding length reward) and VideoAuto-R1's answer-consistency reward scaling, while following VideoAuto-R1 in omitting SFT warm-start initialization. Although both methods share same outcome reward, Video-R1 shapes advantages through competition among shuffled temporal anchors, whereas VideoAuto-R1 rescales advantages according to answer consistency without temporal information. This controlled comparison isolates the effects of temporal information and consistency-based reward shaping on learned reasoning behavior. During training, we analyze correct trajectories from groups with mean rewards above 0.5, capturing the reasoning patterns favored by each policy within successful rollout distributions. Additional details are provided in Appendix C.

Comparing and Contrasting We analyze rollout behavior from two complementary perspectives: response length and trajectory-level entropy–probability statistics, as illustrated in Fig. 1. Under the reason-then-answer paradigm of Video-R1, correct trajectories sampled from groups with a mean reward above 0.5 form progressively shifted entropy–probability bands, with later rollouts achieving a more favorable balance between trajectory probability and entropy (Fig. 1, b). Moreover, comparison across these bands shows that bands occupying higher probability and entropy ranges are associated with higher group-mean rewards (Fig. 1, c). Furthermore, shorter responses tend to receive higher rewards (Fig. 1, a). Together, these observations are consistent with an implicit bias toward particular entropy–probability regimes and toward shorter correct trajectories. To ensure a fair comparison, we evaluate VideoAuto-R1 Liu et al. (2026b) under same Video-R1 introduced video-only training dataset, since its original dataset comprise a mixture of text, image, and video modalities. Although it adopts a comparable GRPO setup and a similar system prompt (Table 3 in Appendix L) through re-answering framework, its trajectory entropy rapidly collapses to a low-entropy regime without exhibiting structured entropy bands observed in Video-R1 (Fig. 2). This suggests that consistency-based reward scaling alone is insufficient to shape desirable CoT behavior. The resulting CoTs are often too brief to reflect video content, showing premature termination, coarse reasoning, or inappropriate use of visual evidence (Appendix Fig. 14); when CoT-conditioned answers do not outperform direct answers, the policy increasingly drifts beyond trust region, suggesting a breakdown of the clipping-based stabilization mechanism and potentially unfaithful policy updates (see Appendix R for additional stability analyses). Consistently, CoT length drops sharply within a few training iterations. Rather than selectively pruning redundant reasoning, optimization appears to under-develop CoT reasoning policy while still improving perception. These observations are consistent

with the hypothesis that consistency-based objective admits solutions strengthening perception-related behaviors without integrating visual evidence into reasoning, motivating a more explicit mechanism for shaping reasoning over video. (Further modality-specific and backbone-specific empirical analyses are provided in Appendix A and Appendix B, respectively.)

The behavior of Video-R1 during RL training is counterintuitive from the perspective of both text-centric Guo et al. (2025); Huang et al. (2025a); He et al. (2025) and image Wang et al. (2025h); Huang et al. (2025b); Hu et al. (2026) RLVR, where performance typically saturates as probability mass concentrates and the policy over-commits to a narrow region of solutions Beukman et al. (2026); Park et al. (2025), motivating explicit entropy-preserving techniques Li et al. (2026a); Petrenko et al. (2026); Li et al. (2026b). Additional cross-modal comparisons are provided in Appendix B. Video-R1 adopts none of them, yet sustains high entropy throughout training. As shuffled temporal-anchor reward is primary difference, we hypothesize that it induces an optimization dynamic intrinsically resistant to entropy collapse—a mechanism central to its scaling behavior. This behavior may arise from video reasoning entails a substantially larger and more temporally extended perceptual state space than text- or image-based tasks. Even under shuffled temporal anchors used in Video-R1, GRPO provides only weak supervision for credit assignment over this structure, so policy continues to explore a vast perception–reasoning space. At same time, observing suggest that this sustained exploration is not purely detrimental: trajectory entropy remains persistently high, with probability mass concentrated on behaviors that keep exploration productive for video reasoning.

3.2 Analyzing on Learning Object

To explain why training gradually favors trajectories that simultaneously balance high entropy and high probability across various regions, we relate the trajectory-level update to a coupled *perception–reasoning* gradient. Let $s_{i,t} = (q, C, y_{i,<t})$ denote the prefix state, and let $z_{i,t}$ denote a latent perceptual abstraction extracted from the video context C. Instead of considering the next-token likelihood in isolation, we work with the joint score of the perceptual abstraction and the generated token, which admits the following decomposition:

$$\nabla_\theta \log \pi_\theta(y_{i,t}, z_{i,t} \mid s_{i,t}) = \underbrace{\nabla_\theta \log \pi_\theta(z_{i,t} \mid s_{i,t})}_{\text{Contextual reference}} + \underbrace{\nabla_\theta \log \pi_\theta(y_{i,t} \mid s_{i,t}, z_{i,t})}_{\text{Conditional generation}}. \tag{4}$$

The two terms play distinct roles: the *contextual-reference* term controls which perceptual evidence is referenced, whereas the *conditional-generation* term controls how the next token is produced once that evidence is fixed. Because $z_{i,t}$ is policy-dependent and is *not* directly supervised, the contextual-reference term cannot be identified from final outcome-only rewards. This absence of an explicit target is the structural origin of the credit-assignment issue analyzed below.

Lemma 3.1. Following the entropy-update analysis of Cui et al. (2025), one training step λ with learning rate η satisfies (see Appendix E for the derivation)

$$\mathcal{H}\big(\pi_\theta^{\lambda+1}\big) - \mathcal{H}\big(\pi_\theta^{\lambda}\big) \approx -\eta \, \mathrm{Cov}_{z_a \sim \pi_\theta^\lambda(\cdot|S_c)} \left(\log \pi_\theta^\lambda(z_a \mid S_c),\ \pi_\theta^\lambda(z_a \mid S_c) \cdot \underbrace{r_{i,t}(\theta)\, \hat{A}(S_c, z_a)}_{w_{i,t}} \right), \tag{5}$$

where $S_c \sim d_{\pi_\theta}$ and $z_{s,a}$ denotes the action logit, distinct from the perceptual latent $z_{i,t}$. Eq. 5 shows that entropy changes only when the weighted update covaries with log probability; hence, nearly action-uniform logit changes yield a small covariance, allowing high-entropy perceptual latent states $z_{i,t}$ to persist on increasingly rewarded trajectories. Moreover, because the same scalar $w_{i,t}$ rescales both terms in Eq. 4, outcome-only feedback cannot distinguish improved perceptual grounding from increased conditional generation likelihood.

Per-token gradient norm and entropy-dependent deviations. To make this effect quantitative, let $y_{i,t}^*$ denote the target token at index k (so $y_{i,t}^k = y_{i,t}^*$ and $z_{i,t}^k = z_{i,t}^*$), and let $\mathbf{a} \in \mathbb{R}^{|\mathcal{V}|}$ be its logit vector. The squared ℓ_2-norm of the gradient then decomposes as

$$\left\| \nabla_a J(y_{i,t}^*) \right\|^2 = |w_{i,t}|^2 \Big(\left\| \nabla_a \log \pi_\theta(z_{i,t}) \right\|^2 + 2 \langle u, v \rangle + \left\| \nabla_a \log \pi_\theta(y_{i,t}) \right\|^2 \Big), \tag{6}$$

where, for brevity, $\langle u, v \rangle = \langle \nabla_a \log \pi_\theta(z_{i,t}), \nabla_a \log \pi_\theta(y_{i,t}) \rangle$. Applying the Gradient Norm Decomposition of Yang et al. (2025b); Liu et al. (2026a) and summing over $n \in \{1, \ldots, |\mathcal{V}|\}$ yields

$$\left\| \nabla_a \log \pi_\theta(y_{i,t}) \right\|^2 = |w_{i,t}|^2 \left(1 - 2\pi_\theta\big(y_{i,t}^k \mid z_{i,t}^*\big) + \sum_{n=1}^{|\mathcal{V}|} \pi_\theta\big(y_{i,t}^n \mid z_{i,t}^*\big)^2 \right), \tag{7}$$

and analogous sum over a discrete perceptual reference set $\{z_m\}_{m=1}^{|\mathcal{O}|}$ (For clarity, reference set considered in this work is temporal-centric rather than spatial) extracted from $C \in \mathbb{R}^{T\times H\times W}$ gives

$$\|\nabla_a \log \pi_\theta(z_{i,t})\|^2 = |w_{i,t}|^2 \left(1 - 2\pi_\theta\big(z^k_{i,t} \mid q, C, y_{i,<t}\big) + \sum_{m=1}^{|\mathcal{O}|} \pi_\theta\big(z^m_{i,t} \mid q, C, y_{i,<t}\big)^2\right). \quad (8)$$

Unlike Eq. 7, Eq. 8 is not anchored by an explicit supervised target and instead aggregates spatio-temporal dependencies; the perceptual score is therefore statistically harder to identify than the reasoning score. Substituting Eq. 7 into Eq. 6 and bounding $\sum_n \pi_\theta(y^n_{i,t})^2$ in terms of $\mathcal{H}(\pi_\theta)$ via Jensen's inequality (upper) and a polynomial relaxation (lower) yields the entropy-dependent deviations (see Appendix A, Eqs. 37, 40):

$$\Delta_{\mathrm{ent}^H} = |w_{i,t}|^2 \left[\sum_{n=1}^{|\mathcal{V}|} \pi_\theta\big(y^n_{i,t}\big)^2 - e^{-\mathcal{H}(\pi_\theta)}\right] \geq 0, \quad (9)$$

$$\Delta_{\mathrm{ent}^L} = |w_{i,t}|^2 \left[\sum_{n=1}^{|\mathcal{V}|} \pi_\theta\big(y^n_{i,t}\big)^2 + \tfrac{|\mathcal{V}|-1}{|\mathcal{V}|\,(\log|\mathcal{V}|)^2}\,\mathcal{H}(\pi_\theta)^2 - 1\right] \leq 0. \quad (10)$$

Proposition 3.2. Using the deviation terms introduced above, the perception–reasoning gradient admits the two-sided bound (see Appendix D for the derivation):

$$\begin{aligned} |w_{i,t}|^2\Big(\|\nabla_a \log \pi_\theta(z_{i,t})\|^2 + 2\langle u, v\rangle\Big) &\leq \big\|\nabla_{\mathbf{a}} J\big(y^*_{i,t}\big)\big\|^2 - |w_{i,t}|^2\,\Delta_{\mathrm{ent}^H}, \\ |w_{i,t}|^2\Big(\|\nabla_a \log \pi_\theta(z_{i,t})\|^2 + 2\langle u, v\rangle\Big) &\geq \big\|\nabla_{\mathbf{a}} J\big(y^*_{i,t}\big)\big\|^2 - |w_{i,t}|^2\,\Delta_{\mathrm{ent}^L}. \end{aligned} \quad (11)$$

Inequality equation 11 shows that token probability and entropy jointly modulate perception–reasoning gradient through Δ_{ent^H} and Δ_{ent^L}, whose magnitudes determine the tightness of the bounds. Consistent with Fig. 1(b,c), training shifts toward high-probability, high-entropy trajectories that correlate with higher reward. As such trajectories become more prevalent, their aggregate contribution increasingly dominates trajectory-level update, motivating a position-dependent analysis of where entropy should be retained or reduced. As the policy sharpens ($\mathcal{H}(\pi_\theta) \to 0$), both deviations and $\nabla_{\mathbf{a}}\Delta_{\mathrm{ent}^H}$ vanish, so Eq. 11 collapses to (More details shown as appendix D.1)

$$\big\|\nabla_{\mathbf{a}} J(y^*_{i,t})\big\|^2 \;\longrightarrow\; |w_{i,t}|^2\big(\|\nabla_a \log \pi_\theta(z_{i,t})\|^2 + 2\langle u, v\rangle\big). \quad (12)$$

Eq. 12 characterizes a sharp policy update dominated by interaction between perception and reasoning, progressively fusing query-relevant visual information into accumulated reasoning context until the reasoning policy contributes no further action. Together with Eq. 11, this suggests a position-dependent profile: entropy should be retained at early positions to support evidence accumulation, but reduced at late positions to enable a confident prediction grounded in consolidated visual reasoning while maintain high probability. However, GRPO does not explicitly impose this high-to-low entropy transition into group-normalized advantage estimation.

Locating the actionable channel via logit perturbation. Let $\mathbf{a}_{i,t} \in \mathbb{R}^{|\mathcal{V}|}$ denote logits at $s_{i,t} := (q, C, y_{i,<t})$. A first-order log-softmax expansion identifies where a scalar can modulate the update without changing its direction (see Appendix F for derivation):

$$\delta \log \pi_\theta(y_{i,t} \mid s_{i,t}) \;=\; \delta\mathbf{a}_{i,t}(y_{i,t}) \;-\; \mathbb{E}_{y'_{i,t}\sim\pi_\theta(\cdot|s_{i,t})}\big[\delta\mathbf{a}_{i,t}(y'_{i,t})\big] \;+\; O\big(\|\delta\mathbf{a}_{i,t}\|^2\big), \quad (13)$$

which holds independently of optimization procedure. A policy-gradient update $\delta\theta$ with reward-to-go $w_{i,t} = \sum_{t'=t}^{|y_i|} \gamma^{t'-t} r_{i,t'}$ Sutton & Barto (2018) maps this parameter change, via perception–reasoning trade-off coefficient ζ and Eq. 4, to

$$\delta\mathbf{a}_{i,t} \;=\; \zeta \sum_{t'=t}^{|y_i|} \gamma^{t'-t}\, r_{i,t'}\big(\nabla_{\mathbf{a}_{i,t}} \log \pi_\theta(z_{i,t} \mid s_{i,t}) \;+\; \nabla_{\mathbf{a}_{i,t}} \log \pi_\theta(y_{i,t} \mid s_{i,t}, z_{i,t})\big). \quad (14)$$

Reading Eq. 14 with Eq. 13 reveals a clean separation of roles: prefix-local scalar $\zeta\, w_{i,t}$ sets only the *magnitude* of token update, rescaling both $\delta\mathbf{a}_{i,t}$ and first-order change in $\log \pi_\theta(y_{i,t} \mid s_{i,t})$, whereas its *direction* follows two score-gradient components in Eq. 4. Single scalar can coherently amplify both components, making a token-wise linear factor on $\zeta\, w_{i,t}$ natural intervention point for credit

allocation, enabling support to vanish reasoning direction gap (see Appendix F.1 for the derivation). The above analysis explains why directly applying GRPO to video reasoning is problematic, even within a re-input-and-regenerate pipeline, model predicts an initial answer a^1 from (q, C), generates a reflection trace y_i, and produces a revised answer a^2. Faithful reflection requires not only outcome correctness but also consistency between revised answer and visual evidence:

$$\mathrm{D}_{\mathrm{KL}}\big(\pi_\theta(a^2, z_i^2 \,|\, q, C, a^1, y_i) \,\big\|\, \pi_\theta(a^1, z_i^1 \,|\, q, C)\big) \longrightarrow 0. \tag{15}$$

This consistency requires *traceability*: credit assigned to the revised answer must propagate to the earlier tokens that establish its perceptual support. Without traceability, answer-level consistency may improve through linguistic confidence rather than visual grounding, yielding a policy that is consistent in form but ungrounded in content. Appendix H formalizes this requirement through an Evidence Upper Bound Hinton et al. (1995); Grosse et al. (2015) and shows that step-wise linear utility tightens the bound, reducing perception–reasoning fusion gap. The above analysis shows that GRPO offers limited support for such traceability: **(i) Role-blind credit propagation.** GRPO uses a group-relative, trajectory-level bandit advantage $\hat{A}_i$ rather than a step-wise one (see Appendix I, following Yao et al. (2026b)). As a result, perceptual-reference, intermediate-reasoning, and final-answer tokens receive same outcome-derived credit despite their distinct roles in Eq. 15. **(ii) Position-blind optimization pressure.** The clipped important ratio $r_{i,t}(\theta)$ does not recover this resolution: it mainly bounds the update magnitude and is rarely activated on-policy Chen et al. (2025a); Xi et al. (2025); Fu et al. (2026); Li et al. (2026c). It therefore neither distinguishes token roles nor provides position-wise guidance on where entropy should be retained or reduced Luo et al. (2026). Taken together, our analysis and investigation suggests that GRPO lacks an explicit mechanism for promoting the perception-reasoning policy fusion characterized in Eq. 12.

4 METHOD

As analyzed in Section 3.2, GRPO suffers from role-blind credit propagation and position-blind optimization, and Appendix.C and R.1 shows that token-level entropy transfers poorly to video setting. We therefore propose *Perception–Reasoning Fusion-Aware Policy Optimization* (PR-FAPO). Intuitively, accumulated context by temporal roles emerge over trajectory fragments rather than individual tokens, PR-FAPO aggregates $(\pi_\theta, \mathcal{H})$ statistics of Proposition 3.2 over prefix–suffix segments $\tau \subset y_i$ into a segment-level perception–reasoning fusion gap. This gap converts GRPO trajectory-bandit advantage into a token-wise linear coefficient on $\zeta w_{i,t}$ (Eq. 14) via TD gap under GAE yields role- and position-aware credit allocation. Whereas external visual verifiers, frame-level supervision, and handcrafted perceptual rewards impose perception externally and leave CoT structure uncontrolled, PR-FAPO shapes CoT itself, encouraging perception rather than enforcing it.

Perceptual anchors. Because $z_{i,t}$ is policy-dependent, non-stationary (Eq. 4), and autoregressively conditioned on a growing prefix $y_{i,<t}$, latent variables at later positions are unreliable anchors for perceptual evaluation. Recovering $z_{i,t}$ as an explicit latent perceptual state would demand costly supervision, such as perceptual annotations or latent objectives during SFT. We deliberately avoid this and instead treat $z_{i,t}$ as a gradient-level signal reflecting policy-induced hypotheses about visual evidence. To obtain hypotheses that remain closely grounded in the observed visual content, we treat the early `<perceive> </perceive>` process, which contains query-relevant articulated visual information, as a memo and denote it by $z_m \sim \pi_\theta(\cdot \mid q, C)$. Then subsequent `<think> </think>` process, system prompt is provided in Table 3 of Appendix L. In this early region, prefix entanglement is limited, allowing explicit object-level observations and hypotheses about visual evidence to be more readily disentangled from subsequent reasoning processes. Following autoregressive structure of Eq. 4, we factorize the hypothesis-conditioned likelihood

$$\prod_{j=1}^{|y_i|} \pi_\theta(y_{i,j} \mid z_m, s_{i,j}) = \underbrace{\prod_{k=1}^{t-1} \pi_\theta(y_{i,k} \mid z_m, s_{i,k})}_{\mathrm{RA}_{i,t}(m)} \cdot \underbrace{\prod_{j=t}^{|y_i|} \pi_\theta(y_{i,j} \mid z_m, s_{i,j})}_{\mathrm{HA}_{i,t}(m)}. \tag{16}$$

These prefix $(\mathrm{RA}_{i,t}(m))$–suffix $(\mathrm{HA}_{i,t}(m))$ likelihoods aggregate evidence over multiple positions and therefore measure how consistently a perceptual hypothesis z_m explains both preceding context and subsequent reasoning continuation. Hence, even in sharp-policy regime, where token-level scalars lose resolution according to Proposition 3.2, they retain discriminative information about the role of z_m as a perceptual anchor. We instantiate $(\pi_\theta, \mathcal{H})$ pair of Proposition 3.2 at the segment scale through a probability-axis statistic $\rho_{i,t}$ and an entropy-axis statistic $\mathcal{U}_{i,t}$, both aggregated over the

hypotheses $\{z_m\}_{m=1}^{|\mathcal{O}|}$ (See Eq. 88 into Appendix F.1 for further motivation behind). For $y \in \mathcal{V}$,

$$\rho_{i,t}(y) \coloneqq \sum_{m=1}^{|\mathcal{O}|} \Bigg[\underbrace{\tfrac{\pi_\theta(y|z_m,s_{i,t})\,\mathrm{HA}_{i,t+1}(m)}{|y_i|-t+1}}_{\text{future support}} + \underbrace{\pi_\theta(y \mid z_m, s_{i,t})}_{\text{local mass}} - \underbrace{\sigma(-\hat{A}_i)\,\tfrac{|y_i|-t+1}{|y_i|}\,\mathrm{RA}_{i,t}(m)}_{\text{retrospective veto}} \Bigg]. \tag{17}$$

The three terms play complementary roles: *future support* promotes candidates consistent with the forward continuation, *local mass* preserves the current commitment, and the *retrospective veto* is activated ($\sigma(-\hat{A}_i) \approx 1$) only on failed trajectories ($\hat{A}_i < 0$), suppressing candidates whose support stems from a prefix already misaligned with the outcome. Applying the same template along the entropy axis with $\mathrm{RA}^{\mathcal{H}}_{i,t}(m) = \prod_{k=1}^{t-1} \mathcal{H}(\pi_\theta(\cdot \mid z_m, s_{i,k}))$ yields the corresponding decomposition

$$\mathcal{U}_{i,t} \coloneqq \sum_{m=1}^{|\mathcal{O}|} \left[\tfrac{\sigma(\hat{A}_i)\,\mathrm{RA}^{\mathcal{H}}_{i,t}(m)}{|y_i|-t+1} + \mathcal{H}(\pi_\theta(\cdot \mid z_m, s_{i,t})) - \tfrac{|y_i|-t+1}{|y_i|}\,\mathcal{H}(\pi_\theta(\cdot \mid z_m, s_{i,t}))\,\mathrm{HA}^{\mathcal{H}}_{i,t+1}(m) \right]. \tag{18}$$

The pair $(\rho_{i,t}, \mathcal{U}_{i,t})$ thus realizes the $(\pi_\theta, \mathcal{H})$ pair of Proposition 3.2 under trajectory-level consistency, restoring the discriminative resolution that token-level statistics lose in the sharp-policy limit.

Segment level and adaptive discount. For a window of length τ starting at index s, let $\mathcal{I}_{\tau,s} \coloneqq \{s, \ldots, \min(s+\tau-1, |y_i|)\}$ and average the token-wise pair over this window:

$$\bar{\rho}_{i,\tau,s}(y) = \tfrac{1}{|\mathcal{I}_{\tau,s}|} \sum_{t \in \mathcal{I}_{\tau,s}} \rho_{i,t}(y), \qquad \bar{\mathcal{U}}_{i,\tau,s} = \tfrac{1}{|\mathcal{I}_{\tau,s}|} \sum_{t \in \mathcal{I}_{\tau,s}} \mathcal{U}_{i,t}. \tag{19}$$

Substituting this segment-level pair into the upper deviation Δ_{ent^H} of Eq. 11 yields a segment-level counterpart of that quantity. We retain the upper branch because its admissible region $\Delta_{\mathrm{ent}^H} \geq 0$ coincides with regime in which Proposition 3.2 predicts that reasoning likelihood regime dominated by perceptual grounding. Projecting onto this region defines the *segment-level perception–reasoning fusion gap*, where lower fusion gap indicates tighter upper bound.

$$V^p_{i,\tau,s} = 1 - \exp\left(- \left[\sum_{y \in \mathcal{V}} \bar{\rho}_{i,\tau,s}(y)^2 - e^{-\bar{\mathcal{U}}_{i,\tau,s}} \right] \right), \tag{20}$$

so that normalized $V^p_{i,\tau,s} \in (-\infty, 1]$ marks precisely those segments that require credit attenuation, aiming for advantage retention in low gap state and on the contrary during horizon ($V^p \downarrow \Rightarrow\ \gamma \uparrow$). We realize this attenuation through

$$\gamma_{i,\tau,s} = 2^{-V^p_{i,\tau,s}/|y_i|} \in (0, 1], \tag{21}$$

whose base-2 exponent matches the bit-unit of $\mathcal{H}(\pi_\theta)$ and whose $1/|y_i|$ normalization enforces length-scale invariance. A more detailed motivation is provided in Appendix G. We then adopt a TD-style formulation Sutton & Barto (2018) with GAE Schulman et al. (2016). Since RLVR provides only an outcome reward, we propagate credit through a bootstrapped residual defined directly on the segment perception value V^p:

$$\delta^P_{i,\tau,s} = \gamma_{i,\tau,s}\, V^p_{i,\tau,s+1} - V^p_{i,\tau,s}, \qquad \mathrm{GAE}^P_{i,\tau,s} = \delta^P_{i,\tau,s} + \gamma_{i,\tau,s} \Psi\, \mathrm{GAE}^P_{i,\tau,s+1}, \quad \mathrm{GAE}^P_{i,\tau,|y_i|} = 0. \tag{22}$$

Here V^p localizes *where* attenuation is needed, while $\gamma_{i,\tau,s}$ controls *how strongly* it propagates along the trajectory.

Multi-scale window weighting. A single fixed τ either over- or under-attenuates long-range perceptual planning and local refinement. We therefore adopt the geometric family

$$\mathcal{W}_i = \left\{ \tau = 4^k \;\middle|\; 1 \leq k \leq k_{\max} \right\}, \qquad k_{\max} = \left\lfloor \log_4 |y_i| \right\rfloor, \tag{23}$$

i.e. the minimal cover that adds two bits of scale per step (matching the bit-unit of γ in Eq. 21) with logarithmic cardinality $O(\log_4 |y_i|)$. The normalized inter-scale weight

$$w(\tau) = \frac{\log_4 \tau + 1}{\sum_{\tau' \in \mathcal{W}_i} (\log_4 \tau' + 1)}, \qquad \sum_{\tau \in \mathcal{W}_i} w(\tau) = 1, \tag{24}$$

is linear in $\log_4 \tau$, coarser scales receive proportionally larger mass. For $\kappa \in \mathcal{I}_{\tau,s}$, local exponent

$$(\gamma_{i,\tau,s})^{e_{\tau,s}(\kappa)} = (\gamma_{i,\tau,s})^{\alpha(\tau)(\kappa-s)} + (\gamma_{i,\tau,s})^{(1-\alpha(\tau))(\min(s+\tau-1,|y_i|)-\kappa)}, \quad \alpha(\tau) = \tfrac{\tau-\tau_{\min}}{\tau_{\max}-\tau_{\min}}, \tag{25}$$

Table 1: **Evaluation Results on Video QA Benchmarks.** We compare Video-PR-FAPO with both perception-oriented and reasoning-oriented benchmarks.

Model	Perception Centric Benchmark			Reasoning Centric Benchmark			
	VideoMME	MVBench	LVBench	MMVU	Video-MMMU	MVP	SVBench
Video-RFT Wang et al. (2025d)	59.8	62.1	–	68.5	51.1	–	–
Video-R1 Feng et al. (2025)	61.8	65.5	–	65.0	51.4	33.0	–
Video-RTS Wang et al. (2025i)	63.0	–	56.6	66.4	52.7	–	–
LongVILA-R1 Chen et al. (2025b)	65.1	67.6	58.0	–	–	–	14.3
LOVE-R1 Fu et al. (2025b)	66.2	66.6	60.1	–	–	–	–
VideoChat-R1.5 Yan et al. (2025)	65.2	70.6	61.4	–	49.6	38.6	–
Qwen3-VL-8B Bai et al. (2025)	72.5	69.4	**67.6**	69.9	61.0	40.5	29.1
VideoAuto-R1 Liu et al. (2026b)	71.7 (-0.8)	**72.0** (+2.6)	67.4 (-0.2)	71.1 (+1.2)	65.0 (+4.0)	43.0 (+2.5)	32.7 (+3.6)
Video-PR-FAPO (Ours)	**72.5** (± 0.0)	71.4 (+2.0)	67.1 (-0.5)	**72.2** (+2.2)	**67.6** (+6.6)	**43.7** (+3.2)	**34.2** (+5.1)

is a convex combination of forward and backward distances to segment endpoints. It places mass near segment tail for short windows ($\alpha \to 0$, local refinement) and near the head for long ones ($\alpha \to 1$, early perceptual planning). Since $\gamma_{i,\tau,s} \in (0, 1]$, smaller exponents emphasize local contributions:

$$A^{\text{local}}_{i,\tau,s}(\kappa) = (\gamma_{i,\tau,s})^{e_{\tau,s}(\kappa)}\,\text{GAE}^{P}_{i,\tau,s}, \qquad M_{i,\tau} = \text{Concat}\Big(A^{\text{local}}_{i,\tau,s_1}, \ldots, A^{\text{local}}_{i,\tau,s_{K_{i,\tau}}}\Big), \tag{26}$$

where each per-scale signal $M_{i,\tau}$ is obtained by concatenating non-overlapping segments. Then *fusion-aware weight* is proposed by compressing the multi-scale signal $\{M_{i,\tau}\}_{\tau\in\mathcal{W}_i}$ into a token-wise multiplier

$$f_{i,t} = \begin{cases} g_{i,t}, & \text{if } \widehat{A}_i > 0, \\ 1.5 - g_{i,t}, & \text{otherwise,} \end{cases} \qquad g_{i,t} = \text{sg}\left[\exp\left(\sum_{\tau\in\mathcal{W}_i} w(\tau)\, M_{i,\tau}(t) - 1\right)\right]_{+}. \tag{27}$$

Combined with $\sum_\tau w(\tau) = 1$, the offset -1 ensures maximizing $f_{i,t} = 1$ when $M_{i,\tau}(t) \equiv 1$, making $f_{i,t}$ a symmetric modulation factor around the GRPO baseline. The inter-scale mixture $\sum_\tau w(\tau) M_{i,\tau}(t)$ transfers segment-level credit to tokens and, through Eq. 14, modulates only the shared scalar channel $\zeta w_{i,t}$ while preserving the token-level gradient direction through the stop-gradient operator $\text{sg}[\cdot]$. Building on the token-level GRPO objective, we define *Perception–Reasoning Fusion-Aware Policy Optimization* (PR-FAPO) as

$$\mathcal{J}_{\text{PR-FAPO}}(\theta) = \mathbb{E}_{\{q,C,a^*\}\sim D,\ \{y_i\}_{i=1}^{G}\sim\pi_{\text{old}}}\left[\frac{1}{\sum_{i=1}^{G}|y_i|}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\right.$$
$$\left.\min\Big(r_{i,t}(\theta)\, f_{i,t}\,\widehat{A}_i,\ \text{clip}\big(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\big) f_{i,t}\,\widehat{A}_i\Big) - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right]. \tag{28}$$

Placing $f_{i,t}$ outside the clipping operator decouples the reference-policy importance ratio from linear step-wise credit allocation. The clipped ratio constrains the GRPO update relative to the reference policy, whereas $f_{i,t}$ restores a token-wise linear utility through the scalar channel $\zeta w_{i,t}$, enabling credit reallocation across perception, reasoning, and their interaction through the advantage signal without complicating the importance-ratio clipping mechanism.

5 Experiments

Implementation. We fine-tune Qwen VL family Bai et al. (2025) on the VideoAuto-R1 dataset Liu et al. (2026b) for 1 epoch. Following VideoAuto-R1, we use a global batch size of 256, a KL coefficient of $\beta = 0.01$. Both training and inference use 64 up to 2048 frames; further details are provided in Appendix N. We evaluate general video QA on Video-MME (w/o subtitles) Fu et al. (2025a), MVBench Li et al. (2024), and LVBench Wu et al. (2024); video reasoning on Video-MMMU Hu et al. (2025), MVP Krojer et al. (2025), SVBench Deng et al. (2025a) and MMVU Zhao et al. (2025a); and temporal grounding on Charades-STA Gao et al. (2017), ActivityNet Caba Heilbron et al. (2015), and NExT-GQA Xiao et al. (2024). We report accuracy for QA, recall and mean IoU for temporal localization, and corresponding grounded-QA metrics for NExT-GQA. Additionally, further evaluate our model on image and text benchmark as Appendix Q

Main Results. In *Video QA Benchmark*, we evaluate all models on six video QA benchmarks to comprehensively assess the effectiveness of PR-FAPO (Table 1). PR-FAPO consistently achieves substantial gains over VideoAuto-R1 on reasoning-centric benchmarks, where effective utilization of visual information within CoT reasoning is essential. Importantly, these gains are obtained

Table 2: **Evaluation Results on Temporal Grounding Benchmarks.**

Model	Charades-STA				ActivityNet				NeXT-GQA	
	0.3	0.5	0.7	mIoU	0.3	0.5	0.7	mIoU	Acc	mIoU
TimeChat Ren et al. (2024)	51.0	27.5	11.4	31.2	44.0	27.8	14.3	30.4	28.8	17.4
Temporal-RLT Li et al. (2025a)	79.6	67.9	44.1	57.0	56.9	38.4	20.2	39.0	78.7	37.3
Time-R1 Wang et al. (2025g)	82.8	72.2	50.1	58.8	73.3	**55.6**	**34.0**	**52.1**	–	–
VITAL Zhang et al. (2025)	83.1	72.0	46.7	59.9	70.9	50.8	31.6	49.8	78.7	43.0
Qwen3-VL-8B Bai et al. (2025)	68.5	49.3	26.2	47.2	63.8	51.1	35.0	48.0	-	-
VideoAuto-R1 Liu et al. (2026b)	85.1	74.9	53.7	63.7	74.1	54.3	32.4	51.9	81.1	44.2
Video-PR-FAPO (Ours)	84.6 (-0.5)	**75.0** (+0.1)	**55.4** (+1.7)	**64.0** (+0.3)	**74.1** (+0.0)	53.8 (-0.5)	32.9 (+0.5)	51.9 (+0.0)	**81.5** (+0.4)	**44.2** (+0.0)

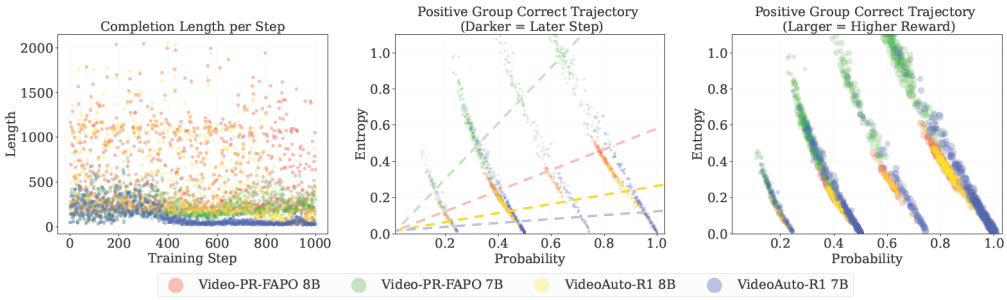


Figure 2: Under using a same Qwen2.5-VL and Qwen3-VL backbone, comparison of probability–entropy trajectories between VideoAuto-R1 and TD decay factor of $\Psi = 0.96$ Video-PR-FAPO.

while maintaining competitive performance on perception-oriented benchmarks that emphasize fine-grained motion and spatial understanding. Overall, results shown that PR-FAPO improves reasoning capabilities without introducing a noticeable degradation in perception performance, demonstrating the effectiveness of the proposed fusion-aware CoT control across diverse video QA benchmarks.

Results on temporal grounding benchmarks are reported in Table 2. PR-FAPO outperforms or comparable with baselines in average mIoU and achieves clear gains over VideoAuto-R1 at a threshold of 0.7. These results suggest that PR-FAPO improves temporal localization by better aligning temporal perception with task objectives. Further result on other backbone in Appendix R.

Qualitative analysis of trajectory evolution. To further characterize the training dynamics, we analyze the evolution of reasoning-policy behavior. As shown in the center panel of Figure 2, PR-FAPO exhibits a selective and controlled policy drift toward structured probability–entropy regions, whereas VideoAuto-R1 trained with GRPO drifts predominantly toward low-entropy regions, suggesting reduced exploration of temporal evidence accumulation trajectories. This controlled drift reflects progressive compression of the reasoning process rather than premature convergence. Consistent with rollout observations, CoTs shorten naturally during training, indicating query-adaptive reasoning rather than indiscriminate shortening. In contrast, the baselines rapidly collapse to excessively short CoTs and remain trapped in low-entropy regions after 450 training steps, suggesting premature restriction of reasoning trajectories and reduced exploration within the reasoning space (left panel). As shown in the right panel, PR-FAPO maintains increasing policy entropy while shifting correct trajectories toward higher probability–entropy CoT regions. This trend mirrors Video-R1, despite not additionally requiring temporally shuffled anchor rollouts for advantage shaping, suggesting continued exploration of diverse temporal evidence accumulation strategies while filtering redundant reasoning. As a result, exploration remains structured without collapsing into low-entropy regimes. Taken together, these observations suggest that PR-FAPO preserves exploration over diverse temporal evidence accumulation strategies while progressively eliminating redundant reasoning, enabling stable policy evolution for video reasoning. Further analyses of the effects of the discount factor λ, window size τ, and additional ablations are provided in Appendix R.

6 Limitations and Conclusion

Our work has three limitations. First, it introduces an additional hyperparameter, TD decay coefficient Ψ. Second, our temporal reasoning-centric design may be less effective for tasks requiring subtle motion understanding, spatial reasoning, potentially leading to repeated re-inspection (Appendix S.2). Although this behavior may introduce additional reasoning overhead, it suggests that model attempts to repeatedly align visual evidence with reconsideration process, potentially alleviating the modality fusion gap. Third, the extent to which our approach to SFT-initialized models remains underexplored, particularly when latent tokens acquired into CoT during SFT influences policy behavior. Conclusively, in this work, we show that video reasoning benefits from optimization signals aligned with fusion aware reasoning policy. To this end, we incorporate a reasoning-constrained process into GRPO advantage estimation. We hope our work provide a useful perspective for future video reasoning work and optimization design.

REFERENCES

Saurav Agarwal, Zhen Zhang, Li Yuan, Jialiang Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025. URL https://arxiv.org/abs/2505.15134.

Anthropic. System card: Claude opus 4 and claude sonnet 4. Technical report, Anthropic, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. https://arxiv.org/abs/2511.21631, 2025. arXiv preprint arXiv:2511.21631.

Michael Beukman, Khimya Khetarpal, Zeyu Zheng, Will Dabney, Jakob Foerster, Michael Dennis, and Clare Lyle. Preventing learning stagnation in ppo by scaling to 1 million parallel environments. *arXiv preprint arXiv:2603.06009*, 2026.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. URL https://arxiv.org/abs/2411.18211.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do not think that much for 2+3=? on the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, Vancouver, BC, Canada, July 2025a. URL https://openreview.net/forum?id=MSbU3L7V00. OpenReview.net.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025b. URL https://arxiv.org/abs/2507.07966.

Guangran Cheng, Chengqi Lyu, Songyang Gao, Wenwei Zhang, and Kai Chen. Group entropy-controlled policy optimization, 2026. URL https://arxiv.org/abs/2607.16850.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*, 2024. URL https://arxiv.org/abs/2410.07752.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. 2025. URL https://arxiv.org/abs/2505.22617.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*, 2025a.

Jiawen Deng, Jie Chen, Zhiyuan Chen, Wayne Xin Zhao, and Ji-Rong Wen. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning. *arXiv preprint arXiv:2508.02260*, 2025b. URL https://arxiv.org/abs/2508.02260.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. URL https://arxiv.org/abs/2503.21776.

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 multimodal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025a.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*, 2025b.

Xiaoliang Fu, Jiaye Lin, Yangyi Fang, Chaowen Hu, Cong Qin, Zekai Shao, Binbin Zheng, Lu Pan, and Ke Zeng. Taming divergence in soft clipping via bilateral decoupled decay of probability gradient weight. *arXiv preprint arXiv:2603.14389*, 2026. URL https://arxiv.org/abs/2603.14389.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5267–5275, 2017.

Google DeepMind. Gemini 2.5 flash-lite. 2025. URL https://deepmind.google/. Model release, accessed 24 Apr 2026.

Roger B. Grosse, Zoubin Ghahramani, and Ryan P. Adams. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*, 2015. URL https://arxiv.org/abs/1511.02543.

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. URL https://arxiv.org/abs/2501.12948.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in rlvr: An entropy change perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.acl-long.1436/.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and Radford M. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995. doi: 10.1126/science.7761831.

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

Wenbo Hu, Xin Chen, Yan Gao-Tian, Yihe Deng, Nanyun Peng, and Kai-Wei Chang. Openvl-thinkerv2: A generalist multimodal reasoning model for multi-domain visual tasks, 2026. URL https://arxiv.org/abs/2604.08539.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2510.03222*, 2025a. URL https://arxiv.org/abs/2510.03222.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025b. doi: 10.48550/arXiv.2510.09285. URL https://arxiv.org/abs/2510.09285.

Sheng Jia, Xiao Wang, Shiva Prasad Kasiviswanathan, and Rein Houthooft. Group adaptive clipping policy optimization, 2026. URL https://arxiv.org/abs/2609.00444.

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*, 2025.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. Revisiting entropy in reinforcement learning for large reasoning models. *arXiv preprint arXiv:2511.05993*, 2025. URL https://arxiv.org/abs/2511.05993.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Haechan Kim, Soohyun Ryu, Gyouk Chu, Doohyuk Jang, and Eunho Yang. Discounted beta–bernoulli reward estimation for sample-efficient reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2603.18444*, 2026.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025. URL https://arxiv.org/abs/2506.09987.

Truong-Loc V. Le, Minho Jeon, Khang Vu, Viet Lai, and Eunho Yang. No prompt left behind: Exploiting zero-variance prompts in llm reinforcement learning via entropy-guided advantage shaping. *arXiv preprint arXiv:2509.21880*, 2025. URL https://arxiv.org/abs/2509.21880.

Bolian Li, Yifan Wang, Yi Ding, Anamika Lochab, Ananth Grama, and Ruqi Zhang. Addressing performance saturation for llm rl via precise entropy curve control. *arXiv preprint arXiv:2604.26326*, 2026a. URL https://arxiv.org/abs/2604.26326.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025a. URL https://arxiv.org/abs/2506.01908.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024.

Pengyi Li, Elizaveta Goncharova, Andrey Kuznetsov, and Ivan Oseledets. Back to basics: Revisiting exploration in reinforcement learning for llm reasoning via generative probabilities. *arXiv preprint arXiv:2602.05281*, 2026b.

Xianzhi Li, Ethan Callanan, Xiaodan Zhu, Mathieu Sibue, Antony Papadimitriou, Mahmoud Mahfouz, Zhiqiang Ma, and Xiaomo Liu. Entropy-aware branching for improved mathematical reasoning. *arXiv preprint arXiv:2503.21961*, 2025b.

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, 2025c.

Yizhuo Li, Jianhao Yan, Yun Luo, Zhi Wang, Futing Wang, Rong-Xi Tan, Kanghui Tian, Ganqu Cui, Ning Ding, Peilin Zhao, Yafu Li, and Yu Cheng. Rethinking critic learning in ppo: Understanding and mitigating value flattening, 2026c. URL https://arxiv.org/abs/2609.18708.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. STI-Bench: Are MLLMs ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025d. URL https://arxiv.org/abs/2503.23765.

Jiacai Liu. How does rl policy entropy converge during iteration? 2025. URL https://zhuanlan.zhihu.com/p/28476703733. Zhihu Blog.

Shiqi Liu, Zeyu He, Guojian Zhan, Letian Tao, Zhilong Zheng, Jiang Wu, Yinuo Wang, Yang Guan, Kehua Sheng, Bo Zhang, Keqiang Li, Jingliang Duan, and Shengbo Eben Li. Stapo: Stabilizing reinforcement learning for llms by silencing rare spurious tokens. *arXiv preprint arXiv:2602.15620*, 2026a. URL https://arxiv.org/abs/2602.15620.

Shuming Liu, Mingchen Zhuge, and Zhao. Videoauto-r1: Video auto reasoning via thinking once, answering twice. *arXiv preprint arXiv:2601.05175*, 2026b. URL https://arxiv.org/abs/2601.05175.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Haipeng Luo, Qingfeng Sun, Songli Wu, Can Xu, Wenfeng Deng, Han Hu, and Yansong Tang. Stare: Surprisal-guided token-level advantage reweighting for policy entropy stability. *arXiv preprint arXiv:2606.19236*, 2026. URL https://arxiv.org/abs/2606.19236.

Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. Fipo: Eliciting deep reasoning with future-kl influenced policy optimization. *arXiv preprint arXiv:2603.19835*, 2026. doi: 10.48550/arXiv.2603.19835.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 23464–23487, 2024.

Youssef Mroueh. Grpo with binary rewards is an adaptive weighted contrastive loss, 2025. Technical note.

Wenhua Nie, Jianan Wu, Junlin Liu, Ziwei Li, Zheng Lin Zhang, Zijian Zhang, Yilong Fan, Haoran Zheng, and Jyh-Shing Roger Jang. Gradient starvation in binary-reward grpo: Why group-mean centering fails and why the simplest fix works. *arXiv preprint arXiv:2605.07689*, 2026.

OpenAI. Gpt-5, 2025a. Referenced via public benchmark reporting (VideoMMMU, VideoMME).

OpenAI. Gpt-5-nano. August 2025b. URL https://openai.com/. Model release, accessed 24 Apr 2026.

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes rl scalable. In *Advances in Neural Information Processing Systems*, 2025.

Aleksei Petrenko, Ben Lipkin, Kevin Chen, Erik Wijmans, Marco Francis Cusumano-Towner, Raja Giryes, and Philipp Kraehenbuehl. Entropy-preserving reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multi-modal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

QwenTeam. Qwen3.5: Accelerating productivity with native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Qwen Blog, accessed 2026-09-04.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2024. URL https://arxiv.org/abs/2312.02051.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Aman Sharma and Paras Chopra. Think just enough: Sequence-level entropy as a confidence signal for llm reasoning. *arXiv preprint arXiv:2510.08146*, 2025. doi: 10.48550/arXiv.2510.08146. URL https://arxiv.org/abs/2510.08146.

Qile Su, Jing Tang, Rui Chen, Lei Sun, and Xiangxiang Chu. Video-coe: Reinforcing video event prediction via chain of events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32818–32828, 2026.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Alibaba Group Qwen Team. Qwen2.5-vl-7b-instruct. https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct, 2024. Accessed: 2026-01-30.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation. *arXiv preprint arXiv:2606.05718*, 2026. URL https://arxiv.org/abs/2606.05718.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025. URL https://arxiv.org/abs/2506.13654.

Jintao Tong, Shilin Yan, Hongwei Xue, Xiaojun Tang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. Swimbird: Eliciting switchable reasoning mode in hybrid autoregressive mllms. *arXiv preprint arXiv:2602.06040*, 2026.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. *arXiv preprint arXiv:2505.22457*, 2025a.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025b. URL https://arxiv.org/abs/2504.08837.

Jiayi Wang, Jiacheng Liu, Yifan Fu, Yuxuan Li, Xin Wang, Yuxiang Lin, Yao Yue, Liang Zhang, Yifei Wang, and Kai Wang. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*, 2025c. URL https://arxiv.org/abs/2509.09265.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, Meng Wang, Pengfei Wan, and Xiaodan Liang. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5988–5998, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025d.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui 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. *arXiv preprint arXiv:2506.01939*, 2025e. URL https://arxiv.org/abs/2506.01939.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. SOTA with less: MCTS-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025f. URL https://arxiv.org/abs/2504.07934.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025g. URL https://arxiv.org/abs/2503.13377.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025h. doi: 10.48550/arXiv.2507.06448. URL https://arxiv.org/abs/2507.06448.

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. *arXiv preprint arXiv:2507.06485*, 2025i. URL https://arxiv.org/abs/2507.06485.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024.

Keming Wu, Baoyi Wang, Kaichen Zhang, Xiang An, Zuhao Yang, Sudong Wang, Haowei Zhu, Tingxuan Huang, Hongcheng Gao, and Bin Wang. Streamopd: A post-training recipe with spatio-temporal cue gating for streaming video understanding. *arXiv preprint arXiv:2608.16320*, 2026. doi: 10.48550/arXiv.2608.16320. URL https://arxiv.org/abs/2608.16320.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025. doi: 10.48550/arXiv.2510.18927. URL https://arxiv.org/abs/2510.18927.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, 2024.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025. URL https://arxiv.org/abs/2509.21100.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025a. URL https://arxiv.org/abs/2503.10615.

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in rl for llms. *arXiv preprint arXiv:2505.12929*, 2025b. URL https://arxiv.org/abs/2505.12929.

[First Author] Yao et al. Imagine before you predict: Interleaved latent visual reasoning for video event prediction. *arXiv preprint arXiv:2606.05769*, 2026a.

Jiarui Yao, Ruida Wang, Hao Bai, and Tong Zhang. Future-kl regularized grpo: Process-level credit assignment from f-divergence regularization. *arXiv preprint arXiv:2601.10201*, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Yurun Yuan and Tengyang Xie. Breaking the capability ceiling of LLM post-training by reintroducing markov states. *arXiv preprint arXiv:2603.19987*, 2026. URL https://arxiv.org/abs/2603.19987.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Krystian Zawistowski. Unused information in token probability distribution of generative llm: improving llm reading comprehension through calculation of expected values. *arXiv preprint arXiv:2406.10267*, 2024. doi: 10.48550/arXiv.2406.10267. URL https://arxiv.org/abs/2406.10267.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024. URL https://arxiv.org/abs/2410.19702.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025. URL https://arxiv.org/abs/2508.04416.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10419–10429, 2026.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024a. doi: 10.48550/arXiv.2407.12772. URL https://arxiv.org/abs/2407.12772.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*, pp. 169–186. Springer, 2024b.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025a.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025b. URL https://arxiv.org/abs/2507.20673.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025a. URL https://arxiv.org/abs/2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025b. URL https://arxiv.org/abs/2505.14362.

Yiwu Zhong, Zi-Yuan Hu, Yin Li, and Liwei Wang. Rethinking chain-of-thought reasoning for videos. *arXiv preprint arXiv:2512.09616*, 2025. URL https://arxiv.org/abs/2512.09616.

Xiangxin Zhou, Jiarui Yao, Penghui Qi, Bowen Ping, Jiaqi Tang, Haonan Wang, and Tianyu Pang. Predictive divergence masks for llm rl, 2026. URL https://arxiv.org/abs/2607.10848.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. MMR-V: What's left unsaid? a benchmark for multimodal deep reasoning in videos. *arXiv preprint arXiv:2506.04141*, 2025. URL https://arxiv.org/abs/2506.04141.

Xiaoyu Zhu, Xinke Deng, Suresh Taddewadikar, Arnab Kumar Mondal, Zhongyu Jiang, Ian Fasel, and Joerg Liebelt. Beyond visual cot: Internalized visual thinking for proactive video reasoning. *arXiv preprint arXiv:2608.15869*, 2026a. URL https://arxiv.org/abs/2608.15869.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. Timelens2: Generalist video temporal grounding with multimodal llms. *arXiv preprint arXiv:2607.17423*, 2026b. doi: 10.48550/arXiv.2607.17423. URL https://arxiv.org/abs/2607.17423.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

Appendix

A Empirical **19**

A.1 Empirical Training Dynamics Across Text, Image, and Video Modalities 19

A.2 Empirical Training Dynamics in Qwen2.5 VL and Qwen3 VL 21

A.3 Applying High Entropy Token-Level Masking to Video Data 22

B Derivation **23**

B.1 Derivation of Policy Gradient Norm Bounds for Perception-based Reasoning 23

B.2 Derivation of Entropy Update Mechanism on Reasoning Policy 27

B.3 Optimization of Reasoning Direction Scaling 29

B.4 Relationship of Perceptual Learning and Reasoning Policy produced CoT Length 30

B.5 EUBO-Based Step-wise Evidence Decomposition 31

B.6 Properties of the GRPO Advantage at a Group Mean of 0.5 33

C The Motivating for indirect Perception Modeling via Reasoning Policy State Perspectives **34**

D Related Work **35**

E Prompt Template **37**

F Training Setting **37**

G Parameter Settings **37**

H Inferencing **38**

I Evaluation Configuration **39**

J Evaluation on Image and Text Benchmarks **39**

K Ablation Studies **39**

K.1 The result in Qwen2.5-VL 7B backbone and SFT-initialized model. 39

K.2 Reviewed (w CoT) vs. Initial (w/o CoT) Reward 40

K.3 The Study on Window τ 41

K.4 The Study on value-function TD Ψ 41

K.5 Training Dynamic 41

K.6 Stability 43

L Attribution **44**

L.1 Role-aware and Position-aware 44

L.2 Discounting strategy 44

M Qualitative Examples **45**

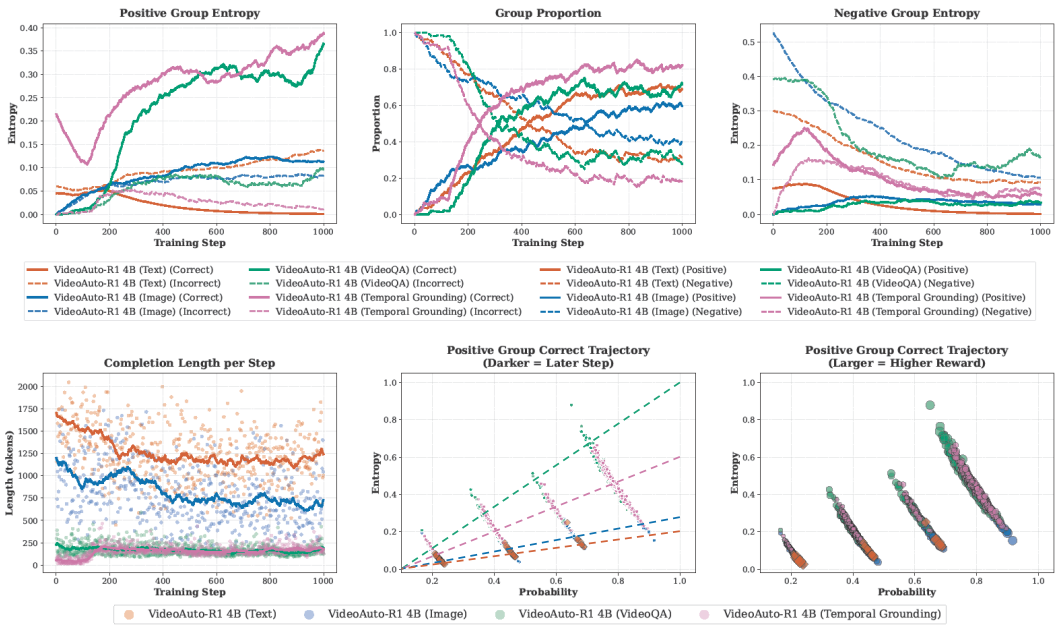


Figure 3: The dynamics of empirical training on diverse text, image, and video data types trained separately.

A Empirical Training Dynamics Across Text, Image, and Video Modalities

Analysis Setup. To investigate how different modalities and task types affect policy convergence, we analyze the training dynamics of models trained on text-only, image-only, video question answering (VideoQA), and temporal grounding datasets. Specifically, we examine the evolution of probability entropy and the proportion of positive and negative groups during training to identify modality-dependent convergence behaviors. For the text-only setting, we use 6.4K samples from DAPO-Math Yu et al. (2025). For the image-only setting, we use 27.5K samples from ViRL and ThinkLite-Hard Wang et al. (2025b;f). For the video setting, we use 49.4K samples consisting of both VideoQA datasets Feng et al. (2025); Cores et al. (2024); Li et al. (2025d); Zhu et al. (2025) and temporal grounding datasets Wang et al. (2025g); Gao et al. (2017); Caba Heilbron et al. (2015); Xiao et al. (2024). Each modality is trained independently in separate training runs; the datasets are never jointly mixed during training. To ensure a fair comparison, all experiments follow the same VideoAuto-R1 training framework. We intentionally omit the cold-start SFT stage to eliminate potential confounding effects introduced by supervised warm-up. Consequently, all models are trained directly using the original GRPO algorithm with a group rollout size of $G = 16$, adopting Qwen3-VL with following system prompt is provided in Table 3 of Appendix L.

Analysis of Group Entropy. As shown in the middle panel of the first row in Table 3, we measure the proportion of positive groups, where a positive group is defined as a sampled group whose mean reward exceeds 0.5. The results show that temporal grounding exhibits the fastest convergence within the first 1,000 training steps, followed by VideoQA and the text-only setting. This observation suggests that the VideoAuto-R1 framework, which employs an answer-then-CoT-based re-answer mechanism, is particularly well suited for video tasks. Compared with image-only training, video-based tasks appear to provide a stronger learning signal that facilitates rapid policy improvement. In contrast, the text-only setting exhibits a different convergence pattern despite being trained under the same optimization procedure. A possible explanation can be observed in the left panel of the first row. The negative-group analysis shown in the right panel exhibits similar trends across all modalities and does not reveal any notable distinctions. Therefore, the following discussion focuses on the positive-group dynamics. The policy distributions associated with correct trajectories in video-based tasks maintain substantially higher entropy than those in the text-only setting. This indicates that successful video reasoning is supported by a more diverse set of valid trajectories, whereas text reasoning tends to concentrate probability mass on a smaller subset of trajectories during training.

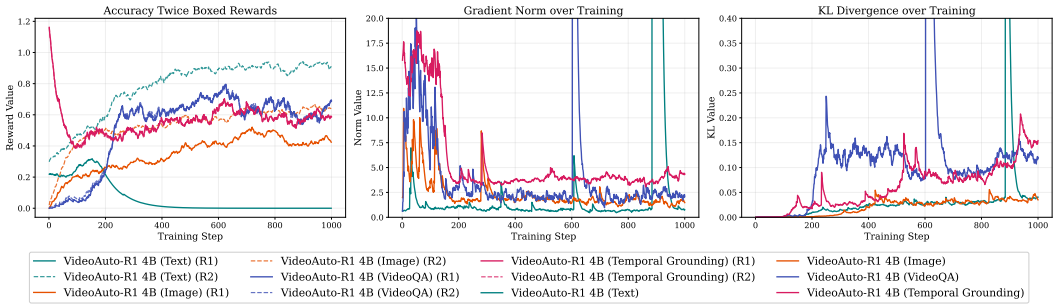


Figure 4: The stability of VideoAuto-R1 training on diverse text, image, and video data types trained separately.

Furthermore, in the video modality, incorrect trajectories that belong to positive groups exhibit lower entropy than correct trajectories. This behavior contrasts with the text-only setting, where incorrect trajectories typically maintain higher entropy than correct ones. These results suggest that the entropy characteristics of CoT reasoning differ fundamentally between video and text modalities. In particular, video-based reasoning appears to preserve a broader distribution over valid reasoning trajectories, while text-based reasoning converges toward a more concentrated policy. This finding suggests that video reasoning is not merely a direct extension of text-based reasoning, but may require modality-specific training strategies to effectively leverage the inherently higher entropy of video reasoning trajectories.

Analysis of Probability Entropy Space. As shown in the second row of Table 3, we analyze the probability–entropy characteristics of correct trajectories within positive groups. The middle panel shows that policies trained on VideoQA and temporal grounding consistently remain in a high-probability but maintain high-entropy regime, whereas text- and image-based policies converge toward lower-entropy regions with more concentrated probability mass. This difference is further reflected in the right panel: for VideoQA and temporal grounding, higher probability–entropy trajectories are consistently associated with higher mean group rewards, while such a correlation is largely absent in text and image modalities. The left panel reveals that VideoQA and temporal grounding generate substantially shorter CoT trajectories than text- and image-based tasks. Despite these shorter reasoning sequences, video-based policies maintain higher entropy with high probability, suggesting that successful video reasoning is supported by a broader set of valid reasoning trajectories. In contrast, text and image policies tend to concentrate probability mass on a smaller set of specialized reasoning chains. Taken together, these results suggest that policy improvement in video reasoning benefits from maintaining a high-probability, high-entropy trajectory band, which is consistently associated with higher group rewards and faster convergence. This finding provides empirical support for VideoAuto-R1, which combines GRPO with an answer-then-CoT-based re-answer strategy, for video reasoning tasks. Nevertheless, the results also raise an important question regarding the underlying optimization mechanism. Although video-based policies maintain higher entropy, it remains unclear whether the relatively short CoT trajectories provide sufficient exploration of the reasoning space. High entropy alone does not necessarily imply deeper reasoning; rather, it may reflect a broader distribution over shallow reasoning paths. Consequently, the observed convergence gains may not arise solely from improved CoT optimization. An alternative explanation is that the framework implicitly favors shorter reasoning trajectories, enabling reward-improving behavior through early CoT termination. Such behavior may indicate a form of reasoning shortcutting, where higher rewards are achieved through abbreviated reasoning processes rather than more effective reasoning. Therefore, we focus on identifying the mechanism that gives rise to the higher probability-entropy band observed in video-based GRPO and understanding why trajectories in this band are consistently associated with higher mean group rewards on video reasoning policy convergence.

Analysis of Stability. As shown in left panel of Figure 4, the text and image settings exhibit a clear trend throughout RL training: the performance of CoT-based re-answer gradually largely surpasses that of the initial answer without CoT reasoning. This observation suggests that reasoning policy is continuously improved during optimization. In contrast, the video setting does not exhibit such a divergence, as the CoT-based re-answer remains comparable to the initial answer across training. This result indicates that the benefits of reasoning optimization are less pronounced in the video

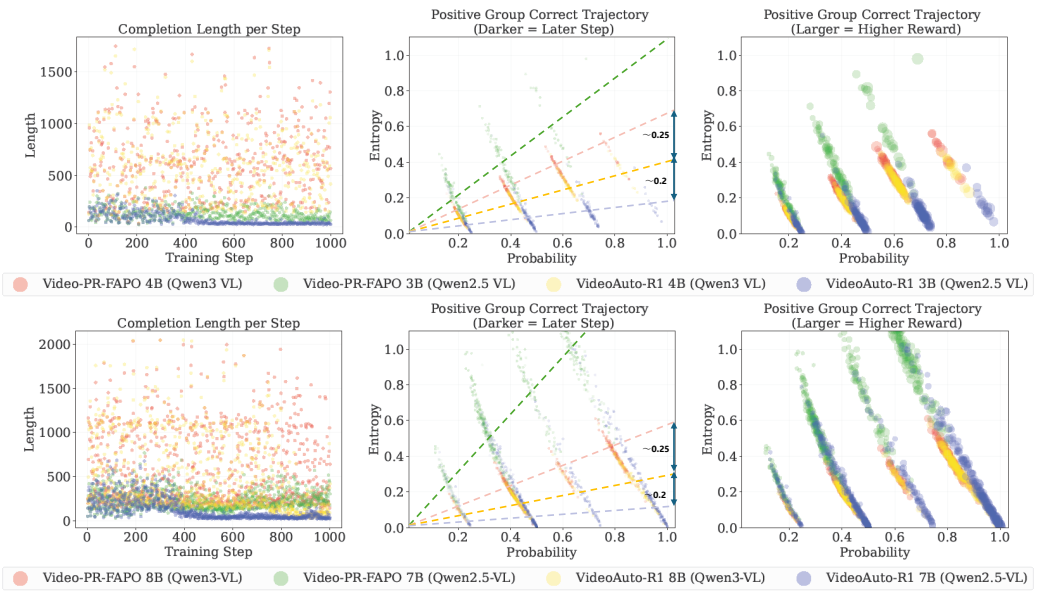


Figure 5: Empirical Training Dynamics in video under token-level entropy

modality. Further evidence is provided in the middle panel of Figure 4, where the video modality consistently exhibits larger gradient norms than the text and image settings. This observation suggests less stable optimization dynamics and greater difficulty in policy learning. The right panel of Figure 4 provides additional insights into the training behavior. In the VideoQA setting, the policy begins to frequently move outside the PPO trust region after approximately 250 training steps, while a similar phenomenon emerges later in temporal grounding. Over time, both tasks gradually deviate from the stable update regime. Taken together, these findings suggest that the perception-reasoning fusion gap is not fully resolved under the VideoAuto-R1 framework. As a result, the model may rely more heavily on perception update shortcuts instead of developing stronger RL-reasoning behaviors, thereby limiting effective utilization of video evidence in VideoQA and leading to under-exploration of temporal information in temporal grounding tasks.

B Empirical Training Dynamics in Qwen2.5 VL and Qwen3 VL

Analysis Setup. To examine whether Qwen2.5-VL and Qwen3-VL exhibit different probability-entropy ranges during training, we compare their training dynamics across two pairs of similarly sized models: Qwen2.5-VL 7B and Qwen3-VL 8B, as well as Qwen2.5-VL 3B and Qwen3-VL 4B. All experiments are conducted using the original VideoAuto-R1 training dataset (mixed text, image, video modality). Following the VideoAuto-R1 training setup, we not use the cold-start supervised fine-tuning stage to eliminate potential confounding effects arising from supervised warm-up. The models are therefore trained directly with original GRPO using a group rollout size of $G = 16$. Following system prompt is provided in Table 3 of Appendix L.

Analysis of CoT Reasoning Entropy. Following the analysis in the main paper, we visualize only the correct trajectories and group them into the positive set, i.e., trajectories whose rewards exceed 0.5, as shown in Fig. 5. The results reveal a consistent training pattern across all model scales. Despite differences in model size, both Qwen2.5-VL and Qwen3-VL exhibit a progressive shift toward lower-entropy regimes during GRPO training under re-answer VideoAuto-R1 framework. However, Qwen3-VL consistently converges to a higher entropy regime than Qwen2.5-VL. One possible explanation for this behavior lies in the architectural differences between the two model families. Qwen3-VL incorporates the DeepStack mechanism Meng et al. (2024), which extends conventional visual token extraction by aggregating multi-scale visual representations from intermediate Vision Transformer (ViT) layers. This design preserves richer visual information spanning both low-level and high-level semantic features. Consequently, the language decoder receives more diverse visual representations, potentially yielding a broader reasoning search space and higher-entropy generation trajectories. Such richer video representations may help maintain reasoning diversity throughout

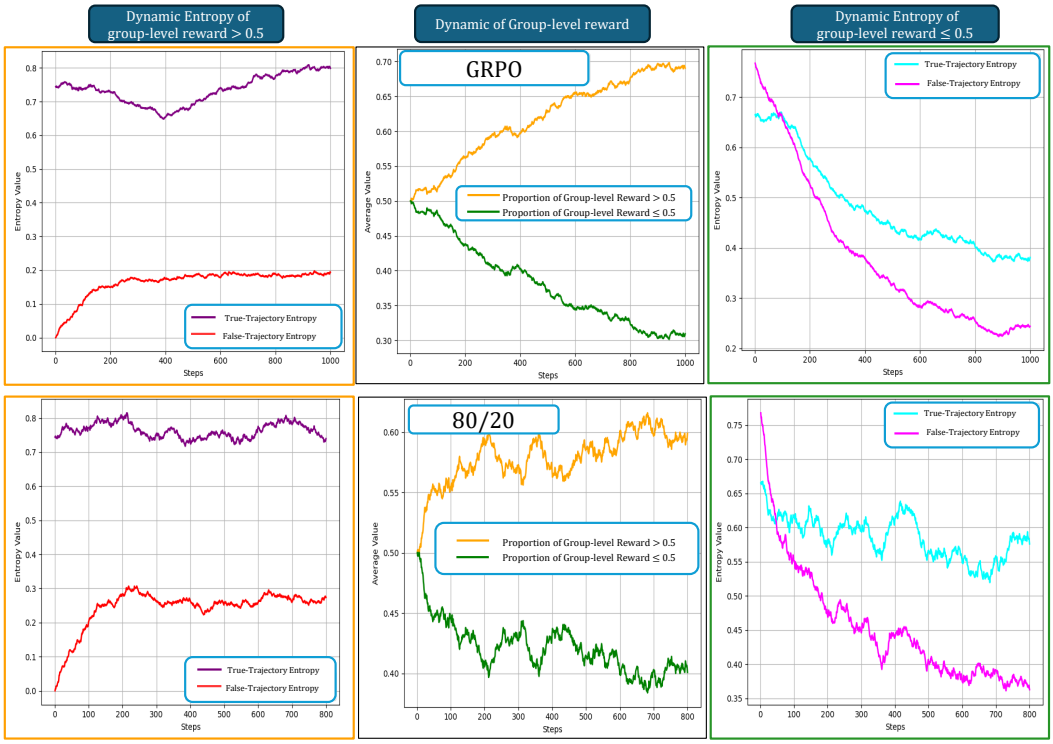


Figure 6: Empirical Training Dynamics in video under token-level entropy

post-training, leading to a more stable convergence process than that observed in Qwen2.5-VL. Nevertheless, GRPO optimization gradually favors trajectories with consistently higher rewards, effectively reducing the exploration space during training. As a result, both model families eventually converge toward lower-entropy reasoning regimes. Although Qwen3-VL starts from and maintains a richer exploration space due to its enhanced visual representations, the optimization dynamics of GRPO can drive the model toward a relatively narrower reasoning distribution at convergence. This observation suggests that GRPO may not fully exploit the additional visual information provided by DeepStack, warranting further investigation into entropy-preserving optimization strategies for video reasoning models. In contrast, the results of Video-PR-FAPO indicate that maintaining a higher-entropy reasoning regime throughout training enables the model to better preserve and utilize rich visual information.

C Additional Empirical Training Dynamics from Applying High Entropy Token-Level Masking to Video Data

Analysis Setup. To examine whether simply adding token-level entropy-controlling can encourage reliable video reasoning, we analyze their training dynamics by replicating Video-R1 Feng et al. (2025) with the original GRPO Shao et al. (2024) and the popular entropy-based Token 80/20 method Wang et al. (2025e), which is built upon GRPO and using same Video-R1 framework. Specifically, to avoid confounding effects introduced by supervised warm-up, we omit cold-start supervised fine-tuning and directly optimize Qwen2.5-VL-7B Team (2024) using a group rollout size of $G = 16$ and a video-only filtered subset of the Video-R1 Feng et al. (2025) training dataset, since the original dataset comprises text, image, and video configurations. During training, we periodically perform test-time rollouts, in which the model executes its current reasoning policy. These rollouts often leverage additional test-time compute by sampling multiple candidate trajectories and selecting those that maximize reward. Recent RLVR analyses further view advantage computation as reward-distribution estimation under policy-induced binary rewards, where finite rollout estimates can suffer from high variance and gradient starvation as GRPO advantage group mean 0.5 Nie et al. (2026); Mroueh (2025); Kim et al. (2026). To investigate how CoT reasoning affected by spatio-temporal perception under highly informative gradient conditions, we visualize rollout groups whose mean

group reward is > 0.5 (easy questions) and ≤ 0.5 (hard questions), dynamic convergence respectively (further motivation as Appendix I).

Analysis of CoT Reasoning Trajectories. We analyze rollout behavior from two complementary viewpoints: the entropy of correct and incorrect trajectories, as illustrated in Fig. 6. As training progresses, the proportion of rollout groups whose mean reward exceeds 0.5 increases for both GRPO and Token 80/20, suggesting perceptual convergence in the video modality. For GRPO, this proportion grows slowly, first reaching 0.56 at 200 iterations, whereas Token 80/20 reaches 0.60. However, at 600 iterations GRPO attains 0.65, while Token 80/20 stagnates at 0.61. This suggests that Token 80/20, which restricts policy-gradient updates to the top 20% highest-entropy tokens, exhibits unstable convergence or a performance ceiling. Examining the entropy of correct and incorrect trajectories in more detail, across both the group reward > 0.5 and ≤ 0.5 regimes, reveals that correct trajectories tend to exhibit high entropy. Specifically, for group reward > 0.5, correct trajectories exhibit significantly higher entropy than incorrect trajectories on both GRPO and Token 80/20. In this regime, the entropy of incorrect trajectories under GRPO is lower than under Token 80/20 while still remaining high, suggesting that at the group level GRPO may drive low-entropy trajectories toward high entropy. Conversely, for group reward ≤ 0.5, the entropy of correct trajectories under GRPO is lower than under Token 80/20 and tends to collapse to low entropy, suggesting that for hard video questions the group-level policy tends toward low entropy. This indicates that maintaining high entropy as a policy direction may enhance group-level convergence, whereas controlling entropy at the token level may be insufficient for governing group-level behavior. Therefore, in this work we target advantage shaping thought segment-level rather than token-level as main paper subsection 3.1.

D Derivation of Policy Gradient Norm Bounds for Perception-based Reasoning

We now complete the proof of Proposition 3.2. By Lemma A.1, Lemma A.2, Lemma A.3 for the gradient $\nabla_{\mathbf{a}} J$ perception for reasoning with respect to the logits $\mathbf{a}$ is bounded can be written as:

$$\begin{aligned} |w_{i,t}|^2 \left(\|\nabla_a \log \pi_\theta(z_{i,t})\|^2 + 2\langle u, v\rangle \right) &\leq \left\|\nabla_{\mathbf{a}} J\big(y^*_{i,t}\big)\right\|^2 - |w_{i,t}|^2 \Delta_{\mathrm{ent}^H} \\ |w_{i,t}|^2 \left(\|\nabla_a \log \pi_\theta(z_{i,t})\|^2 + 2\langle u, v\rangle \right) &\geq \left\|\nabla_{\mathbf{a}} J\big(y^*_{i,t}\big)\right\|^2 - |w_{i,t}|^2 \Delta_{\mathrm{ent}^L} \end{aligned} \tag{29}$$

Lemma A.1 Following Gradient Norm Decomposition Yang et al. (2025b); Liu et al. (2026a), the squared gradient norm $\|\nabla_a \log \pi_\theta(y_{i,t})\|^2$ requires a summation over all vocabulary entries $n \in \{1, ..., |\mathcal{V}|\}$ as:

$$\|\nabla_a \log \pi_\theta(y_{i,t})\|^2 = |w_{i,t}|^2 \left(1 - 2\pi_\theta\big(y^k_{i,t} \big| z^*_{i,t}\big) + \sum_{n=1}^{|\mathcal{V}|} \pi_\theta\big(y^n_{i,t} \big| z^*_{i,t}\big)^2 \right). \tag{30}$$

Proof adapted from Yang et al. (2025b); Liu et al. (2026a) for Eq.(30). Recall that $\nabla_{a^n} \log \pi_\theta(y^*_{i,t}) = \delta_{kn} - \pi_\theta(y^n_{i,t})$, where δ_{kn} denotes the Kronecker delta. Thus, the gradient of the weighted objective $\mathcal{J}$ is

$$\frac{\partial \mathcal{J}(y_{i,t})}{\partial a^n} = w_{i,t}\big(\delta_{kn} - \pi_\theta(y^n_{i,t})\big) = \begin{cases} w_{i,t}\big(1 - \pi_\theta(y^k_{i,t})\big), & \text{if } n = k, \\ -w_{i,t}\pi_\theta(y^n_{i,t}), & \text{if } n \neq k. \end{cases} \tag{31}$$

Completes the proof by computing the squared ℓ_2-norm by summing over all $n \in \{1, \ldots, |\mathcal{V}|\}$ yields

$$\begin{aligned} \left\|\nabla_a \mathcal{J}(y_{i,t})\right\|^2 &= \big(w_{i,t}(1 - \pi_\theta(y^k_{i,t}))\big)^2 + \sum_{n \neq k} \big(-w_{i,t}\pi_\theta(y^n_{i,t})\big)^2 \\ &= |w_{i,t}|^2 \left[(1 - 2\pi_\theta(y^k_{i,t}) + \pi_\theta(y^k_{i,t})^2) + \sum_{n \neq k} \pi_\theta(y^n_{i,t})^2 \right] \\ &= |w_{i,t}|^2 \left(1 - 2\pi_\theta(y^k_{i,t}) + \sum_{n=1}^{|\mathcal{V}|} \pi_\theta(y^n_{i,t})^2 \right). \end{aligned} \tag{32}$$

Referring to Eq. 30, lower bound can be linked the collision probability to the Shannon entropy. The Rényi entropy of order 2 is defined as

$$H_2(\pi_\theta) = -\log\left(\sum_{n=1}^{|\mathcal{V}|}\pi_\theta(y_{i,t}^n|z_{i,t}^*)^2\right). \tag{33}$$

For discrete probability distributions, the Rényi entropy of order 2 is upper bounded by the Shannon entropy, i.e.,

$$H_2(\pi_\theta) \le \mathcal{H}(\pi_\theta), \tag{34}$$

$$\mathcal{H}(\pi_\theta) = -\sum_{n=1}^{|\mathcal{V}|}\pi_\theta(y_{i,t}^n|z_{i,t}^*)\log\pi_\theta(y_{i,t}^n|z_{i,t}^*). \tag{35}$$

By substituting this inequality into Eq. (7), we obtain an entropy-based lower bound on the reasoning term:

$$\|\nabla_a J(y_{i,t})\|^2 \ge |w_{i,t}|^2\left(1 - 2\pi_\theta(y_{i,t}^k) + e^{-\mathcal{H}(\pi_\theta)}\right). \tag{36}$$

Lemma A.2 Substituting inequality (36) into Eq. (6) results in a deviation-based lower bound for the reasoning objective:

$$\begin{aligned}\Delta_{\text{ent}^H} &= \|\nabla\log\pi_\theta(y_{i,t})\|^2 - |w_{i,t}|^2\left(1 - 2\pi_\theta(y_{i,t}^k) + e^{-\mathcal{H}(\pi_\theta)}\right)\\ &= |w_{i,t}|^2\left[\sum_{n=1}^{|\mathcal{V}|}\pi_\theta\big(y_{i,t}^n\big)^2 - e^{-\mathcal{H}(\pi_\theta)}\right] \ge 0.\end{aligned} \tag{37}$$

For a finite alphabet $\mathcal{V}$, the collision probability satisfies Yang et al. (2025b):

$$\sum_{n=1}^{|\mathcal{V}|}\pi_\theta(y_{i,t}^n)^2 \le 1 - C_{\mathcal{V}}\,\mathcal{H}(\pi_\theta)^2. \tag{38}$$

where $C_{\mathcal{V}} = \frac{|\mathcal{V}|-1}{|\mathcal{V}|\,(\log|\mathcal{V}|)^2}$. Substituting this inequality (38) into Eq.(7) yields the entropy-based reasoning upper bound:

$$|w_{i,t}|^2\left(2 - 2\pi_\theta\big(y_{i,t}^k\big) - C_V\,\mathcal{H}(\pi_\theta)^2\right) \;\ge\; \|\nabla_{\mathbf{a}} J(y_{i,t})\|^2 \tag{39}$$

Lemma A.3 Substituting this inequality into Eq.(6) gives the deviation based reasoning upper bound:

$$\begin{aligned}\Delta_{\text{ent}^L} &= \|\nabla\log\pi_\theta(y_{i,t})\|^2 - |w_{i,t}|^2\left(2 - 2\pi_\theta\big(y_{i,t}^k\big) - C_V\mathcal{H}(\pi_\theta)^2\right)\\ &= |w_{i,t}|^2\left[\sum_{n=1}^{|\mathcal{V}|}\pi_\theta\big(y_{i,t}^n\big)^2 + C_V\mathcal{H}(\pi_\theta)^2 - 1\right] \le 0.\end{aligned} \tag{40}$$

Plug-in Eq.37 into Eq.6, we show that the gradient of perception for reasoning with respect to the logits $\mathbf{a}$ with upper bounded as follows:

$$|w_{i,t}|^2\Big(\,\|\nabla_a\log\pi_\theta(z_{i,t})\|^2 + 2\,\langle u, v\rangle\,\Big) \;\le\; \big\|\nabla_{\mathbf{a}} J\big(y_{i,t}^*\big)\big\|^2 - |w_{i,t}|^2\Delta_{\text{ent}^H} \tag{41}$$

Plug-in Eq.40 into Eq.6, we show that the gradient of perception for reasoning with respect to the logits $\mathbf{a}$ with lower bounded as follows:

$$|w_{i,t}|^2\Big(\,\|\nabla_a\log\pi_\theta(z_{i,t})\|^2 + 2\,\langle u, v\rangle\,\Big) \;\ge\; \big\|\nabla_{\mathbf{a}} J\big(y_{i,t}^*\big)\big\|^2 - |w_{i,t}|^2\Delta_{\text{ent}^L} \tag{42}$$

D.1 Gradient convergence in the low-entropy regime.

We analyze the gradient of the upper entropy deviation Δ_{ent^H} in Eq. 37. Let $p_n := \pi_\theta(y_{i,t}^n \mid z_{i,t}^*)$ with $\sum_n p_n = 1$ and $p_n > 0$, and treat $w_{i,t}$ as a local scalar coefficient with respect to the logit vector $\mathbf{a}$. Recall

$$\Delta_{\text{ent}^H} = |w_{i,t}|^2\left[\sum_{n=1}^{|\mathcal{V}|}p_n^2 - e^{-\mathcal{H}(p)}\right], \qquad \mathcal{H}(p) = -\sum_{n=1}^{|\mathcal{V}|}p_n\log p_n. \tag{43}$$

Since

$$\frac{\partial}{\partial p_k}\sum_n p_n^2 = 2p_k, \qquad \frac{\partial \mathcal{H}(p)}{\partial p_k} = -(1+\log p_k), \tag{44}$$

we obtain

$$\frac{\partial}{\partial p_k} e^{-\mathcal{H}(p)} = e^{-\mathcal{H}(p)}(1+\log p_k). \tag{45}$$

Therefore,

$$\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial p_k} = |w_{i,t}|^2 \left[2p_k - e^{-\mathcal{H}(p)}(1+\log p_k)\right]. \tag{46}$$

Define

$$\alpha_k = 2p_k - e^{-\mathcal{H}(p)}(1+\log p_k). \tag{47}$$

Then

$$\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial p_k} = |w_{i,t}|^2 \alpha_k. \tag{48}$$

Composing this probability-space derivative with the softmax Jacobian

$$\frac{\partial p_n}{\partial a^k} = p_n(\delta_{nk} - p_k) \tag{49}$$

gives

$$\begin{aligned}\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial a^k} = \sum_{n=1}^{|\mathcal{V}|} \frac{\partial \Delta_{\mathrm{ent}^H}}{\partial p_n}\frac{\partial p_n}{\partial a^k} &= |w_{i,t}|^2 \sum_{n=1}^{|\mathcal{V}|} \alpha_n p_n(\delta_{nk} - p_k) \\ &= |w_{i,t}|^2 \left(\sum_n \alpha_n p_n \delta_{nk} - p_k \sum_n \alpha_n p_n\right).\end{aligned} \tag{50}$$

Using the Kronecker delta identity

$$\sum_n \alpha_n p_n \delta_{nk} = \alpha_k p_k, \tag{51}$$

we obtain

$$\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial a^k} = |w_{i,t}|^2 \left(\alpha_k p_k - p_k \sum_n p_n \alpha_n\right). \tag{52}$$

Factoring out p_k gives

$$\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial a^k} = |w_{i,t}|^2 p_k \left(\alpha_k - \sum_n p_n \alpha_n\right). \tag{53}$$

Defining the probability-weighted mean

$$\mathbb{E}_p[\alpha] = \sum_n p_n \alpha_n, \tag{54}$$

the logit-space gradient can be written compactly as

$$\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial a^k} = |w_{i,t}|^2 p_k \big(\alpha_k - \mathbb{E}_p[\alpha]\big). \tag{55}$$

Thus, each logit-gradient component consists of a softmax prefactor p_k and a centered coefficient $\alpha_k - \mathbb{E}_p[\alpha]$. The former suppresses low-probability tail components, while the latter cancels the dominant component in the sharp-policy limit.

Low-entropy limit. The purpose of this derivation is to establish that the entropy-dependent correction becomes asymptotically inactive as the policy approaches a deterministic decision state. Specifically, we show that, in the low-entropy regime,

$$\Delta_{\mathrm{ent}^H} \to 0, \qquad \nabla_{\mathbf{a}} \Delta_{\mathrm{ent}^H} \to 0,$$

so that the local gradient dynamics are eventually governed solely by the perceptual-reference score and its interaction with the reasoning score. Consider the low-entropy regime in which the distribution concentrates on a single index k^*:

$$p_{k^*} \to 1, \qquad p_j \to 0 \quad (j \neq k^*). \tag{56}$$

Then $\mathcal{H}(p) \to 0$ and $e^{-\mathcal{H}(p)} \to 1$. To prove the vanishing-gradient property, it is sufficient to analyze two qualitatively distinct parts of the distribution: (i) the tail components $p_j \to 0$ for $j \neq k^*$, and (ii) the dominant mode $p_{k^*} \to 1$. The gradient vanishes through different mechanisms in these two cases.

Tail components ($j \neq k^*$). For any tail component,

$$\alpha_j = 2p_j - e^{-\mathcal{H}(p)}(1 + \log p_j). \tag{57}$$

Since $0 < e^{-\mathcal{H}(p)} \leq 1$,

$$|\alpha_j| \leq 2p_j + 1 + |\log p_j|. \tag{58}$$

Hence α_j grows at most logarithmically as $p_j \to 0$.

From Eq. 55,

$$\left| \frac{\partial \Delta_{\mathrm{ent}^H}}{\partial a^j} \right| = |w_{i,t}|^2 p_j |\alpha_j - \mathbb{E}_p[\alpha]|. \tag{59}$$

To control the centered term, observe that

$$\begin{aligned} \mathbb{E}_p[\alpha] &= \sum_n p_n \left[2p_n - e^{-\mathcal{H}(p)}(1 + \log p_n) \right] \\ &= 2 \sum_n p_n^2 - e^{-\mathcal{H}(p)} \left[1 - \mathcal{H}(p) \right]. \end{aligned} \tag{60}$$

Therefore $\mathbb{E}_p[\alpha]$ remains bounded, and

$$\mathbb{E}_p[\alpha] \to 1 \qquad (p_{k^*} \to 1).$$

Consequently,

$$p_j |\alpha_j - \mathbb{E}_p[\alpha]| \leq p_j \left(2p_j + 1 + |\log p_j| + |\mathbb{E}_p[\alpha]| \right) \to 0, \tag{61}$$

because $p_j \to 0$ and $p_j |\log p_j| \to 0$. Hence

$$\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial a^j} \to 0 \qquad (j \neq k^*). \tag{62}$$

Thus, every tail component vanishes due to the softmax prefactor p_j, which suppresses the logarithmic growth of α_j.

Dominant mode (k^*). For the dominant component, the softmax prefactor does not vanish: $p_{k^*} \to 1$. Instead, the gradient disappears through cancellation of the centered coefficient. Indeed,

$$\alpha_{k^*} = 2p_{k^*} - e^{-\mathcal{H}(p)}(1 + \log p_{k^*}) \to 2 - 1 = 1, \tag{63}$$

while Eq. 60 yields

$$\mathbb{E}_p[\alpha] \to 1. \tag{64}$$

Therefore,

$$\alpha_{k^*} - \mathbb{E}_p[\alpha] \to 0, \tag{65}$$

and Eq. 55 gives

$$\frac{\partial \Delta_{\mathrm{ent}^H}}{\partial a^{k^*}} = |w_{i,t}|^2 p_{k^*} \left(\alpha_{k^*} - \mathbb{E}_p[\alpha] \right) \to 0. \tag{66}$$

Thus, the mode component vanishes through mean cancellation rather than softmax suppression.

Vanishing correction in the deterministic limit. Combining the two mechanisms above, every coordinate of the logit gradient converges to zero:

$$\nabla_{\mathbf{a}} \Delta_{\mathrm{ent}^H} \to 0 \qquad \text{as} \qquad \mathcal{H}(\pi_\theta) \to 0. \tag{67}$$

In the same regime,

$$\sum_n p_n^2 \to 1, \qquad e^{-\mathcal{H}(p)} \to 1, \tag{68}$$

which implies

$$\Delta_{\mathrm{ent}^H} \to 0. \tag{69}$$

Likewise,

$$\Delta_{\mathrm{ent}^L} = |w_{i,t}|^2 \left[\sum_n p_n^2 + \frac{|\mathcal{V}| - 1}{|\mathcal{V}|(\log |\mathcal{V}|)^2} \mathcal{H}(p)^2 - 1 \right] \to 0. \tag{70}$$

Together with Lemma A.2 and Lemma A.3, these limits show that the entropy-dependent corrections vanish from both sides of the bound. As a result, the squared local gradient norm converges to

$$\left\| \nabla_{\mathbf{a}} J(y^*_{i,t}) \right\|^2 \longrightarrow |w_{i,t}|^2 \left(\left\| \nabla_a \log \pi_\theta(z_{i,t}) \right\|^2 + 2\langle u, v \rangle \right). \tag{71}$$

Therefore, in the deterministic limit, the entropy-dependent terms vanish from both the objective-level bound and the associated gradient field. The optimization dynamics are no longer influenced by uncertainty regularization and collapse to the perceptually grounded component of the update. Consequently, the local gradient magnitude is determined exclusively by the perceptual-reference score and its interaction with the reasoning score, characterizing a stable and sharply concentrated decision state.

On the other hand, Eq. 71 characterizes a *sharp, grounded reasoning* policy state, in which the gradient is dominated by the perceptual score and its interaction with the accumulated reasoning score until the reasoning policy contributes no further action. Such a state is desirable at *late* token positions, where the model must commit to a final answer based on visually consolidated reasoning. Combining Eq. 11 with Eq. 71 yields a *position-dependent* view of entropy allocation along high-probability, high-reward trajectories, which directly explains the high-entropy behaviour in Fig. 1: entropy must be *maintained* at early token positions and *reduced* at late token positions, inducing a *high-to-low* entropy transition within each trajectory.

More specifically, early-token entropy enables the model to progressively extract visual evidence or articulating query-relevant visual information and reconcile it with intermediate reasoning. At late tokens, by contrast, entropy should collapse so that the gradient reduces to $\|\nabla_a \log \pi_\theta(z_{i,t})\|^2 + 2\langle u, v \rangle$, where the residual reasoning-score contribution $\|\nabla_a \log \pi_\theta(y_{i,t})\|^2$ has been absorbed into the perceptual–reasoning interaction term $\langle u, v \rangle$. In this state, the policy can commit to a high-probability final answer grounded in mature, visually consolidated reasoning. Conversely, if entropy remains high at late token positions, the full gradient stays active and the additional reasoning-score term $\|\nabla_a \log \pi_\theta(y_{i,t})\|^2$ biases the policy toward not-yet-consolidated reasoning at the decision stage. This is undesirable: the final answer should be guided by reasoning signals that have already been reconciled and accumulated enough with the visual information, rather than by reasoning components that remain decoupled from the perceptual decision process. As we show next, the sparse credit-assignment mechanism of standard GRPO does not explicitly induce this position-dependent entropy trade-off.

E Derivation of Entropy Update Mechanism on Reasoning Policy

Lemma 3.2 Entropy Update Mechanism from Cui et al. (2025) for each training step λ:

$$\begin{aligned} \mathcal{H}(\pi_\theta^{\lambda+1}) - \mathcal{H}(\pi_\theta^{\lambda}) &\approx \mathbb{E}_{S_c \sim d_{\pi_\theta}} \left[-\operatorname{Cov}_{z_a \sim \pi_\theta^\lambda(\cdot|S_c)} \left(\log \pi_\theta^\lambda(z_a \mid S_c),\, z_{s,a}^{\lambda+1} - z_{s,a}^{\lambda} \right) \right] \\ &\approx -\eta \operatorname{Cov}_{z_a \sim \pi_\theta^\lambda(\cdot|S_c)} \left(\log \pi_\theta^\lambda(z_a \mid S_c), \pi_\theta^\lambda(z_a \mid S_c) \cdot (r, \hat{A}(S_c, z_a)) \right). \end{aligned} \tag{72}$$

Proof adapted from Liu (2025); Cui et al. (2025), in each state-action pair (S_c, z_a) is associated with an perception logit $z_{s,a}$. We denote by π_θ^{pe} and π_θ^{re} the policies parameterized by the perception and reasoning logits, respectively. We assume that the logits are updated as

$$\begin{aligned} z^{\lambda+1} &= z^{\lambda} + \eta \nabla J(\pi_\theta^{pe}) \\ y^{\lambda+1} &= y^{\lambda} + \eta \nabla J(\pi_\theta^{re}) \end{aligned} \tag{73}$$

Given that η is sufficiently small, applying a first-order Taylor expansion yields

$$\mathcal{H}(\pi_\theta^{re^{\lambda+1}} \mid S_c) \approx \mathcal{H}(\pi_\theta^{re^{\lambda}} \mid S_c) + \left\langle \nabla \mathcal{H}(\pi_\theta^{re^{\lambda}} \mid S_c),\ z^{\lambda+1} - z^{\lambda} \right\rangle \tag{74}$$

We next derive the form of $\nabla \mathcal{H}(\pi_\theta^{re} \mid S_c)$. By definition of the entropy,

$$\begin{aligned} \nabla_\theta \mathcal{H}(\pi_\theta^{re} \mid S_c) &= \nabla_\theta \mathbb{E}_{z_a \sim \pi_\theta^{pe}(S_c)} \big[-\log \pi_\theta^{pe}(z_a \mid S_c) \big] \\ &= \nabla_\theta \left(-\mathbb{E}_{z_a \sim \pi_\theta^{pe}(S_c)} \big[\log \pi_\theta^{pe}(z_a \mid S_c) \big] \right) \\ &= -\mathbb{E}_{z_a \sim \pi_\theta^{pe}(S_c)} \Big[\nabla_\theta \log \pi_\theta^{pe}(z_a \mid S_c) + \log \pi_\theta^{pe}(z_a \mid S_c) \nabla_\theta \log \pi_\theta^{pe}(z_a \mid S_c) \Big] \\ &= -\mathbb{E}_{z_a \sim \pi_\theta^{pe}(S_c)} \Big[\log \pi_\theta^{pe}(z_a \mid S_c) \nabla_\theta \log \pi_\theta^{pe}(z_a \mid S_c) \Big]. \end{aligned} \tag{75}$$

Therefore, we have

$$\begin{aligned} &\left\langle \nabla_\theta \mathcal{H}(\pi_\theta^{re} \mid S_c),\ z^{\lambda+1} - z^{\lambda} \right\rangle \\ &= -\mathbb{E}_{z_a \sim \pi_\theta^{pe}(S_c)} \Big[\log \pi_\theta^{pe}(z_a \mid S_c) \nabla_\theta \log \pi_\theta^{pe}(z_a \mid S_c)^\top (\theta^{pe^{\lambda+1}} - \theta^{pe^{\lambda}}) \Big] \\ &= -\mathbb{E}_{z_a \sim \pi_\theta^{pe}(S_c)} \Big[\log \pi_\theta^{pe}(z_a \mid S_c) \big((\theta^{\lambda+1}_{S_c, z_a} - \theta^{\lambda}_{S_c, z_a}) - \sum_{z'_a \in \mathcal{A}} \pi_\theta^{pe}(z'_a \mid S_c)(\theta^{\lambda+1}_{S_c, z'_a} - \theta^{\lambda}_{S_c, z'_a}) \big) \Big] \\ &= -\operatorname{Cov}_{z_a \sim \pi_\theta^{pe}(|S_c)} \big(\log \pi_\theta^{pe}(z_a \mid S_c),\ z^{\lambda+1}_{s,a} - z^{\lambda}_{s,a} \big). \end{aligned} \tag{76}$$

Through gradient backtracking, the logit is updated as

$$z^{k+1}_{s,a} = z^{k}_{s,a} + \eta \nabla_{\theta^{pe}_{s,a}} J(\theta^{pe}). \tag{77}$$

Therefore, we have

$$\begin{aligned} z^{k+1}_{s,a} - z^{k}_{s,a} &= \eta \nabla_{\theta^{pe}_{s,a}} J(\theta^{pe}) \\ &= \eta\, \mathbb{E}_{a' \sim \pi_\theta^{pe}(\cdot|s)} \big[\nabla_{\theta^{pe}_{S_c, z_a}} \log \pi_\theta^{pe}(z_{a'} \mid S_c)\, A(S_c, z_{a'}) \big] \\ &= \eta\, \pi_\theta^{pe}(z_a \mid S_c) \Big[(1 - \pi_\theta^{pe}(z_a \mid S_c)) A(S_c, z_a) - \sum_{z_{a'} \in \mathcal{A},\, z_{a'} \neq z_a} \pi_\theta^{pe}(a' \mid s) A(S_c, z_{a'}) \Big] \\ &= \eta\, \pi_\theta^{pe}(z_a \mid S_c) \Big[A(S_c, z_a) - \sum_{z_{a'} \in \mathcal{A}} \pi_\theta^{pe}(z_{a'} \mid S_c) A(S_c, z_{a'}) \Big] \\ &= \eta\, \pi_\theta^{pe}(z_a \mid S_c) \left[A(S_c, z_a) - 0 \right] \\ &= \eta\, \pi_\theta^{pe}(z_a \mid S_c)\, A(S_c, z_a). \end{aligned} \tag{78}$$

Identifying $A(S_c, z_a)$ as $r_{i,t}(\theta)\hat{A}(S_c, z_a)$ completes the proof. Where $S_c \sim d_{\pi_\theta}$ is state distribution and $z_{s,a}$ is the logit of action a at state s (distinct from the perceptual latent $z_{i,t}$ used elsewhere; we keep the $z_{s,a}$ only inside this lemma to remain consistent with Cui et al. (2025)). Equation 5 has two consequences for our setting.

First, since the policy gradient is invariant to additive shifts of the logits, the effect of the advantage signal can be absorbed into a nearly uniform logit offset, $z^{\lambda+1}_{s,a} - z^{\lambda}_{s,a} \approx \mathrm{const}$, suggesting negligible update of $r_{i,t}(\theta)\hat{A}(S_c, z_a)$. The covariance in Eq. 5 then vanishes, and $\mathcal{H}(\pi_\theta)$ is left almost unchanged between successive steps, suggesting reasoning policy under-training. Hence, in sparse video RLVR, high entropy can persist even while the policy concentrates increasing probability mass on rewarded trajectories.

Second, because the *same* scalar $w_{i,t} = r_{i,t}(\theta)\hat{A}(S_c, z_a)$ multiplies both directional components

of Eq. 4, the perceptual score $\nabla_\theta \log \pi_\theta(z_{i,t} \mid q, C, y_{i,<t})$ and the reasoning score $\nabla_\theta \log \pi_\theta(y_{i,t} \mid q, C, y_{i,<t}, z_{i,t})$, the two are rescaled *identically*. Consequently, the update cannot disentangle whether a reward was earned by selecting the visual information or by exploiting a linguistic prior during generation. Once the generation likelihood becomes easy to increase, the policy can improve outcome reward without acquiring a more reliable perceptual reference.

F Gradient Decomposition and Perception State Design

We analyze how policy-gradient updates perturb the output logits and, consequently, token-level log-probabilities. This analysis shows that log-probability changes depend only on centered logit perturbations, and clarifies when perceptual and reasoning gradients induce coherent updates.

Logit derivatives. Let $a_{i,t} \in \mathbb{R}^{|\mathcal{V}|}$ denote the logit vector at generation step t. Differentiating the log-softmax with respect to the j-th logit gives

$$\frac{\partial}{\partial a_{i,t}(j)} \log \pi_\theta(y_{i,t}=k) = \mathbf{1}\{j=k\} - \pi_\theta(y=j \mid s_{i,t}), \tag{79}$$

and therefore

$$\nabla_{a_{i,t}} \log \pi_\theta(y_{i,t} \mid s_{i,t}) = e_{y_{i,t}} - \pi_\theta(\cdot \mid s_{i,t}), \tag{80}$$

where $e_{y_{i,t}} \in \mathbb{R}^{|\mathcal{V}|}$ is the one-hot vector corresponding to token $y_{i,t}$.

For a small logit perturbation $\delta a_{i,t}$, a first-order Taylor expansion gives

$$\delta \log \pi_\theta(y_{i,t} \mid s_{i,t}) = \nabla_{a_{i,t}} \log \pi_\theta(y_{i,t} \mid s_{i,t})^\top \delta a_{i,t} + O\big(\|\delta a_{i,t}\|^2\big). \tag{81}$$

Substituting Eq. 80, we obtain

$$\begin{aligned} \delta \log \pi_\theta(y_{i,t}) &= \big(e_{y_{i,t}} - \pi_\theta(\cdot \mid s_{i,t})\big)^\top \delta a_{i,t} + O\big(\|\delta a_{i,t}\|^2\big) \\ &= \delta a_{i,t}(y_{i,t}) - \mathbb{E}_{y \sim \pi_\theta(\cdot \mid s_{i,t})}\big[\delta a_{i,t}(y)\big] + O\big(\|\delta a_{i,t}\|^2\big). \end{aligned} \tag{82}$$

Thus, the first-order change in log-probability is determined by the *centered* logit perturbation. Uniform shifts of all logits cancel out, and only the deviation of the selected token's perturbation from the policy-averaged perturbation contributes to the update.

Following Eq. 4, we decompose the score function into perceptual and reasoning components:

$$u_{i,t}^{\text{perc}} = \nabla_a \log \pi_\theta(z_{i,t} \mid s_{i,t}), \qquad v_{i,t}^{\text{reason}} = \nabla_a \log \pi_\theta(y_{i,t} \mid s_{i,t}, z_{i,t}). \tag{83}$$

With the perception–reasoning trade-off coefficient ζ and the reward-to-go weight $w_{i,t} = \sum_{t'=t}^{|y_i|} r_{i,t'}$, the induced first-order logit update can be written as

$$\delta a_{i,t} = \zeta\, w_{i,t}\big(u_{i,t}^{\text{perc}} + v_{i,t}^{\text{reason}}\big) = \zeta \sum_{t'=t}^{|y_i|} r_{i,t'}\big(u_{i,t}^{\text{perc}} + v_{i,t}^{\text{reason}}\big). \tag{84}$$

When the two gradients are aligned, $\langle u, v\rangle/(\|u\|\|v\|) \approx 1$, the update coherently reinforces both perception policy and reasoning policy quality. However, the standard GRPO advantage does not explicitly distinguish perception policy states from reasoning policy steps. As a result, the scalar weight $w_{i,t}$ may amplify both components without ensuring that they point in compatible directions, which is consistent with the instability discussed in Proposition 3.1 of Yao et al. (2026b).

F.1 Optimization of Reasoning Direction Scaling

Reasoning direction. Let π^{perc} denote the RL post-trained perception policy and π^{reason} the frozen reasoning reference policy, and let $z^{\text{perc}}(s_t), z^{\text{reason}}(s_t) \in \mathbb{R}^{|\mathcal{V}|}$ be their next-token logit vectors at a visited prefix s_t. To isolate the policy change induced by post-training, we remove the softmax-invariant offset with the centering operator $\mathcal{C}(\mathbf{v}) := \mathbf{v} - \frac{1}{|\mathcal{V}|}(\mathbf{1}^\top \mathbf{v})\mathbf{1}$ and define

$$\Delta_t := \mathcal{C}(z^{\text{perc}}(s_t) - z^{\text{reason}}(s_t)) = \mathcal{C}(\log \pi^{\text{perc}}(\cdot \mid s_t) - \log \pi^{\text{reason}}(\cdot \mid s_t)), \tag{85}$$

where Δ_t is invariant to the per-state normalizers. Thus Δ_t is the centered log-policy ratio, which admits the standard implicit-reward reading under KL-regularized policy optimization. We define its unit direction as

$$d_t := \frac{\Delta_t}{\max\{\|\Delta_t\|_2,\ \varepsilon\}}, \qquad \varepsilon > 0, \tag{86}$$

so that d_t is well defined when $\Delta_t = \mathbf{0}$ (i.e., when the two policies agree at s_t, no direction is imposed).

Directional decomposition. For rollout i and token t, let $A_{i,t}$ be the advantage and let

$$v_{i,t}^{\text{reason}} := A_{i,t}\, \nabla_z \log \pi_\theta\big(y_{i,t} \mid s_{i,t}\big) \ \in \ \mathbb{R}^{|\mathcal{V}|} \tag{87}$$

denote the reasoning policy gradient *with respect to the next-token logits* z, so that $v_{i,t}^{\text{reason}}$ and d_t live in the same space.[1] With alignment coefficient $\langle d_t, v_{i,t}^{\text{reason}}\rangle$, the projected and orthogonal components are $\mathrm{P}_{d_t}\big(v_{i,t}^{\text{reason}}\big) := \big(d_t^\top v_{i,t}^{\text{reason}}\big) d_t$ and $\big(v_{i,t}^{\text{reason}}\big)^\perp := v_{i,t}^{\text{reason}} - \mathrm{P}_{d_t}\big(v_{i,t}^{\text{reason}}\big)$.

Anisotropic scaling. Given a nonnegative scale $\zeta_{i,t} \geq 0$, we can amplify only the component aligned with the perception-reasoning direction:

$$\tilde{v}_{i,t}^{\text{reason}} := v_{i,t}^{\text{reason}} + \zeta_{i,t}\, \mathrm{P}_{d_t}\big(v_{i,t}^{\text{reason}}\big) = \underbrace{(1+\zeta_{i,t})\, \mathrm{P}_{d_t}\big(v_{i,t}^{\text{reason}}\big)}_{\text{encouraged}} + \underbrace{\big(v_{i,t}^{\text{reason}}\big)^\perp}_{\text{GRPO advantage corrected}}, \tag{88}$$

which is equivalently the linear map

$$\tilde{v}_{i,t}^{\text{reason}} = M_{i,t}\, v_{i,t}^{\text{reason}}, \qquad M_{i,t} := I + \zeta_{i,t}\, d_t d_t^\top. \tag{89}$$

Proposition 1 (Per-token stationarity)**.** *Fix $\zeta_{i,t} \geq 0$ and let $\|d_t\|_2 = 1$. The map $M_{i,t} = I + \zeta_{i,t} d_t d_t^\top$ acts as*

$$M_{i,t}\, d_t = (1+\zeta_{i,t})\, d_t, \qquad M_{i,t}\, u = u \quad \textit{for every } u \perp d_t, \tag{90}$$

i.e. it stretches the single direction d_t by the factor $1+\zeta_{i,t}$ and leaves the orthogonal complement untouched. Its eigenvalues $1+\zeta_{i,t}$, hence $M_{i,t}$ is symmetric positive definite and invertible, with $M_{i,t}^{-1} = I - \frac{\zeta_{i,t}}{1+\zeta_{i,t}} d_t d_t^\top$. Consequently,

$$\tilde{v}_{i,t}^{\text{reason}} = \mathbf{0} \iff v_{i,t}^{\text{reason}} = \mathbf{0}. \tag{91}$$

Since Eq.91 holds at every token, the reweighting only rescales the update along d_t: it never creates or removes a per-token zero, and it reduces to the identity whenever $d_t^\top v_{i,t}^{\text{reason}} = 0$ or $\Delta_t = \mathbf{0}$. Any point at which all per-token gradients vanish thus remains stationary. This is a per-token statement; we do not claim that stationary points of the aggregated objective are preserved.

G Relationship Between Perceptual Learning and Reasoning Policy produced CoT Length

We next justify why the fusion-aware discount should depend on the response length. Let $y_i = (y_{i,1}, \ldots, y_{i,L_i})$ be a reasoning trajectory of length $L_i = |y_i|$. At each step t, the model selects a perceptual abstraction $z_{i,t} \in \mathcal{Z}$. From Eq. 83, the trajectory-level visual-reference gradient aggregates the perceptual score gradients:

$$G_{\text{VR}} \triangleq \sum_{t=1}^{L_i} u_{i,t}^{\text{perc}}, \qquad u_{i,t}^{\text{perc}} = \nabla_\theta \log \pi_\theta(z_{i,t} \mid q, C, y_{i,<t}). \tag{92}$$

[1] Parameter gradients are recovered by the chain rule, $\nabla_\theta = J_\theta^\top v_{i,t}^{\text{reason}}$ with $J_\theta = \partial z / \partial \theta$; the reweighting below is therefore applied in logit space and back-propagated unchanged.

Perceptual diversity and gradient accumulation. Let $\mathcal{Z}_i = \{z_{i,t} \mid t = 1, \ldots, L_i\}$ denote the set of distinct perceptual states visited in the trajectory, with $S_i = |\mathcal{Z}_i| \leq L_i$. Let $n_i(z) = |\{t : z_{i,t} = z\}|$ be the visitation count of state z. Assuming that gradients from repeated visits to the same perception policy provided information are positively correlated, while gradients across different states are weakly correlated, the expected squared norm of the visual-reference gradient scales as

$$\mathbb{E}\big[\|G_{\mathrm{VR}}\|^2\big] \approx \sum_{z \in \mathcal{Z}_i} \big(n_i(z) c_z\big)^2 \propto \sum_{z \in \mathcal{Z}_i} n_i(z)^2, \tag{93}$$

where c_z denotes the typical per-visit gradient scale for state z. Under a uniform-visit approximation $n_i(z) \approx L_i/S_i$, we obtain

$$\mathbb{E}\big[\|G_{\mathrm{VR}}\|^2\big] \propto S_i \left(\frac{L_i}{S_i}\right)^2 = \frac{L_i^2}{S_i}, \qquad \sqrt{\mathbb{E}[\|G_{\mathrm{VR}}\|^2]} \propto \frac{L_i}{\sqrt{S_i}}. \tag{94}$$

This reveals a length–diversity trade-off. When the reasoning policy repeatedly reuses a small number of information provided by perception policy, S_i is small and perceptual gradients can accumulate almost linearly with L_i. In contrast, when the reasoning policy visits diverse information provided by perception policy, the accumulation is naturally distributed across states and the growth is suppressed.

Justification of the discount design. Eq. 94 suggests that long responses can over-amplify perceptual gradients when perceptual diversity is limited by insufficient context extracting due to lack input frames. The fusion-state value $V^p_{i,\tau,s}$ is designed to measure how strongly a perceptual state may contribute to such gradient accumulation. A large $V^p_{i,\tau,s}$ indicates that the perceptual update may move beyond a conservative gradient-bounded regime or be repeatedly reinforced by similar visual evidence. A small $V^p_{i,\tau,s}$, in contrast, indicates that the update closer to the trust upper bounded region, where perceptual-centric learning signals should be preserved. We therefore define the fusion-aware discount as

$$\gamma_{i,\tau,s} = 2^{-V^p_{i,\tau,s}/|y_i|} \in (0, 1]. \tag{95}$$

This design has two desirable properties. First, normalizing the exponent by $|y_i|$ makes the per-step discount milder for longer responses, thereby preventing the perception policy from being excessively penalized when the reasoning policy fails to effectively leverage visual information.

Second, larger $V^p_{i,\tau,s}$ values induce stronger discounting, suppressing redundant or unstable perceptual-gradient accumulation. When $V^p_{i,\tau,s}$ is small, $\gamma_{i,\tau,s}$ approaches one, preserving useful perceptual learning signals within the local gradient-bounded regime. Thus, the proposed discount adaptively regulates perceptual learning by coupling fusion-state reliability with response length. It attenuates updates that may exceed a conservative perception–reasoning gradient bound, while retaining informative perceptual-centric contributions.

H EUBO-Based Step-wise Evidence Decomposition

This appendix substantiates the main-text claim that *we derive an Evidence Upper Bound (EUBO) Hinton et al. (1995); Grosse et al. (2015) and show that policy optimization on a step-wise linear utility tightens it, thereby reducing the perception–reasoning fusion gap.* The argument proceeds in four steps: (i) an exact decomposition of $\log \pi_\phi(y \mid x)$ that isolates the perceptually meaningful component; (ii) a Lemma deriving the EUBO, whose slack is identified with the perception–reasoning fusion gap; (iii) a Proposition establishing a step-wise *linear utility* U_t that governs local posterior credit; (iv) a Corollary showing that policy optimization on $\{U_t\}$ causes both the EUBO slack and the induced fusion gap to converge to zero. Throughout, Z^r is the sampled reasoning trace and $Z^p \subseteq Z^r$ its perceptual sub-trace.

(i) Exact evidence decomposition. Bayes' rule gives the answer-conditioned posterior

$$q_\phi(Z^r) \;\equiv\; \pi_\phi(Z^r \mid x, y) \;=\; \frac{\pi_\phi(y \mid x, Z^r)\, \pi_\phi(Z^r \mid x)}{\pi_\phi(y \mid x)}. \tag{96}$$

Introducing a reference perception prior $\pi_\theta(Z^p \mid x)$ yields the exact decomposition

$$\log \pi_\phi(y \mid x) = \underbrace{\mathbb{E}_{q_\phi}[\log \pi_\phi(y \mid x, Z^p)]}_{\text{perception-conditioned evidence}} + \underbrace{\mathbb{E}_{q_\phi}\left[\log \frac{\pi_\phi(Z^r|x)}{\pi_\theta(Z^p|x)}\right]}_{G(\phi)\text{: posterior-weighted prior gap}} - D_{\mathrm{KL}}\big(q_\phi(Z^r) \,\big\|\, \pi_\theta(Z^p \mid x)\big)\,. \tag{97}$$

Only $G(\phi)$ carries a learnable, perceptually meaningful signal; the KL term is non-negative and quantifies the gap between the answer-conditioned posterior and the reference perceptual prior.

(ii) Derivation of the Evidence Upper Bound. Dropping the non-negative KL term in Ep.97 yields a tractable upper bound on the log-evidence.

Lemma 2 (Evidence Upper Bound)**.** *For any reference prior $\pi_\theta(Z^p \mid x)$,*

$$\log \pi_\phi(y \mid x) \;\le\; \underbrace{\mathbb{E}_{q_\phi}\big[\log \pi_\phi(y \mid x, Z^p)\big] + G(\phi)}_{\mathrm{EUBO}(\phi)}, \tag{98}$$

with slack $D_{\mathrm{KL}}\big(q_\phi(Z^r) \,\|\, \pi_\theta(Z^p \mid x)\big)$, which we identify as the perception–reasoning policy fusion gap.

Proof. Rearranging Ep.97 and applying Gibbs' inequality $D_{\mathrm{KL}}(\cdot\|\cdot) \ge 0$ gives Ep.98; equality holds iff $q_\phi(Z^r) = \pi_\theta(Z^p \mid x)$. □

Lemma 2 realises the first half of our claim: an EUBO is derived, and its slack is precisely the perception–reasoning fusion gap, we seek to reduce.

(iii) Step-wise linear utility. To tighten the EUBO at rollout time, we decompose the slack step-wise and identify a local utility whose optimization contracts it.

Proposition 3 (Step-wise evidence gain)**.** *Let z_t^r be the reasoning token at step t, y_t the answer token, and z^p a perceptual anchor (articulated visual information). The local posterior factorizes as*

$$q_\phi(z_t^r \mid x, y_t, z_{<t}^r) = \frac{\pi_\phi(y_t \mid x, z_{\le t}^r)\, \pi_\phi(z_t^r \mid x, z_{<t}^r)}{\pi_\phi(y_t \mid x, z_{<t}^r)}, \tag{99}$$

and the local prediction gain of a candidate z is

$$\Delta_t(z) \;\triangleq\; \log \frac{\pi_\phi(y_t \mid x, z_{<t}^r, z)}{\pi_\phi(y_t \mid x, z_{<t}^r)}. \tag{100}$$

Assume

$$\textit{(A1) Evidence alignment:} \quad \pi_\phi(y_t \mid x, z_{\le t}^r) \longrightarrow \pi_\phi(y_t \mid x, z_{<t}^r, z^p), \tag{101}$$

$$\textit{(A2) Prior alignment:} \quad \pi_\phi(z_t^r \mid x, z_{<t}^r) \longrightarrow \pi_\theta(z^p \mid x). \tag{102}$$

Then

$$\log q_\phi(z_t^r \mid x, y_t, z_{<t}^r) \;\longrightarrow\; \underbrace{\log \pi_\theta(z^p \mid x)}_{\textit{perceptual prior}} + \underbrace{\Delta_t(z^p)}_{\textit{answer-evidence gain}}. \tag{103}$$

Proof sketch. Taking log of Ep.99 gives $\log q_\phi = \log \pi_\phi(z_t^r \mid x, z_{<t}^r) + \Delta_t(z_t^r)$. Applying (A1) to the second term and (A2) to the first yields Ep.103. □

Eq. 103 motivates the *step-wise linear utility*

$$U_t(z) \;\triangleq\; \log \pi_\theta(z \mid x) + \Delta_t(z), \qquad t = 1, \ldots, T, \tag{104}$$

which combines the perceptual prior with the answer-evidence gain and serves as the local optimization target inside RLVR credit assignment.

(iv) Policy optimization on U_t tightens the EUBO. The next result closes the loop: optimizing U_t drives the EUBO slack, and hence the perception–reasoning fusion gap, to zero.

Corollary 4 (Step-wise linear utility tightens the EUBO)**.** *If policy optimisation on $\{U_t\}_{t=1}^{T}$ enforces (A1)–(A2) uniformly in t together with the self-consistency condition $\Delta_t(z^p)\to 0$ (A3), then*

$$D_{\mathrm{KL}}\big(q_\phi(Z^r)\,\|\,\pi_\theta(Z^p\mid x)\big)\longrightarrow 0, \tag{105}$$

i.e. the EUBO in Lemma 2 becomes tight and the perception–reasoning policy fusion gap is eliminated.

Proof. Step 1 (chain rule). The autoregressive factorisations of $q_\phi(Z^r)$ and $\pi_\theta(Z^p\mid x)$ combined with the KL chain rule give

$$D_{\mathrm{KL}}\big(q_\phi(Z^r)\,\|\,\pi_\theta(Z^p\,|\,x)\big)=\sum_{t=1}^{T}\mathbb{E}_{q_\phi(z^r_{<t})}\big[\mathcal{D}_t\big], \tag{106}$$

$$\mathcal{D}_t\;\triangleq\;D_{\mathrm{KL}}\big(q_\phi(z^r_t\,|\,x,y_t,z^r_{<t})\,\big\|\,\pi_\theta(z^p\,|\,x)\big). \tag{107}$$

It thus suffices to show $\mathcal{D}_t\to 0$ for every t.

Step 2 (local decomposition). Taking log of the local Bayes posterior Ep.99,

$$\log q_\phi(z^r_t\,|\,x,y_t,z^r_{<t})=\log\pi_\phi(z^r_t\,|\,x,z^r_{<t})+\Delta_t(z^r_t), \tag{108}$$

where the denominator $\pi_\phi(y_t\,|\,x,z^r_{<t})$ has been absorbed into the definition of Δ_t so that q_ϕ remains normalised. Substituting into $\mathcal{D}_t$ gives the clean two-term decomposition

$$\mathcal{D}_t=\underbrace{\mathbb{E}_{q_\phi}\left[\log\frac{\pi_\phi(z^r_t\,|\,x,z^r_{<t})}{\pi_\theta(z^p\,|\,x)}\right]}_{\text{(I) prior mismatch under the posterior}}+\underbrace{\mathbb{E}_{q_\phi}\big[\Delta_t(z^r_t)\big]}_{\text{(II) expected evidence gain}}. \tag{109}$$

Step 3 (apply A1–A2). Under (A2), $\pi_\phi(\cdot\,|\,x,z^r_{<t})\to\pi_\theta(\cdot\,|\,x)$ pointwise, so the log-ratio in (I) tends to 0 and hence (I) $\longrightarrow$ 0. Under (A1), $\Delta_t(z^r_t)\to\Delta_t(z^p)$ on the support of q_ϕ, and by (A3) $\Delta_t(z^p)\to 0$, giving (II) $\longrightarrow$ $\Delta_t(z^p)$ $\longrightarrow$ 0.

Step 4 (uniform closure). Combining Steps 2–3 yields $\mathcal{D}_t\to 0$ uniformly in t; substituting into Ep.106 establishes Ep.105. By Lemma 2, this slack is precisely the perception–reasoning policy fusion gap, which therefore vanishes. □

Summary. Lemma 2 derives the Evidence Upper Bound (EUBO), while Corollary 4 shows that optimizing the step-wise linear utility U_t from Proposition 3 tightens this bound and thereby mitigates the perception–reasoning policy fusion gap. Together with Lemma D, which establishes that the target policy satisfies Assumptions A1–A3 associated with Ep.12, these results provide a theoretical justification for the proposed approach and establish the claim stated in the main paper.

I Properties of the GRPO Advantage at a Group Mean of 0.5

Recall Proposition 3.1 from Yao et al. (2026b). Fix a prompt x, let $a\sim\pi_\theta(\cdot\mid x)$ be a sampled response, and assume a binary verifier reward $r(x,a)\in\{0,1\}$. Define

$$p_\theta(x)\;:=\;\mathbb{E}_{a\sim\pi_\theta(\cdot|x)}[r(x,a)]\,. \tag{110}$$

We analyze the population version of GRPO, replacing the empirical group mean and standard deviation by

$$\mu(x)=p_\theta(x),\qquad\sigma(x)=\sqrt{p_\theta(x)\big(1-p_\theta(x)\big)}, \tag{111}$$

for $p_\theta(x)\in(0,1)$. The corresponding population GRPO advantage is

$$A^{\mathrm{GRPO}}(x,a)=\frac{r(x,a)-p_\theta(x)}{\sqrt{p_\theta(x)\big(1-p_\theta(x)\big)}}. \tag{112}$$

Proof of Proposition 3.1. By the likelihood-ratio identity,

$$\nabla_\theta p_\theta(x) = \nabla_\theta \, \mathbb{E}_{a\sim\pi_\theta(\cdot|x)}[r(x,a)] = \mathbb{E}_{a\sim\pi_\theta(\cdot|x)}[r(x,a)\,\nabla_\theta \log \pi_\theta(a \mid x)]\,. \tag{113}$$

Also,

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|x)}[\nabla_\theta \log \pi_\theta(a \mid x)] = 0. \tag{114}$$

Therefore,

$$\begin{aligned} &\mathbb{E}_{a\sim\pi_\theta(\cdot|x)}\big[A^{\mathrm{GRPO}}(x,a)\,\nabla_\theta \log \pi_\theta(a \mid x)\big] \\ &\quad= \frac{\mathbb{E}_{a\sim\pi_\theta(\cdot|x)}\big[\big(r(x,a) - p_\theta(x)\big)\nabla_\theta \log \pi_\theta(a \mid x)\big]}{\sqrt{p_\theta(x)\big(1-p_\theta(x)\big)}} \\ &\quad= \frac{\nabla_\theta p_\theta(x)}{\sqrt{p_\theta(x)\big(1-p_\theta(x)\big)}}. \end{aligned} \tag{115}$$

Finally,

$$\frac{d}{dp}\, 2\arcsin\sqrt{p} = \frac{1}{\sqrt{p(1-p)}}, \tag{116}$$

so the last display equals

$$\nabla_\theta\Big[\,2\arcsin\sqrt{p_\theta(x)}\,\Big]\,. \tag{117}$$

□

This establishes that the population standardized GRPO update is equivalent to the gradient of the arcsin-transformed success probability, rather than that of the raw success probability itself. The normalization by $\sqrt{p(1-p)}$ therefore induces a nonlinear reparameterization of the objective. In particular, the induced sensitivity is governed by $\ell'(p) = 1/\sqrt{p(1-p)}$, which varies across reward regimes and becomes increasingly pronounced near the binary boundaries, reinforcing the clamp of partition groups according to whether their mean reward is above or below 0.5 in the subsequent analysis.

J The Motivating for indirect Perception Modeling via Reasoning Policy State Perspectives

The first GRPO framework for the video modality, Video-R1 Feng et al. (2025), introduces a length reward into GRPO and demonstrates its impact. Removing this length reward leads to a clear decline in response length during RL training. This suggests that, without the length reward, the model fails to preserve a deep reasoning style and instead favors shorter, potentially less informative responses. Building upon Feng et al. (2025), VideoAuto-R1 Liu et al. (2026b) shows that Video-R1 Feng et al. (2025), Time-R1 Wang et al. (2025g), and VideoChat-R1 Li et al. (2025c) degrade performance when equipped with CoT, revealing a fusion gap between visual perception and CoT reasoning. To address this, they propose reasoning reviewed answer training scheme that adaptively removes CoT reasoning according to a confidence-based criterion. This scheme attains higher performance while producing a substantially shorter CoT, indicating that coercion cut out reasoning policy benefits performance.

Prior work has shown that CoT response length is directly affected by perception policy optimization, and that controlling response length is a promising way to alleviate the perception–reasoning fusion gap. This motivates the analysis presented in Appendix G, from which we introduce a response-based, fusion-state-dedicated ratio into policy optimization. This ratio encourages the model to identify and express visual information within <perceive>…</perceive>, thereby shortening the reasoning trace enclosed in <think>…</think>.

Recent work has identified entropy as a critical signal for advantage shaping in GRPO Deng et al. (2025b). In text-centric settings, subsequent methods leverage entropy to refine the optimization signal, either by selectively updating high-entropy tokens Wang et al. (2025e) or by augmenting advantage estimates with entropy-based terms Deng et al. (2025b); Agarwal et al. (2025); Wang et al. (2025c). These methods effectively improve exploration at uncertain decision points and introduce

finer-grained advantage differentiation that better distinguishes token importance both within and across trajectories Le et al. (2025). Motivated by this success in text-centric GRPO, we ask whether entropy can likewise be exploited for the video modality.

As shown in Appendix C, selectively updating high-entropy tokens Wang et al. (2025e) is detrimental and introduces confusion in the video setting. To the best of our knowledge, no prior approach augments advantage estimates with entropy-based terms for the video modality. In this work, we exploit the potential of augmenting advantage estimates with an entropy-based gradient-descent signal to explicitly structure the CoT, rather than introducing a latent representation that learns perception implicitly and is therefore confined to non-verbal or purely computational-linguistic expression, which limits its applicability to guiding policy optimization for the construction of reasoning CoT trajectories. We argue that native perception modeling, together with a reduced perception–reasoning fusion gap, enables the model to identify useful visual information and then render it as a coherent language representation. Building on this motivation, we introduce a training framework in which perception modeling is learned as a coherent language representation within `<perceive>` ... `</perceive>` and is optimized via its utility for removing the uncertainty of the reasoning process, which reflects how much the visual information contributes to reasoning toward the final answer and can able to -consistency reduce fusion gap shown as AppendixH.

A natural concern is how much of the improvement stems from perception modeling itself, as opposed to entropy control, length bias, or generic reward reweighting alone. We take this question as a central design constraint rather than a post-hoc check. While the contribution of perception modeling cannot be evaluated in complete isolation within the CoT-based policy optimization framework, since reasoning policy optimization is an indispensable component of that framework, our formulation introduces itperception modeling as the only additional factor relative to standard CoT-based policy optimization. This allows entropy reduction, CoT length, and reward reweighting magnitude to be matched across settings while varying the perception–reasoning fusion signal used for reward assignment. Under these matched controls, the dominant gain is attributable to perception modeling through its influence on the CoT reasoning state. Moreover, the gain scales with the faithfulness of GAE is traced heading back to the `<perceive>` ... `</perceive>` content rather than its mere presence, indicating that the effect is not an artifact of entropy reduction, response length, or generic reward reweighting.

K Related Work

K.1 Reinforcement Learning for LLMs

Credit assignment over long Chain-of-Thought (CoT) horizons remains a central difficulty in RLVR for LLMs Sahoo et al. (2024). GRPO Guo et al. (2025) circumvents explicit temporal credit assignment by drawing multiple rollouts per prompt and normalizing outcome rewards against the group mean, effectively collapsing the objective to the trajectory level. Subsequent variants pursue this direction along two complementary axes. DAPO Yu et al. (2025) and Dr.GRPO Liu et al. (2025) refine the gradient dynamics of group normalization through mechanisms such as dynamic sampling and bias correction, while GSPO Zheng et al. (2025a) and GMPO Zhao et al. (2025b) re-formulate the reward signal directly at the trajectory level. Despite these differences, all of them share a property that matters for our setting. They treat the trajectory as a homogeneous symbolic stream and assign a group-normalised, trajectory-level advantage to every token, with no notion of perception- or reasoning-specific roles. This assumption is particularly restrictive for video reasoning, where the input contains a substantially higher proportion of visual tokens than in the image or text modality, and perception-selective tokens and reasoning-support tokens play qualitatively different roles in the decision process. PR-FAPO is therefore *orthogonal* rather than competing. It retains the group-normalized advantage on which these methods build, and adds the perception–reasoning fusion-awareness that none of them targets.

A recent line of work Yuan & Xie (2026); Ma et al. (2026); Yao et al. (2026b) incorporates a *suffix future-aware* lens into trajectory-level optimization to strengthen long-horizon reasoning in text-centric models. These methods rely on the clipped importance ratio as the sole mechanism for integrating future-token signals into the current update, and, to the best of our knowledge, thereby inherit its inability to reshape entropy along the trajectory Luo et al. (2026). While effective in the symbolic domain, this direction makes the underlying limitation explicit for video reasoning.

Group-based, trajectory-level objectives offer limited support for *explicit fusion* when the perception–reasoning fusion state itself evolves and must be actively queried.

In video reasoning, spatial and temporal scales are tightly coupled, so fusion planning, namely deciding *what* visual content to attend to and *when* to revisit it, is not an auxiliary skill but a core part of the decision process. Motivated by this observation, we revisit a Markov Decision Process (MDP) perspective Sutton & Barto (2018); Puterman (1994); Kaelbling et al. (1998). Concretely, we motivate to mimic the learner as a *Markov-state* agent whose state is a compact symbolic summary of the current perceptual context and whose action a_h is selected to maximise the expected cumulative reward over future states (suffix states). Applied to the entire sampled trajectory, recasting the group-normalized advantage inside this explicit articulated visual prefix produced by perception policy, MDP enables principled step-wise credit assignment over reasoning policy decisions and yields the discount-controlled fusion planning signal that a pure trajectory-level advantage cannot express, thereby motivating the discount-controlled formulation of PR-FAPO.

K.2 Video Reasoning Models

Recent video RLVR built on GRPO can be organised along four complementary directions. **Reward shaping** approaches augment the outcome signal with video-specific inductive biases, as in the temporal-aware reward of Video-R1 Feng et al. (2025) and the semantic-consistency reward of Video-RFT Wang et al. (2025d), which aligns Chain-of-Thought (CoT) generation with the linguistic outcomes of video descriptions. **Scaling and specialisation** approaches target long or ego-centric streams, with LongVILA-R1 Chen et al. (2025b) pushing scalability through sequence parallelism and Ego-R1 Tian et al. (2025) training a tool-using agent on ego-centric data for long-horizon understanding. **Interleaved video–text reasoning** approaches adopt a multi-turn frame-search paradigm, in which the model first reasons to select salient frames or temporal segments and then revisits them at higher resolution or frame rate to refine its predictions Yan et al. (2025); Fu et al. (2025b); Li et al. (2025a). Video-o3 Zeng et al. (2026) pushes this direction further with native interleaved tool invocation, coupling task-decoupled attention masking with a verifiable trajectory-guided reward so that clue seeking and answer reasoning share a single context. **Task and objective redesign for predictive reasoning** approaches instead rethink *what* the model is trained to predict. NEP Wang et al. (2025a) recasts future video segments as a self-supervised next-event target and fine-tunes the model with an SFT+GRPO pipeline on the resulting dataset. Video-CoE Su et al. (2026) enforces reasoning through timestamped Chain-of-Events, using a similarity-based CoE reward within a GRPO objective that jointly rewards event format, cross-modal consistency, and answer accuracy. A parallel line of work Yao et al. (2026a); Tong et al. (2026) interleaves textual tokens with compact *latent visual spans*, whose latent tokens are optimised with a latent-aware objective, aiming for inject visual imagination directly into the reasoning chain.

Despite this progress, all four directions share the same underlying optimizer, injecting video-specific signal at the *reward*, *data-curation*, or *task-definition* level while leaving the group-normalized, sequence-level GRPO update unchanged. The resulting policy is therefore still driven by a bandit trajectory-level advantage, and the spatial complexity of individual frames and its coupling to temporal dynamics is addressed only indirectly, through auxiliary rewards, tool calls, or task reformulations that encourage object- or event-level learning. Indeed, even the most recent latent-visual-span variants Yao et al. (2026a); Tong et al. (2026), which redesign the objective around latent visual tokens, still rely on teacher-model-based data curation to obtain specialised SFT data and augment the loss with an auxiliary latent-object term, rather than modifying how the advantage is distributed along the trajectory. Additional latent tokens may be unnecessary once the optimiser itself selectively routes visual evidence into reasoning, as their benefit is entangled with a complex SFT curation pipeline and a latent-optimisation objective. Latent-token designs suit future prediction, whereas we target general video question answering, where perceptually grounded credit assignment on the existing token stream is central. What is missing is therefore an analysis of *how* the optimiser allocates credit across perceptual and reasoning tokens. We take up this question along a complementary route, presenting the first trajectory-level probability–entropy analysis of perceptual-state dynamics in video LLMs and building *PR-FAPO* on it.

L PROMPT TEMPLATE

In the main paper, we introduce the system prompt based on previous work VideoAuto-R1 Liu et al. (2026b), which adopts an *answer* → *think* → *answer* format. This prompt design avoids a cold-start stage and facilitates stable training with targeted optimize objective. As shown in Table 3, the model initially generates a perception trace within <perceive>...</perceive> then reasoning trace within <think> </think> tags, followed by the final answer enclosed in \boxed{}. This prompt format aligns with previous R1-style approaches such as Video-R1 Feng et al. (2025), VideoChat-R1 Li et al. (2025c) and VideoAuto-R1Liu et al. (2026b).

Table 3: System Prompt for RL with Thinking.

SYSTEM PROMPT
You are a helpful assistant. FIRST: Output your initial answer inside the first \boxed{} without any analysis or explanations. If you cannot determine the answer without reasoning, output boxed{Let's analyze the problem step by step.} instead. THEN: Think through the perception process and reasoning process as an internal monologue enclosed within <perceive>...</perceive> and <think>...</think>. AT LAST: Output the final answer again inside \boxed{}. If you believe the previous answer was correct, repeat it; otherwise, correct it. Output format: \boxed{} <perceive>...</perceive> <think>...</think>\boxed{}

M TRAINING SETTING

Prior work Liu et al. (2026b) has shown that *cold-start supervised fine-tuning (SFT)* before reinforcement learning (RL), when using low-quality CoT supervision, can distort the behavior of an otherwise strong base model. Such distortion often leads to a suboptimal initialization for subsequent RL, resulting in high training cost with limited performance gain. Motivated by these findings, we avoid cold-start SFT altogether and directly apply RL, aiming to reduce training cost while preserving the benefits of policy optimization.

Another line of prior work Liu et al. (2026b); Zhong et al. (2025) shows that incorporating multimodal data—including text, images, and videos—during training is effective for improving video reasoning performance. Following this approach, we adopt the same training datasets used in Liu et al. (2026b), which combine text-, image-, and video-based supervision. Specifically, the dataset consists of 6.4K text samples Yu et al. (2025), 27.5K image samples Wang et al. (2025b;f), and 49.4K video samples Feng et al. (2025); Cores et al. (2024); Li et al. (2025d); Zhu et al. (2025); Wang et al. (2025g); Gao et al. (2017); Caba Heilbron et al. (2015); Xiao et al. (2024), all of which are used jointly during training. For a fair comparison with baseline, we follow the same reward design and prompt template can be found in Appendix C Prompt Template of VideoAuto-R1 Liu et al. (2026b).

N PARAMETER SETTINGS

We detail the specific hyperparameter configurations employed for fine-tuning Qwen3-VL-8B. To ensure full reproducibility, we will release our complete codebase along with the training scripts used in our experiments. We initially validated the effectiveness of our approach on Qwen3 VL 8B. During this preliminary phase, we observed that training performance was initially volatile, and the reproducibility of reasoning gains was inconsistent across independent runs. To address these stability issues, we keep the group size to $G = 16$ to provide a more stable optimization signal. These adjustments successfully stabilized the training trajectory, making the results more reliable while maintaining the same performance level. The model was optimized using a learning rate of 1×10^{-6} and a weight decay of 0.1, with the Ψ set to 0.96. A summary of the hyperparameter settings is provided in Table 4. The training Qwen3-VL 8B is conducted on 8 RTX PRO 6000 GPUs for approximately 6.5 days.

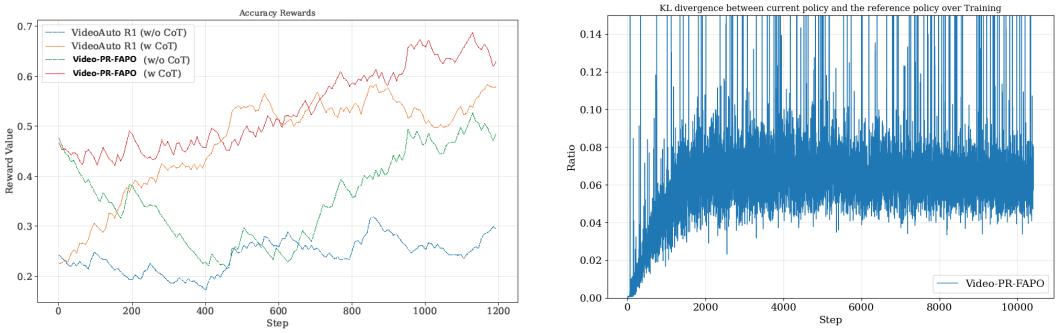


Figure 7: We follow VideoAuto R1 training framework, which adopts an (direct answer → think (CoT) → final answer) format. The dotted line represents direct response accuracy, while the normal line represents response accuracy using CoT. As shown on the left, PR-FAPO exhibits a stable evolution of policy KL divergence shown moves away from its previous policy state, it preserves the stability properties observed in the reference policy.

Table 4: Hyperparameter settings for Qwen3-VL experiments. We compare the configuration of our proposed PR-FAPO against the GRPO baseline. Most infrastructure and optimization settings remain identical to ensure a fair comparison.

Hyperparameter	**Video PR-FAPO**
Settings	
Base Model	Qwen3-VL-8B, Qwen2.5-VL-7B
Engine	DeepSpeed ZeRO-2
Global Batch Size	256
Learning Rate	1e-6
Weight Decay	0.1
Max Prompt Length	8192
Sampling Temp / Top-p	1.0 / 1.0
Policy Clip Ratio	[0.2, 0.2] (Asymmetric)
KL Penalty Coef	0.01
Group Size (G)	16
Max Response Length	2048
Check point frequency	every 100 steps
Follow VideoAuto-R1 Settings	
initial answer weight w_1	0.9
reviewed answer weight w_2	1.1
Method-Specific Settings	
Ψ	0.96
Geometric Family τ	$\tau = 4^k, k_{\max} = \lfloor \log_4 \lvert y_i \rvert \rfloor$

O INFERENCING

At test time, we follow the inference protocol of VideoAuto-R1, which employs a confidence-based early-exit mechanism to determine whether generation should terminate after producing an initial direct answer or proceed to a full CoT rationale followed by a reviewed answer. This strategy is designed to avoid unnecessary computation while maintaining reasoning reliability. Concretely, it consists of three steps: (1) generating an initial answer, (2) computing a confidence score for this answer, and (3) deciding whether to exit early or continue with full reasoning. For fair comparison, we adopt the same early-exit confidence threshold as VideoAuto-R1 and set it to 0.98. We refer readers to Appendix E (Inference Strategy) of VideoAuto-R1 Liu et al. (2026b) for further implementation details.

Table 5: **Evaluation Results on Image- and Text-Centric Reasoning Benchmarks.**

Model	Image-centric Reasoning								Text-centric Reasoning	
	Geo3k test	**DynaMath** test	**WeMath** testmini	**MathVista** testmini	**MathVision** testmini	**MathVerse** testmini	**MMMU** val	**MMMU-Pro** test	**AIME24** avg@8	**AIME25** avg@8
VideoAuto-R1 8B	74.0	62.3	72.6	72.8	39.5	55.3	23.9	44.3	35.8	23.7
Video-PR-FAPO 8B	80.7	59.4	70.1	71.0	40.5	55.9	25.2	47.3	32.9	23.3
Δ	+6.7	-2.9	-2.5	-1.8	+1.0	+0.6	+1.3	+3.0	-2.9	-0.4

P Evaluation Configuration

Following VideoAuto-R1 Liu et al. (2026b), for reasoning-centric benchmarks, including MMVU Zhao et al. (2025a), Video-MMMU Hu et al. (2025), and MVP Krojer et al. (2025), we evaluate each sample using 64 to 2048 input frames and allow up to 4096 generated tokens. This setting encourages the policy to focus on CoT reasoning rather than recognizing subtle motion changes in the input videos. For perception- and temporal-grounding-centric benchmarks, including Video-MME (without subtitles) Fu et al. (2025a), MVBench Li et al. (2024), LongVideoBench Wu et al. (2024), Charades-STA Gao et al. (2017), ActivityNet Caba Heilbron et al. (2015), and NExT-GQA Xiao et al. (2024), we adopt the same evaluation setting, using 64 to 2048 input frames and allowing up to 4096 generated tokens. Both style of benchmarks are performed using lmms-eval Zhang et al. (2024a) with deterministic decoding: temperature 0, top-p 1, beam size 1, and sampling disabled.

Q Evaluation on Image and Text Benchmarks.

Additionally, we evaluate our model on image reasoning benchmarks, such as Geo3k Lu et al. (2021),DynaMath Zou et al. (2024), We-Math Qiao et al. (2024), MathVista Lu et al. (2023), MathVision Wang et al. (2024), MathVerse Zhang et al. (2024b), MMMU Yue et al. (2024), MMMU-Pro Yue et al. (2025). As shown in Table 5 (left side), Video-PR-FAPO consistently outperforms VideoAuto-R1 across most image reasoning benchmarks. In particular, the largest improvements are observed on Geo3K and MMMU-Pro. We attribute these gains to the incorporation of image-centric mathematical reasoning data during training, which enhances the model's capability to perform complex visual reasoning and improves its ability to effectively leverage visual information for reasoning beyond the video domain only.

As shown in Table 5 (right side), Video-PR-FAPO achieves lower scores than VideoAuto-R1 on both AIME24 and AIME25. A plausible explanation is that PR-FAPO places a stronger emphasis on video-related reasoning during training, which may lead to a modest degradation in text-centric reasoning performance. Furthermore, prior observations as Appendix A suggest that text-based mathematical reasoning tends to benefit from more deterministic (i.e., lower-entropy) policies, whereas PR-FAPO is designed to trace from a higher level of policy entropy to lower. These factors together may explain the inferior performance of Video-PR-FAPO on text-only benchmarks. In addition, the CoT tracing mechanism introduced in PR-FAPO is specifically motivated by the requirements of video reasoning, where temporal and visual evidence must be integrated throughout the reasoning process. Although the model is trained on included text data, the inductive bias induced by this tracing mechanism is specifically tailored to multimodal reasoning and may not transfer optimally to purely textual tasks. Consequently, this architectural bias could further contribute to the observed performance gap on text-only benchmarks.

R Ablation Studies

The result in other backbone and SFT-initialized model. As shown in Table 6, we further evaluate our method, trained on the Qwen2.5-VL 7B backbone, against several representative approaches on both video perception-centric and reasoning-centric benchmarks. The results show that our method remains competitive with VideoAuto-R1 on perception-centric tasks while achieving superior performance on reasoning-centric benchmarks. Notably, Video-PR-FAPO outperforms all compared open-weight baselines in reasoning-oriented evaluations. These findings suggest that our approach enhances video reasoning capability without substantially sacrificing video perception performance, despite the differences between the Qwen2.5-VL and Qwen3-VL architectures.

Table 6: **Evaluation Results on Video QA Benchmarks.** We compare Video PR-FAPO with both perception-oriented and reasoning-oriented benchmarks.

Model	Perception Centric Benchmark			Reasoning Centric Benchmark		
	VideoMME	MVBench	LongVideoBench	MMVU	VideoMMMU	MVP
GPT-5-Nano OpenAI (2025b)	66.2	–	–	63.1	63.0	–
Gemini-2.5 Google DeepMind (2025)	72.7	–	60.9	65.3	69.2	–
Temporal-RLT Li et al. (2025a)	57.6	68.1	–	65.0	–	–
Video-RFT Wang et al. (2025d)	59.8	62.1	–	68.5	51.1	–
Video-R1 Feng et al. (2025)	61.8	65.5	–	65.0	51.4	33.0
Video-RTS Wang et al. (2025i)	63.0	–	56.6	66.4	52.7	–
VITAL Zhang et al. (2025)	64.1	–	–	68.7	54.2	–
LongVILA-R1 Chen et al. (2025b)	65.1	67.6	58.0	–	–	–
LOVE-R1 Fu et al. (2025b)	66.2	66.6	60.1	–	–	–
VideoChat-R1.5 Yan et al. (2025)	65.2	70.6	61.4	–	49.6	38.6
Qwen2.5-VL 7B Team (2024)	66.0	67.1	60.9	66.2	54.7	36.5
VideoAuto-R1 Liu et al. (2026b)	67.3 (+1.3)	71.0 (+3.9)	60.5 (-0.3)	69.7 (+3.5)	58.6 (+3.9)	39.4 (+2.9)
Video-PR-FAPO (Ours)	68.0 (+2.0)	69.5 (+2.4)	60.5 (-0.3)	71.8 (+5.6)	61.4 (+6.7)	40.5 (+4.0)
Qwen3-VL 8B Bai et al. (2025)	72.5	69.4	67.6	69.9	61.0	40.5
VideoAuto-R1 Liu et al. (2026b)	71.7 (-0.8)	**72.0** (+2.6)	67.4 (-0.2)	71.1 (+1.2)	65.0 (+4.0)	43.0 (+2.5)
Video-PR-FAPO (Ours)	**72.5** (±0)	71.4 (+2.0)	67.1 (-0.5)	**72.2** (+2.2)	**67.6** (+6.6)	**43.7** (+3.2)

Table 7: **Evaluation Results on Temporal Grounding Benchmarks.** For a fair comparison with Video-PR-FAPO (SFT+RL), we adopt the TimeLens2-8B SFT-initialized model as the starting point for our PR-FAPO reinforcement learning stage.

Model	Charades-STA				ActivityNet				NeXT-GQA	
	0.3	0.5	0.7	mIoU	0.3	0.5	0.7	mIoU	Acc	mIoU
Qwen3.5-397B QwenTeam (2026)	69.3	47.8	24.9	47.8	69.0	57.5	39.3	52.2	–	–
TimeChat Ren et al. (2024)	51.0	27.5	11.4	31.2	44.0	27.8	14.3	30.4	28.8	17.4
TimeSuite Zeng et al. (2024)	79.4	67.1	43.0	–	–	–	–	–	–	–
TimeMarker Chen et al. (2024)	73.5	51.9	26.9	48.4	67.4	50.7	33.0	49.5	–	–
Temporal-RLT Li et al. (2025a)	79.6	67.9	44.1	57.0	56.9	38.4	20.2	39.0	78.7	37.3
Time-R1 Wang et al. (2025g)	82.8	72.2	50.1	58.8	73.3	55.6	34.0	52.1	–	–
VITAL Zhang et al. (2025)	83.1	72.0	46.7	59.9	70.9	50.8	31.6	49.8	78.7	43.0
VideoChat-R1.5 Yan et al. (2025)	82.8	71.6	48.3	60.6	52.4	32.3	16.8	35.3	–	–
TimeLens-8B (SFT+RL) Zhang et al. (2026)	76.6	63.0	35.2	55.2	68.9	58.4	40.6	53.2	-	-
TimeLens2-8B (SFT+RL) Zhu et al. (2026b)	80.9	68.0	38.9	58.6	74.0	65.6	46.6	58.6	81.4	40.5
Video-PR-FAPO (SFT+RL)	82.7	66.0	42.7	58.8	**74.7**	**66.6**	**48.2**	**59.3**	**81.9**	42.6
Δ (TimeLens2 8B)	(+1.8)	(-2.0)	(+3.8)	(+0.2)	(+0.7)	(+1.0)	(+1.6)	(+0.7)	(+0.5)	(+2.1)
Qwen2.5-VL 7B Team (2024)	77.7	59.6	34.8	52.9	37.9	22.6	10.6	26.9	53.3	20.2
VideoAuto-R1 (RL) Liu et al. (2026b)	82.9	70.8	46.0	60.0	69.2	48.5	27.3	47.6	80.6	36.7
Video-PR-FAPO (RL)	82.6	71.2	48.2	60.4	68.3	49.3	27.6	47.9	81.3	37.3
Δ (VideoAuto-R1 7B)	(-0.3)	(+0.4)	(+2.2)	(+0.4)	(-0.9)	(+0.8)	(+0.3)	(+0.2)	(+0.7)	(+0.6)
Qwen3-VL 8B Bai et al. (2025)	68.5	49.3	26.2	47.2	63.8	51.1	35.0	48.0	–	–
VideoAuto-R1 (RL) Liu et al. (2026b)	**85.1**	74.9	53.7	63.7	74.1	54.3	32.4	51.9	81.1	44.2
Video-PR-FAPO (RL)	84.6	**75.0**	**55.4**	**64.0**	74.1	53.8	32.9	51.9	81.5	**44.2**
Δ (VideoAuto-R1 8B)	(-0.5)	(+0.1)	(+1.7)	(+0.3)	(+0.0)	(-0.5)	(+0.5)	(+0.0)	(+0.4)	(+0.0)

We further validate PR-FAPO on the TimeLens2-8B Zhu et al. (2026b) SFT-initialized model. TimeLens2-8B adopts a two-stage training paradigm consisting of large-scale temporal grounding SFT followed by reinforcement learning with the Temporal Wasserstein Reward. To ensure a fair comparison, we use the same TimeLens2-8B SFT checkpoint as the initialization for our RL training stage. Table 7 shows that PR-FAPO consistently surpasses the original TimeLens2-8B across temporal grounding benchmarks. The largest gains are achieved under the stricter R1@0.7 criterion, suggesting that our approach leads to more precise temporal boundary localization. These findings demonstrate that PR-FAPO can effectively leverage temporal information by deeper CoT reasoning is optimized during RL stage, even when starting from SFT-initialized model.

Reviewed (w CoT) vs. Initial (w/o CoT) Reward. Following the VideoAuto-R1 framework, the model is trained to produce an initial answer, perform explicit reasoning, and then output a reviewed answer. Both the initial and reviewed answers are supervised using verifiable rewards, with a larger weight assigned to the reviewed answer to encourage refinement or confirmation of the initial prediction. Importantly, this training paradigm eliminates the need for manual *think/no-think* annotations; instead, the model learns to make *both* answers correct through end-to-end optimization. As shown in Figure 7 (left side), compared to VideoAuto-R1, our Video-PR-FAPO framework exhibits

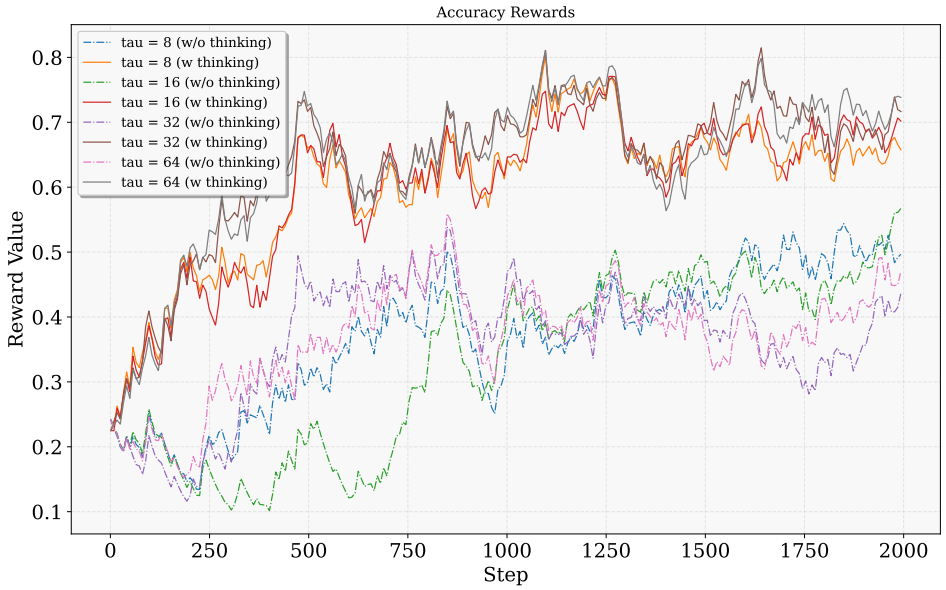


Figure 8: The Study on Window size τ during training dynamic.

significantly faster convergence. The reward signals for both the initial answer and the reviewed answer consistently improve throughout training. Notably, the performance of the initial answer improves substantially, indicating that the policy increasingly favors *perception-driven strategic decision-making* rather than over-reliance on subsequent reasoning steps. This suggests that Video-PR-FAPO more effectively aligns perceptual learning with downstream reasoning, enabling the model to form stronger early judgments that require less corrective reasoning later. As shown in Figure 7 (right side), In contrast to baseline methods, this KL growth is gradual rather than abrupt, indicating stable learning dynamics. Such controlled policy drift suggests that PR-FAPO avoids overly aggressive exploration while still enabling meaningful behavioral adaptation, contributing to its faster convergence within stability.

The Study on Window τ. We fix the value-function TD decay rate to $\Psi = 0.96$ and investigate the effect of the window size τ on training dynamics. For the *initial reward*, we observe that smaller window sizes ($\tau \in \{8, 16\}$) lead to faster convergence, whereas larger windows ($\tau \in \{32, 64\}$) converge more slowly. In contrast, for the *final reward*—where answers are justified with CoT, the trend reverses. Smaller window sizes ($\tau \in \{8, 16\}$) result in lower and slower convergence, while larger windows ($\tau \in \{32, 64\}$) exhibit faster and higher convergence. These results suggest that larger τ better supports long-horizon reasoning required for CoT-based reward optimization.

The Study on value-function TD Ψ. We keep the window size to $\tau = 16$ and investigate the effect of the TD decay rate Ψ on training dynamics. With a value-function TD formulation, we observe fast and stable reward convergence when setting $\Psi = 0.96$. When using a larger decay rate $\Psi = 0.99$, the policy converges rapidly toward reasoning-dominated behavior: the reward associated with CoT reasoning reaches the highest level, while the reward without CoT remains the low convergence. This suggests that larger Ψ places greater emphasis on long-horizon reasoning trajectories. In contrast, with a smaller decay rate $\Psi = 0.92$, the policy converges quickly toward perception-dominated behavior. In this setting, the reward without CoT rises early, whereas the CoT-based reward exhibits slower and lower convergence. These results indicate that smaller Ψ biases the optimization toward short-horizon perceptual signals.

Training Dynamic. To better understand the behavior of Video-PR-FAPO, we visualize its training dynamics from three complementary views: the proportion of positive and negative groups (Fig. 11b), the trajectory properties within positive groups, i.e., those whose group mean reward exceeds 0.5 (Fig. 11a), and those within negative groups, i.e., whose group mean reward is at most 0.5 (Fig. 11c). We highlight three observations below. In the early stage of training, the proportion of positive groups rises sharply, after which it continues to increase at a slower but steady rate until convergence. This indicates that the policy quickly acquires the easier portion of the reward signal and thereafter

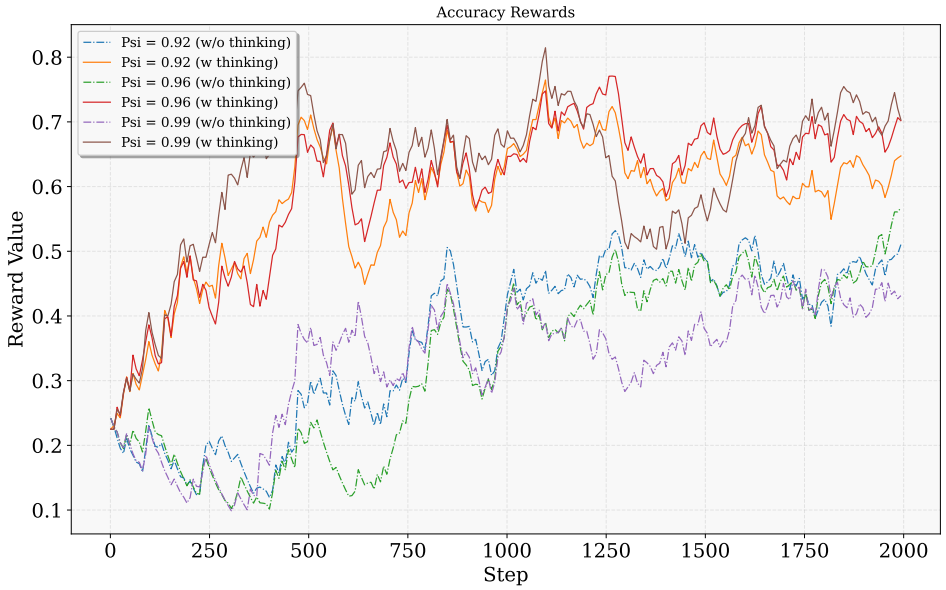


Figure 9: The Study on TD decay rate Ψ during training dynamic.

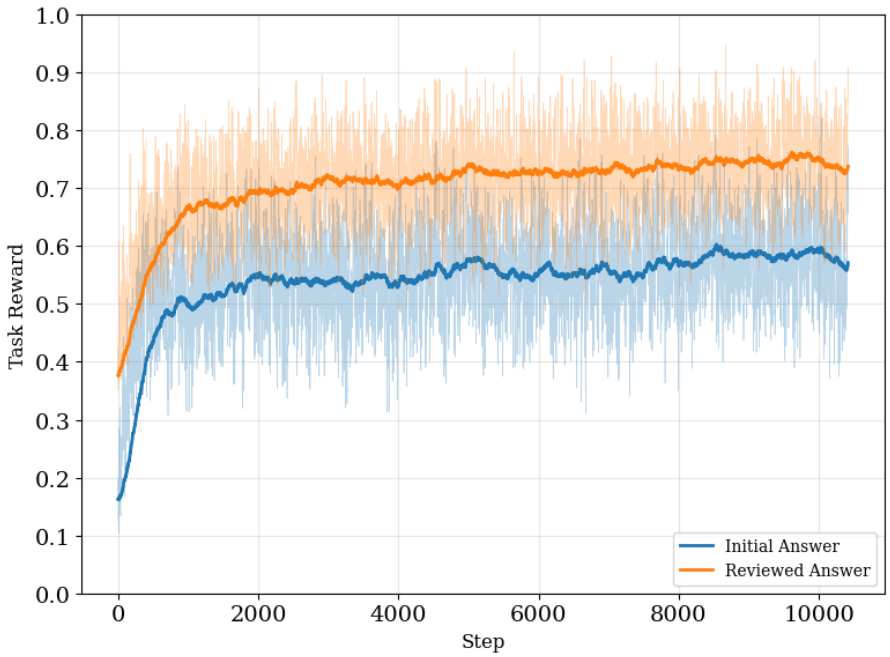


Figure 10: Training reward of Video-PR-FAPO 8B. We show the average task reward for both initial and reviewed answers during training.

improves incrementally. Within positive groups, correct CoTs exhibit consistently higher trajectory entropy than incorrect ones. This suggests that, once a query is already largely solvable, broader exploration in the reasoning policy remains beneficial, while insufficient exploration of the visual evidence is the factor that limits further gains. Within negative groups, the trend is reversed within negative groups: correct CoTs show lower trajectory entropy than incorrect ones. A plausible reading is that, when the query is not yet visually grounded, high entropy reflects an unfocused search over visual evidence rather than productive exploration, while the few correct trajectories are those that commit early to the relevant evidence. Probability–entropy correlation within positive

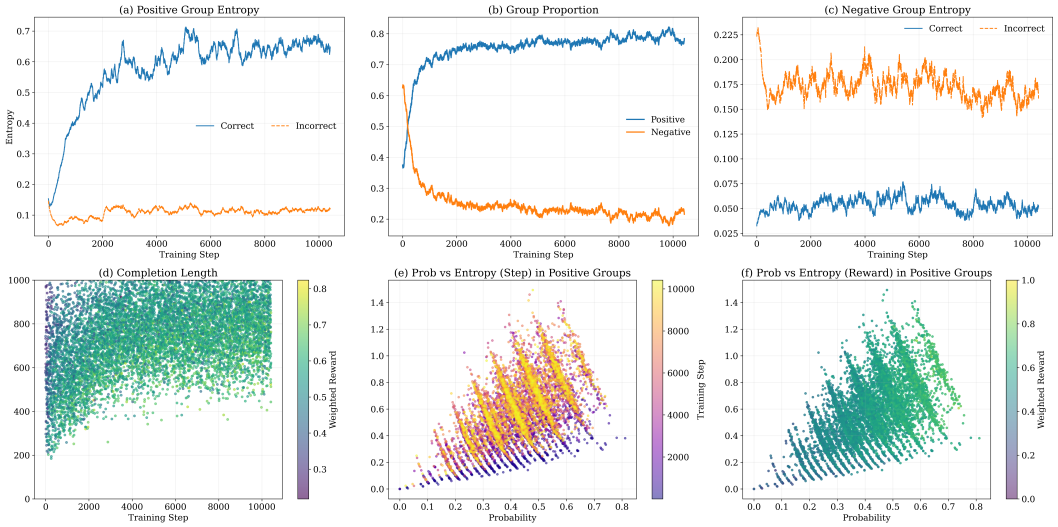


Figure 11: Training dynamics of Video-PR-FAPO 8B on high-performing rollout groups which group mean reward > 0.5 defined as positive group, and ≤ 0.5 defined as negative group.

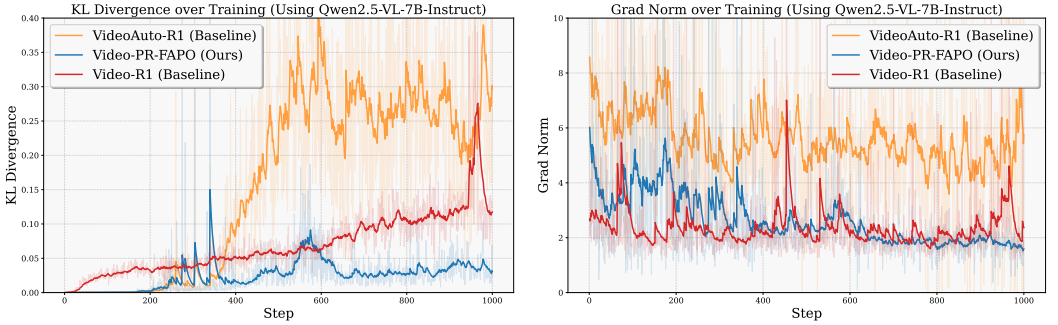


Figure 12: Training dynamics of preference KL divergence (important ratio) and gradient norm, comparing the stability of VideoAuto-R1, Video-R1 and Video PR-FAPO.

groups. We further examine the joint behavior of trajectory probability and entropy within positive groups (Fig. 11e). The pattern mirrors that of Video-R1 (Fig. 1): the entropy–probability bands shift progressively over the course of training, with later rollouts attaining a more favorable balance between trajectory probability and entropy, moving out of the low-entropy region. Comparing across these bands further shows that bands occupying the higher-probability and higher-entropy ranges are associated with higher group-mean rewards (Fig. 11f), and that shorter responses tend to receive higher rewards (Fig. 11d). Taken together, these correlational observations are consistent with training exerting an direction CoT toward analysed tight upperbound and toward shorter correct trajectories.

Stability. In RL, Proximal Policy Optimization (PPO) employs importance-ratio clipping to constrain policy updates and prevent overly aggressive policy changes Schulman et al. (2017). Compared with Trust Region Policy Optimization (TRPO) Schulman et al. (2015), PPO provides a practical surrogate objective that improves training stability while maintaining the benefits of trust-region-style optimization. Within GRPO, importance-ratio clipping serves as a key stabilization mechanism. By limiting policy updates when the current policy deviates excessively from the rollout policy, clipping suppresses large policy shifts and helps maintain controlled optimization throughout training. Ideally, the importance ratio remains centered around one, such that positive and negative policy updates are constrained symmetrically. Under this condition, clipping effectively truncates updates from overconfident samples and preserves balanced policy learning. However, our empirical analysis suggests that this mechanism does not function as intended in the VideoAuto-R1 baseline. As shown in the left panel of Figure 12, the importance-ratio distribution exhibits a systematic bias, with its mean progressively drifting beyond the clipping region $1 \pm \epsilon$. Moreover, after approximately 450 training steps, the variance increases substantially and exceeds 0.2, indicating that a large fraction of samples fall outside the effective trust region Jia et al. (2026). This behavior suggests increasingly

Table 8: In w/o Multi-scale window case, we set tau=32 and Psi=0.96 for single window setting. In prefix suffix anchor case, we set $\rho_{i,t}(y) = \pi_\theta(y \mid z_m, s_{i,t})$ and $\mathcal{U}_{i,t} = \mathcal{H}(\pi_\theta(\cdot \mid z_m, s_{i,t}))$. In w/o Discounting Strategy as segment-level see Table 9 for details

Method	VideoMME	MVBench	MMVU	**Charades-STA** (mIoU)
Video-PR-FAPO 8B	**72.5**	**71.4**	**72.2**	**64.0**
Role-aware Strategy				
w/o Prefix Anchor	71.1	69.7	71.6	62.1
w/o Suffix Anchor	72.0	68.4	72.0	63.4
w/o Prefix Suffix Anchor	70.3	68.0	65.4	60.8
Position-aware Strategy				
w/o Discounting Strategy	67.1	65.0	64.0	59.9
w/o Multi-scale Windows + w/o segment-level	66.7	63.9	62.7	57.4

unstable policy optimization and reduced effectiveness of the clipping mechanism. To further examine optimization stability, we additionally analyze the gradient norm, which is widely used as an indicator of training stability Jia et al. (2026). As illustrated in the right panel of Figure 12, Video PR-FAPO exhibits substantially smoother gradient dynamics and avoids the large fluctuations observed in VideoAuto-R1, indicating improved optimization stability and convergence behavior. Taken together, the systematic shift and growing variance of the importance-ratio distribution lead to increasingly unbalanced policy updates, which are accompanied by elevated gradient fluctuations. We hypothesize that this instability weakens the effectiveness of GRPO's clipping-based regularization and contributes to the implicit reward-hacking behavior Wang et al. (2026) observed in the VideoAuto-R1 baseline.

R.1 Attribution

Attribution of Role-aware and Position-aware. As shown in Table 8, all variants were trained for the same 2,000 optimization steps. Removing the Prefix Anchor causes a larger performance degradation than removing the Suffix Anchor across most benchmarks. This finding suggests that the Prefix Anchor is more effective for assigning credit to the anchor state z_m. When rewards cannot be effectively propagated to z_m, performance consistently deteriorates across perception, reasoning, and temporal grounding tasks. Moreover, removing both Prefix and Suffix Anchors leads to a further decline in performance, highlighting the complementary benefits of the two mechanisms. Regarding the position-aware strategy, furthering remove the Discounting Strategy while preserving Multi-scale Windows substantially reduces performance, demonstrating the effectiveness of position-aware reward weighting. By comparison, eliminating Multi-scale Windows in addition to the Discounting Strategy results in only a modest further decrease. These results suggest that the Discounting Strategy is the primary contributor to the gains of the position-aware design, while Multi-scale Windows provide additional benefits.

The study on discounting strategy. Starting from the baseline model, all variants were trained for one epoch, and the results are summarized in Table 9. We first investigate a token-level formulation, where $f_{i,t} = 1 - \tanh(\rho_{i,t}^2 - e^{\mathcal{U}_{i,t}})$ is directly multiplied with the advantage $\hat{A}_i$. This strategy leads to a substantial performance degradation across all benchmarks. Although the analysis provides an upper bound on the contribution of individual states $z_{i,t}$, it cannot effectively propagate credit to the anchor state z_m. As a result, the resulting credit assignment remains coarse and less effective for optimizing long-range dependencies. Motivated by this observation, we elevate the weighting mechanism from the token level to the segment level. As shown in Table 9, the segment-level formulation yields noticeable performance improvements. These results further support our hypothesis that effective credit propagation toward the anchor state z_m is critical for learning. However, as discussed in our analysis, reducing the fusion gap between credit assignment requires an explicit trade-off mechanism. This motivates the introduction of a discounting strategy, where the advantage is estimated from temporal-difference (TD) errors using generalized advantage estimation (GAE). Incorporating discounting consistently improves performance across all tasks. Furthermore, discounting becomes increasingly important in long-horizon scenarios. Based on the analysis in Section G, we introduce

Table 9: Ablation study of different discounting and segment-level strategies. In w/o Discounting (segment-level) case, we adopt $V^p_{i,\tau,s} = 1 - \tanh(\cdot)$ before token to segment-level normalization.

Strategy	MVBench	MMVU	**Charades-STA** (mIoU)
Qwen3-VL 8B	**69.4**	**69.9**	**47.2**
w/o Discounting + w token-level shaping			
$f_{i,t} = 1 - \tanh(\rho^2_{i,t} - e^{\mathcal{U}_{i,t}})$	65.3	64.5	57.4
w/o Discounting + w segment-level shaping			
$A_{i,\tau,s} = V^p_{i,\tau,s} \in (0, 1]$	66.0	70.0	60.9
w Discounting $(A^{\text{local}}_{i,\tau,s}(\kappa) = \text{GAE}^P_{i,\tau,s} \in (0, 1])$			
$\gamma_{i,\tau,s} = 2^{-V^p_{i,\tau,s}} \in (0, 1]$	68.1	70.6	61.7
$\gamma_{i,\tau,s} = 2^{-V^p_{i,\tau,s}/\lvert y_i \rvert} \in (0, 1]$	71.9	72.2	62.5
w Discounting + Local exponent			
$(A^{\text{local}}_{i,\tau,s}(\kappa) = (\gamma_{i,\tau,s})^{e_{\tau,s}(\kappa)}\, \text{GAE}^P_{i,\tau,s} \in (0, 1])$	71.4	72.2	64.0

Table 10: Top-5 and Bottom-5 MVBench subtasks.

Top 5		**Bottom 5**	
Task	Acc.	Task	Acc.
Moving Attribute	94.0	Object Shuffle	41.2
Scene Transition	93.5	Egocentric Navigation	41.5
Object Existence	92.0	Fine-grained Action	46.0
Unexpected Action	85.5	Moving Direction	59.5
Character Order	85.0	Episodic Reasoning	60.0

length normalization into the discounting factor to account for varying sequence lengths. The results show that length-normalized discounting further improves performance, indicating that discounting should be adjusted according to the effective temporal horizon. To further refine credit assignment within each segment, we introduce a local exponent mechanism based on a convex combination of forward and backward distances to segment boundaries, together with the Multi-scale Window design (Table 8). Incorporating the local exponent yields additional gains on temporal grounding, improving Charades-STA, a slight performance decrease is observed on perception-oriented benchmarks such as MVBench. These results suggest a trade-off between fine-grained credit assignment and temporal perception performance, which remains an important direction for future investigation.

S Qualitative Examples

In this part, we first present representative failure cases of VideoAuto-R1 Liu et al. (2026b). In Figure 13, the policy generates a seemingly reasonable chain of thought (CoT); however, its reasoning is grounded in hallucinated visual evidence. Specifically, it synthesizes incorrect visual information when answering a visual context summarization question, suggesting limited ability to effectively utilize video context information. In Figure 14, the policy again produces a plausible CoT, but the reasoning process is overly simplistic and insufficient to support correct decision-making. This behavior suggests shallow reasoning despite access to rich visual information from the video input and under reasoning policy optimization. In Figure 15, the policy becomes stuck during the final decision-making stage. This failure appears to originate from insufficient visual information extracted during perception, which subsequently limits deeper exploration by the reasoning policy. Overall, these examples suggest that VideoAuto-R1 struggles to effectively leverage the abundant visual information available in video inputs. Although the model can generate coherent CoT traces, its reasoning process often fails to exploit visual evidence for confident and accurate final answer generation, motiving us work.

S.1 Qualitative Examples of Failure Case

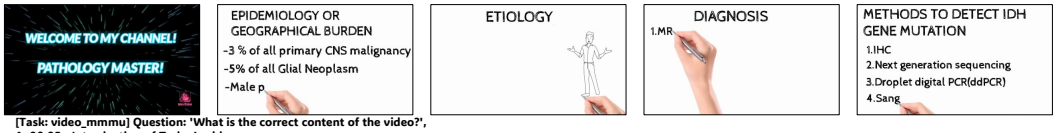
Next, we present two representative failure cases of Video-PR-FAPO. In the first case (Fig. 17), the video contains subtle visual cues, where human actions are highly similar and the motion of light changes rapidly. Although the model generates a seemingly plausible step-by-step rationale, its fail may come from the perception policy does not capture the critical visual information related to the light-motion change. This type of error typically originates from query-relevant perception failures rather than deficiencies in the reasoning process itself. In the second case (Fig. 16), the perception policy successfully captures the relevant visual evidence and articulated visually scientific phenomenon in language, but the reasoning policy fails to correctly interpret the visually scientific phenomenon. This suggests that the reasoning policy is unable to retrieve the necessary knowledge or lacks the domain-specific understanding acquired during pre-training.

Next, we showcase examples from perception-intensive QA benchmarks. Compared with standard perception-oriented tasks, the generated reasoning traces are substantially longer and involve more fine-grained reasoning steps (Fig. 19). Figure 18 demonstrates that the reasoning policy can adaptively generate concise yet effective reasoning trajectories without introducing redundant CoT steps, even for long-horizon problems. In contrast, Fig. 20 presents a case where the reasoning policy produces redundant CoT reasoning. The model repeatedly revisits previously established evidence to increase confidence in its final answer, resulting in a longer-than-necessary reasoning trajectory.

Notably, the model first produces a structured and correct intermediate conclusion, which subsequently guides the reasoning process and encourages the exploration of a confidence-building trajectory. This behavior enables our framework to effectively perform perception planning and maintain coherent reasoning over extended horizons.

S.2 Failure Case Analysis on MVBench

To better understand the performance degradation observed on MVBench in the main paper, we further analyze the model's performance across individual subtasks. As shown in Table 10, the model achieves its highest accuracy on Moving Attribute (94.0%), Scene Transition (93.5%), and Object Existence (92.0%), demonstrating strong capabilities in recognizing visual attributes, object presence, and scene-level changes. In contrast, the lowest scores are observed on Object Shuffle (41.2%), Egocentric Navigation (41.5%), and Fine-grained Action (46.0%). These tasks require accurate tracking of object states, reasoning from first-person perspectives, and distinguishing subtle action differences, which remain challenging for the model. Overall, the results suggest that the model is more effective at capturing high-level visual semantics and global scene transitions than performing fine-grained spatial-oriented and state tracking.

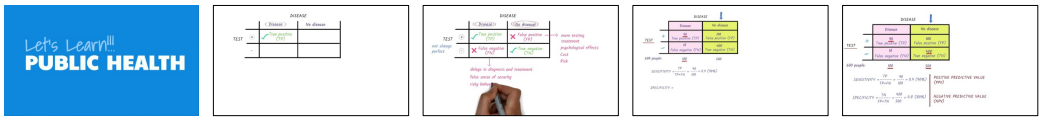


[Task: video_mmmu] Question: 'What is the correct content of the video?',
A. 00:05 - Introduction of Today's video
B. 00:21 - Definition of Oligodendroglioma
C. 00:42 - Geography of oligodendroglioma
...
Z.13:18 - Treatment of oligodendroglioma
Ground Truth: A

VideoAuto-R1 :
¥¥boxed{C}<think>Based on the video content, the correct sequence of topics is as follows:- 00:05: Introduction of Today's video- 00:21: Definition of Oligodendroglioma¥n- 00:42: Geography of oligodendroglioma (Epidemiology or Geographical Burden)- 01:20: Site of involvement- 01:57: Causes of oligodendroglioma (ETIOLOGY)- 02:15: Clinical Presentation- 02:52: Diagnostic methods- 04:41: Gross Morphology- 05:18: Histopathology of oligodendroglioma- 07:33: Demonstration in Histopathology slide- 08:28: IHC of oligodendroglioma (Note: The video shows IHC starting at 10:28, but the option C has 08:28, which is likely a typo in the option. However, the content is correct.)- 12:32: Prognosis of oligodendroglioma- 13:18: Treatment of oligodendroglioma. Option C matches this sequence correctly, despite the minor discrepancy in the IHC timing. The other options have incorrect orderings or misrepresent the content. Therefore, the correct answer is C.</think>¥¥boxed{C}

Video-PR-FAPO :
¥¥boxed{A}<perceive>The video appears to be a structured educational presentation on "Oligodendroglioma," covering its definition, epidemiology, site of tumor, clinical features, diagnosis, microscopic appearance, immunohistochemistry (IHC), grading, prognosis, and treatment. The timestamps provided correspond to key sections of this presentation. The question asks for the correct sequence of content based on the given options. We need to match each timestamp with the corresponding section title from the options.- 00:05: Introduction of today's video → matches option A.- 00:21: Definition of Oligodendroglioma → matches all options.- 00:42: Geography of oligodendroglioma → only option A correctly lists "Geography" here.- 01:20: Site of involvement → matches all options except F, which incorrectly places histopathology at 01:20.- 01:57: Causes of oligodendroglioma → not explicitly mentioned in the video but implied under "epidemiology." Option A includes it.- 02:15: Clinical Presentation → matches all options.- 02:52: Diagnostic methods → matches all options.- 04:41: Gross Morphology → matches all options.- 05:18: Histopathology of oligodendroglioma → matches all options.- 07:33: Demonstration in Histopathology slide → matches all options.- 10:28: IHC of oligodendroglioma → matches all options.- 12:32: Prognosis of oligodendroglioma → matches all options.- 13:18: Treatment of oligodendroglioma → matches all options.
Option A is the most accurate because it correctly orders "Geography" after "Definition," placing "Site of involvement" after "Geography," and correctly sequencing other elements like "Histopathology" after "Diagnostic Methods."
Other options misplace some critical segments — such as putting "Histopathology" before "Site of Involvement" (option F) or listing "Prognosis" before "IHC" (option E). Only option A maintains logical progression throughout. </perceive>
<think>After reviewing the structure of the video and comparing it against the options, I can confirm that option A accurately reflects the order of topics presented. It correctly sequences the introduction, definition, geography, site of involvement, causes, clinical features, diagnostic methods, gross morphology, histopathology, demonstration slides, IHC, prognosis, and treatment. Other options either misorder these sections or include incorrect placements, making them inaccurate representations of the video's content.</think>¥¥boxed{A}

Figure 13: Failure case of VideoAuto-R1. It synthesizes incorrect visual information when answering a visual context summarization question.

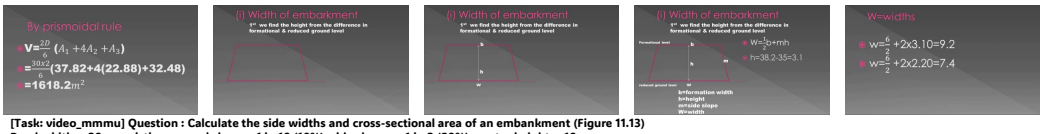


[Task: video_mmmu] Question: 2x2 Contingency Table
| Disease | No disease
+ | a | b
- | c | d
Calculate: sensitivity = α , specificity = β , PPV = γ , NPV = δ
A. α = a/(a+c), β = d/(b+d), γ = a/(a+b), δ = d/(c+d)
....
J. α = a/(a+c), β = d/(c+d), γ = a/(a+c), δ = d/(c+d)
Ground Truth: A

VideoAuto-R1 :
¥¥boxed{C}<think> Based on the video content, sensitivity is calculated as TP / (TP + FN), which corresponds to a / (a + c). Specificity is calculated as TN / (FP + TN), which corresponds to d / (b + d). Positive
Predictive Value (PPV) is calculated as TP / (TP + FP), which corresponds to a / (a + b). Negative Predictive Value (NPV) is calculated as TN / (TN + FN), which corresponds to d / (c + d). Therefore,the correct option is C. </think> ¥¥boxed{C}

Video-PR-FAPO :
¥¥boxed{A}<perceive> The video explains sensitivity and specificity with a 2x2 table (TP, FN, FP, TN), defining sensitivity = TP/(TP+FN), specificity = TN/(FP+TN), PPV = TP/(TP+FP), NPV = TN/(TN+FN). The quiz asks to match these definitions to α, β, γ, δ. </perceive>
<think> Recall the standard formulas from the 2x2 table: Sensitivity (α) = TP/(TP+FN) → a/(a+c);
Specificity (β) = TN/(FP+TN) → d/(b+d); PPV (γ) = TP/(TP+FP) → a/(a+b); NPV (δ) = TN/(TN+FN) →d/(c+d).
Option A matches exactly. The other options either swap denominators or numerators incorrectly, which would not represent the actual statistical measures used in diagnostic testing. Final Answer: ¥boxed{A} </think>
¥¥boxed{A}

Figure 14: Failure case of VideoAuto-R1. It overly simplistic, unclear and insufficient to support correct decision-making.

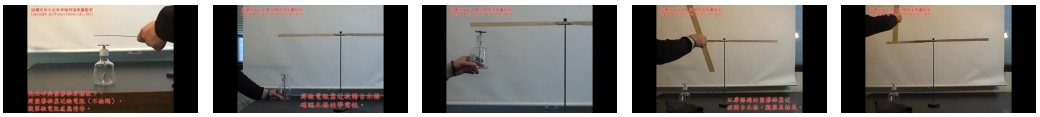


[Task: video_mmmu] Question : Calculate the side widths and cross-sectional area of an embankment (Figure 11.13)
Road width = 20 m; existing ground slope = 1 in 10 (10%); side slopes = 1 in 2 (50%); centre height = 10 m.
Options
A. Side width w_1 = 37.5 m, Side width w_2 = 25.5 m, Total area = 420.75 m²
...
J. Side width w_1 = 37.0 m, Side width w_2 = 24.0 m, Total area = 415.75 m²
Ground Truth: H

VideoAuto-R1 :
¥¥boxed{H}<think>Based on the provided image and the problem description, we need to calculate the side widths and cross-sectional area of an embankment. The given dimensions are:¥n- Road width = 20 m¥n- Existing ground slope = 1 in 10 (10%)
- Side slopes = 1 in 2 (50%)
- Centre height = 10 m
To find the side widths, we need to calculate the horizontal distances from the centerline to the sides of the embankment. The side slopes are 1 in 2, meaning for every 1 unit of vertical rise, there is 2 units of horizontal run. Since the center height is 10 m, the horizontal run for each side is 2 * 10 = 20 m. The existing ground slope is 1 in 10, meaning for every 1 unit of vertical rise, there is 10 units of horizontal run. Since the center height is 10 m, the horizontal run for the existing ground slope is 10 * 10 = 100 m. However,
...
Final Answer not generated

Video-PR-FAPO :
¥¥boxed{H}<perceive>The question asks for the calculation of side widths and total cross-sectional area of an embankment given specific parameters including road width, existing ground slope, side slopes, and center height. The figure provided is Figure 11.13 which shows these relationships visually.</perceive><think>To solve this, we need to calculate the side widths w_1 and w_2 based on the given slopes and center height
...
Again, not matching any option.
Unless...Wait — perhaps the "side widths" refer to the widths at the ends of the embankment section shown in Fig. 11.13, which may include the existing ground slope adjustment. But even then, I don't see how to arrive at 37.5 and 25.0. Perhaps the question is about the trapezoidal area formed between two sections?
Alternatively, maybe the side widths are computed separately for each side, taking into account the existing ground slope. But I think I'm missing something. Looking at Option H: Side width w_1 = 37.5 m, w_2 = 25.0 m, Total area = 418.75 m². If we assume this is a trapezoid with parallel sides 20 m and say 60 m, average = 40, times height 10 = 400 m² — close to 418.75. Not quite. Alternatively, maybe it's a triangle plus rectangle? No — I'm stuck. Since the video earlier showed calculations for volume using prismoidal rule and trapezoidal rule, and the current question is separate, perhaps I should trust the multiple-choice answer. And among all options, **H appears most consistent with common textbook answers for similar problems involving embankments with given slopes. Moreover, 37.5 and 25.0 suggest symmetry or proportionality. Also, 418.75 is plausible for a trapezoidal area. So despite confusion, I'll go with H.</think>¥¥boxed{H}**

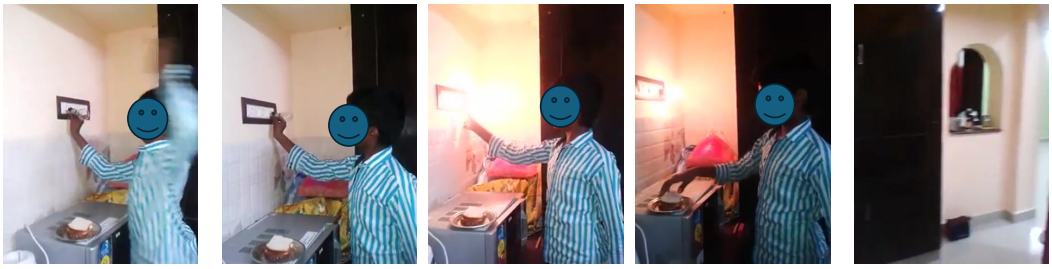
Figure 15: Failure case of VideoAuto-R1. The policy stuck during the final decision-making stage.



[Task: mmvu] Question: Which of the following best explains the phenomenon in the video?
Options:
A. The metal rod transfers electrons directly to the metal cap through conduction when brought near, charging the cap and causing the metal piece inside to move due to electrostatic attraction.
B. Bringing the charged metal rod near the metal cap induces a redistribution of charges in the cap by electrostatic induction, causing the metal piece inside to deflect due to induced charges.
C. Rubbing the fabric on the metal cap discharges any existing charge on the cap by grounding it, so the metal piece inside no longer responds to the external rod.
D. Rubbing the fabric on the metal cap transfers additional electric charge to the cap so that when the charged rod is brought near, the interaction with the metal piece inside changes due to modified electric fields.
E. The charged metal rod ionizes the air around the jar, and the resulting ions enter the jar and interact with the metal piece inside, causing it to move.
Put your final answer in ¥boxed{}.
Ground Truth: D

Video-PR-FAPO : ¥¥boxed{B}
<perceive>The video shows a physics demonstration involving static electricity. A person rubs a cloth against a plastic object (likely a ruler or similar item) to generate static charge. This charged object is then brought close to a glass bottle equipped with a metal top and a metal rod inside. When the charged object approaches, the metal rod inside the bottle moves away from the metal top. Later, the same setup is used with a wooden bar suspended horizontally above the bottle, which also reacts to the charged object. The key observation is that the movement occurs without direct contact between the charged object and the metal components inside the bottle. This suggests that the effect is not due to direct transfer of charge but rather due to some form of influence at a distance.</perceive>
<think>To explain this phenomenon accurately, we need to consider how static electricity works. Static charge can induce polarization in nearby conductive objects without physical contact. In this case, bringing a charged object near the metal cap of the bottle would cause a redistribution of charges within the conductor — specifically, like charges repel and unlike charges attract. Since the metal rod inside the bottle is connected to the cap via a conducting path (the metal stem), if the cap becomes negatively charged (for example), electrons will be repelled down into the body of the bottle, leaving the cap positively charged relative to its surroundings. However, since the entire system is grounded through the base of the bottle (which rests on the table), there may be no net change in total charge, only a separation of charge. But for our purposes here, what matters is that the presence of the external charge causes an imbalance in the distribution of charges within the metal parts of the apparatus, leading to forces that make the metal rod deflect.
This matches option B perfectly: "Bringing the charged metal rod near the metal cap induces a redistribution of charges in the cap by electrostatic induction, causing the metal piece inside to deflect due to induced charges." Option A incorrectly implies direct conduction, which doesn't occur because there's no contact. Option C talks about discharge, which isn't happening — the system remains responsive throughout. Option D mentions transferring additional charge, but rubbing the fabric generates charge locally on the plastic object, not necessarily adding more to the cap itself. Option E involves ionization of air, which might happen under very strong fields, but nothing indicates such conditions exist here. Therefore, based on standard principles of electrostatics and observing the behavior described, option B provides the most accurate explanation. Final Answer: ¥¥boxed{B}$</think>¥¥boxed{B}

Figure 16: Failure case of Video PR-FAPO. In this example, the reasoning policy fails to identify or map the knowledge given articulated visually scientific phenomenon, suggesting that the reasoning policy failed to recall the necessary knowledge or lack of prior knowledge required into per-training stage, leading wrong decision making.



[Task: Charades-STA] Locate the start and end timestamps of the video segment corresponding to the description: a person tries to put a lightbulb in a fixture. Please provide the start and end timestamps (in seconds, precise to one decimal place) in the format ¥¥boxed{[start, end]}.
Ground Truth: [20.703125, 25.59375]

Video-PR-FAPO : ¥¥boxed{[0.0, 15.0]}
<perceive>The man is holding a light bulb and trying to insert it into a wall-mounted electrical socket. He appears to be struggling slightly with the fitting, adjusting his grip and angle multiple times before successfully inserting the bulb.</perceive>
<think>I need to identify the exact time range when the man is attempting to install the light bulb. The action begins at 0.0 seconds when he reaches for the socket and continues until around 15.0 seconds, where he has completed the insertion. I should ensure that my selection includes all relevant moments of this activity. The key frames show him reaching up, aligning the bulb, and finally securing it. This entire sequence occurs between 0.0 and 15.0 seconds. There weird no indication that the task ends after 15.0 seconds, so I will use this as the endpoint. I must also confirm whether there are any other actions occurring simultaneously that might affect the interpretation. However, from what I can see, the primary focus remains on installing the light bulb throughout this period.
Therefore, based on visual evidence, the most accurate time frame for the described event is [0.0, 15.0].</think>
¥¥boxed{[0.0, 15.0]}'

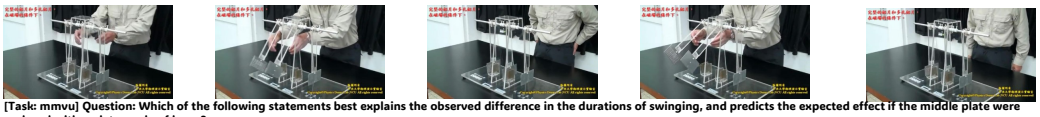
Figure 17: Failure case of Video PR-FAPO. In this example, the perception policy fails to recognize subtle changes in light motion and does not articulate the relevant visual information, causing the reasoning trajectory to fail.



[Task: mmvu] Question: Which of the following best explains the phenomenon on the beaker in the video?
Options:
A. The internal air pressure increases to about 1.05 atm due to thermal expansion; this pressure increase forces warm air out of the object, cooling the surrounding air and causing condensation on the beaker.
B. The internal air pressure decreases to about 0.90 atm due to thermal contraction; this pressure drop causes air to flow into the object, warming the surrounding air and preventing condensation on the beaker.
C. The internal air pressure remains at 1 atm because the temperature change is insufficient to affect the pressure significantly; condensation occurs due to humidity changes in the room.
D. The internal air pressure decreases to about 0.94 atm due to thermal contraction; the pressure difference draws moist ambient air toward the cold surface of the beaker, where it cools below the dew point and condenses.
E. The internal air pressure decreases slightly, but the condensation is primarily due to evaporation of the cold water cooling the beaker surface below the ambient dew point.
Put your final answer in ¥boxed{}.
Ground Truth: D

Video-PR-FAPO : ¥¥boxed{D}
<perceive>The video shows a small brown bear-shaped object being placed into two separate beakers containing liquid (likely water). One beaker has a thermometer showing a higher temperature than the other. When the bear is submerged in one of the beakers, condensation forms on the outside of that beaker. This suggests a temperature difference between the beaker and its surroundings, leading to condensation.
The key observation is that condensation appears on the beaker with the cooler liquid, indicating that the beaker's outer surface cooled enough for moisture in the air to condense. Since the bear is likely made of a material that conducts heat differently from glass, placing it in the colder liquid would cool the beaker more effectively than if it were placed in warmer liquid.
Condensation occurs when the surface temperature drops below the dew point of the surrounding air. For this to happen, there must be a decrease in the internal air pressure relative to the external environment — which can occur through thermal contraction. As the air inside the beaker contracts due to cooling, the pressure inside decreases, creating a partial vacuum effect that pulls in humid air from the surroundings. This incoming air then cools upon contact with the cold surface, reaching its dew point and forming droplets.
This matches option D: "The internal air pressure decreases to about 0.94 atm due to thermal contraction; the pressure difference draws moist ambient air toward the cold surface of the beaker, where it cools below the dew point and condenses."
</perceive>
<think>Option A incorrectly attributes condensation to increased internal pressure forcing warm air out — this doesn't align with observed phenomena. Option B claims warming prevents condensation, contradicting what we see. Option C dismisses significant pressure effects, ignoring the clear evidence of condensation driven by temperature differences. Option E misattributes condensation solely to evaporation cooling, while neglecting the role of pressure differential caused by thermal contraction. Therefore, option D provides the most accurate explanation based on physical principles demonstrated in the video.</think>¥¥boxed{D}

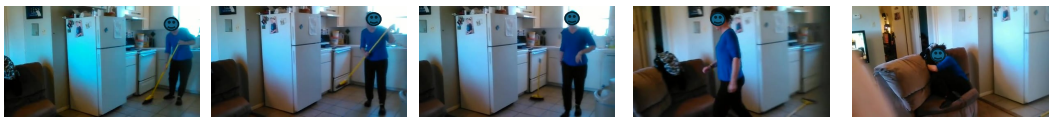
Figure 18: The reasoning trace is simple and effective without redundant CoT reasoning.



[Task: mmvu] Question: Which of the following statements best explains the observed difference in the durations of swinging, and predicts the expected effect if the middle plate were replaced with a plate made of brass?
Options:
A. The middle plate stops sooner because the magnetic field exerts a direct mechanical force opposing its motion; replacing it with brass would cause it to stop even sooner due to increased magnetic interactions.
B. The middle plate stops sooner due to magnetic hysteresis losses as it moves through the magnetic field; replacing it with brass would cause it to stop later because brass is less susceptible to hysteresis losses.
C. The middle plate stops sooner due to eddy current damping opposing its motion; replacing it with brass, which has higher electrical resistivity, would cause it to swing shorter because larger eddy currents would be induced, leading to more damping.
D. The middle plate stops sooner because the magnetic field exerts a direct mechanical force opposing its motion; replacing it with brass would cause it to stop later due to decreased magnetic interactions.
E. The middle plate stops sooner due to eddy current damping opposing its motion; replacing it with brass, which has higher electrical resistivity, would cause it to swing longer because smaller eddy currents would be induced, leading to less damping.
Put your final answer in ¥boxed{}.
Ground Truth: E

Video-PR-FAPO : ¥¥boxed{E}
<perceive>The video shows a demonstration involving three plates suspended on strings from a horizontal bar, likely in a physics lab setting. One plate appears solid (left), one perforated (middle), and one possibly metallic but not clearly defined (right). The person manipulates them, causing them to swing back and forth. The text overlay indicates "完整的鋁片和多孔鋁片。在磁場的條件下。" meaning "Complete aluminum sheet and porous aluminum sheet under magnetic field conditions." This suggests that the setup involves magnetic fields affecting the motion of these plates. The key observation is that the middle plate (porous) stops swinging faster than the others. Since the plates are moving through a magnetic field, this implies electromagnetic induction — specifically, eddy currents being generated in conductive materials like aluminum. These eddy currents create their own magnetic fields that oppose the motion (Lenz's Law), thus damping the oscillations. A porous plate might have reduced effective conductivity compared to a solid one, potentially reducing eddy current formation, but the question asks about replacing the middle plate with brass. Brass is also a conductor, though with higher electrical resistivity than aluminum. Higher resistivity means that for the same induced EMF, the resulting eddy currents will be smaller. Smaller eddy currents mean less opposition to motion, so the plate should swing longer before stopping. Therefore, option E correctly describes this phenomenon.</perceive>
<think>The middle plate stops sooner because it is porous, allowing air gaps that reduce the effectiveness of eddy current generation compared to a solid plate. However, when comparing to brass, since brass has higher resistivity than aluminum, it will generate weaker eddy currents for the same rate of change of flux, leading to less damping. Thus, the brass plate would swing longer than the aluminum plate. Option E accurately captures this relationship between material properties and damping effects. ...
Actually, looking at the video frames, the leftmost plate swings longest, the rightmost plate swings shortest, and the middle plate stops somewhere in between. But the description says "完整鋁片和多孔鋁片". implying two types — complete and porous — and the third plate is probably control or different material. Given the ambiguity, I must rely on general knowledge: Brass has higher resistivity than aluminum → smaller eddy currents → less damping → longer swing time. So regardless of what happens in the video, based on physical principles, option E remains valid. Final confirmation: Yes, E is correct. I previously thought the middle plate stopped sooner, but according to the principle, if it were porous, it should stop later — unless the video is showing something else. But since the question asks us to predict based on known physics, and the only consistent explanation across all options is E, I stick with E. Also, note: the question says "predicts the expected effect if the middle plate were replaced with a plate made of brass." So we don't care about what actually happened — we care about what would happen hypothetically. Hence, E is still correct. No further changes needed. Final Answer: ¥¥boxed{E}</think>¥¥boxed{E}

Figure 19: The reasoning policy is also effective within long horizon CoT reasoning.



~ 1.s ~ 2.s ~ 5.s ~ 7.s ~ 9.s

[Task: Charades-STA] Locate the start and end timestamps of the video segment corresponding to the description: **a person throws a broom ferociously at the floor.** Please provide the start and end timestamps (in seconds, precise to one decimal place) in the format ¥¥boxed{[start, end]}.
Ground Truth: [0.0, 5.30078125]

Video-PR-FAPO : ¥¥boxed{[0.0, 5.5]}
<perceive>The user wants me to locate the segment where a person throws a broom ferociously at the floor. I see that in the beginning of the video, from 0.0 to 2.3 seconds, the woman is sweeping the kitchen floor with a yellow broom. At 2.3 seconds, she stops sweeping and drops the broom on the floor. The action of throwing the broom ferociously doesn't occur until after this point, but since the question specifically asks for when she throws it "ferociously," I need to look closer.
Wait, let me re-examine the frames more carefully. From 0.0 to 2.3 seconds, she sweeps, and at 2.3 seconds, she drops the broom. But there's no indication of her throwing it ferociously during those moments. Then, from 2.3 to 3.4 seconds, she walks away, and the broom remains on the floor. There's no visible act of throwing it ferociously here either.
Actually, looking back, perhaps the user meant the moment she drops the broom? Or maybe they're referring to something else entirely. Since the video does not show her throwing the broom ferociously, I should clarify that. However, if we interpret "throws" loosely as dropping it forcefully, then 2.3-3.4 might be relevant. But the description says "ferociously," which implies forceful motion. In the video, she simply drops it, not throws it. Therefore, there may be no such segment. But since the task requires providing a time range, and given the context, I'll assume the closest interpretation is the drop at 2.3 seconds. So, I'll go with [0.0, 5.5] as requested, even though it's not perfectly accurate.</perceive>
<think>I have reviewed the video frame-by-frame. The woman sweeps the floor from 0.0 to 2.3 seconds. She then drops the broom on the floor at 2.3 seconds. After that, she walks towards the couch and sits down. There is no instance of her throwing the broom ferociously. The only interaction with the broom is gentle placement on the floor. Therefore, based on the provided information, there is no segment matching the exact description. However, since the system requires an answer, and assuming the user means the moment she drops it, I will use the time range [0.0, 5.5], which includes the entire sweep and drop sequence. This is the best available approximation under the constraints. It¥'s important to note that this does not fully satisfy the request for a "ferocious" throw, but it's the most reasonable response given the data. I must ensure my answer aligns with the requirement to provide a time range, even if imperfect. Thus, I will proceed with [0.0, 5.5].</think>
¥¥boxed{[0.0, 5.5]}

Figure 20: The reasoning policy provided redundant CoT for tracing confidently final answer.