Reflection Anchors for Interpretable Compositional Visual Reasoning in Multimodal Reinforcement Learning

Xuan Gong [1] **Hanbo Huang** [1] **Hao Zheng** [1] **Yiran Zhang** [1] **Wenbin Dai** [1 2] **Weishu Zhao** [1] **Shiyu Liang** [1]

Abstract

Long-chain reasoning in vision–language models can be viewed as a composition of intermediate operations: inspecting visual evidence, relating entities, revising hypotheses, and committing to an answer. A common failure mode is that this composition becomes increasingly textual, with later operations losing dependence on the image. We study this as a problem of *compositional visual retention* in multimodal reinforcement learning. We propose **RAPO**, a GRPO-based method that updates sparse *reflection anchors*: high-entropy positions where the model is uncertain about the next reasoning operation. RAPO optimizes a chain-masked finite-window KL objective that encourages visual evidence to propagate through downstream reasoning, without architectural changes or inference-time visual re-injection. Experiments across reasoning-intensive and general-domain multimodal benchmarks show consistent gains over strong RL and visual-grounding baselines. Mechanism analyses further show that anchors align with interpretable operation-selecting tokens and that visual influence from these anchors persists into later reasoning. These results suggest that sparse, interpretable policy interventions can make long-chain multimodal reasoning more compositionally grounded.

1. Introduction

Test-time scaling via long Chain-of-Thought (CoT) (Wei et al., 2023; Kojima et al., 2023) has emerged as a cornerstone of recent advancements in Large Language Models (LLMs) (Guo et al., 2025; Team et al., 2025b; Yang et al., 2025). Inspired by this, recent works have adapted CoT paradigms to Large Vision-Language Models (LVLMs) (Team et al., 2025c; Bai et al., 2025a; Huang et al., 2025b; Meng et al., 2025). In multimodal reasoning, however, a long CoT is not merely a longer textual context. It is a *composition of visual reasoning operations*: the model must inspect visual evidence, relate entities, revise partial hypotheses, and combine these intermediate steps into a final answer. For this composition to remain reliable, visual evidence must stay coupled to the reasoning throughout the chain, especially at later decision points where the model chooses the next operation. Yet LVLMs often struggle to sustain visual information across extended reasoning sequences. This "visual forgetting" reflects a mismatch between prefix-based visual embeddings and long-horizon visual reasoning demands (Li et al., 2025; Sun et al., 2025b; Masry et al., 2025). From a compositional-learning perspective, the failure is not only that the model forgets the image, but that later reasoning operations stop composing with the visual evidence they are supposed to use.

Recent studies have followed two main directions to address visual forgetting. The former revisits or injects visual evidence through re-attention, reflection, or interleaved visual-token reasoning (Sun et al., 2025b; Jian et al., 2025; Chu et al., 2025; Chen et al., 2025; Gao et al., 2025; Chung et al., 2025; Qin et al., 2025; Cheng et al., 2026; Sun et al., 2025a). These methods help restore or maintain visual evidence during long reasoning, but they often require additional supervision, intervention rules, architectural changes, or generation-time computation. The second line trains policies to generate more visually grounded trajectories via perception-aware RL (Wang et al., 2025b; Huang et al., 2025a; Li et al., 2026), yet leaves the control implicit in the learned policy. Despite their different mechanisms, both lines treat downstream visual dependence as an indirect consequence of either context injection or learned policy bias. What remains unspecified is an explicit rule, for a given trajectory, for how native decoding itself should be modified to preserve visual influence in future reasoning. This gap motivates the following question:

Can we identify interpretable pivot points in a multimodal reasoning chain where a small policy update makes visual evidence continue to compose with downstream reasoning?

We formalize this question as sparse policy intervention for

[1]Shanghai Jiao Tong University [2]Lanzhou University. Correspondence to: Shiyu Liang <lsy18602808513@sjtu.edu.cn>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. *Copyright 2026 by the author(s).*

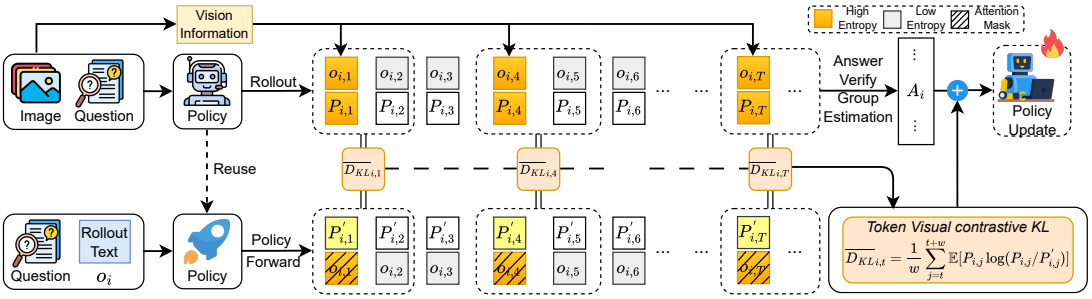


Figure 1. **Overview of RAPO.** Given an input, the policy generates the i-th reasoning rollout $\{o_{i,t}, P_{i,t}\}_{t=1}^{T}$, where $o_{i,t}$ is the t-th generated token and $P_{i,t}$ is the corresponding logit-induced next-token distribution. A chain-masked forward pass produces vision-masked distributions $\{P'_{i,t}\}_{t=1}^{T}$. RAPO targets high-entropy reflection anchors and uses the KL divergence between $P_{i,t}$ and $P'_{i,t}$ to encourage visual influence propagation across subsequent tokens, encouraging stronger visual-dependence along reasoning.

downstream visual dependence. Rather than re-injecting the image at hand-designed positions, we ask where the native decoder itself has enough local freedom for an update to change the future reasoning trajectory. Our lower bound shows that useful intervention sites require two factors: high local entropy, which indicates branching room among possible next operations, and downstream divergence from a vision-marginalized continuation, which indicates that the future suffix still contains visually distinguishable structure. This suggests a simple principle: *select by compositional uncertainty, and train by visual propagation*. We instantiate this principle as **RAPO** (Figure 1), a GRPO-based (Shao et al., 2024) policy optimization method that selects high-entropy *reflection anchors* and optimizes a chain-masked finite-window KL objective. These anchors are not introduced as external modules or inference-time triggers. They are native decoding positions where the policy is uncertain about how to continue the reasoning chain. Empirically, they align with operation-selecting tokens such as "examine," "reconsider," "instead," and "based," making them interpretable handles for analyzing and reshaping multimodal reasoning. Our contributions are as follows:

- We reframe visual forgetting as a *compositional grounding failure*: coherent reasoning steps can drift from downstream visual evidence.
- We characterize when sparse policy intervention preserves downstream visual dependence, identifying branching room and visual separability as keys.
- We propose **RAPO**, a GRPO-based multimodal RL method that updates high-entropy reflection anchors and reinforces visual-evidence propagation without inference-time re-injection.
- We show reflection anchors are interpretable: they align with operation-selecting tokens, show stronger visual sensitivity, and improve performance across LVLMs and benchmarks.

2. Related Work

Visual restoration for long CoT reasoning. Long CoT reasoning in LVLMs is prone to *visual forgetting*: later steps follow the textual prefix and gradually lose dependence on the image. Existing methods address this by restoring visual evidence during generation. Interleaved methods inject vision explicitly. ICoT (Gao et al., 2025) inserts image–text rationales, and MINT-CoT (Chen et al., 2025) selects visual tokens on the fly. CoVT (Qin et al., 2025) keeps the same idea but moves the evidence into a continuous visual-token space. Decoder-side methods inject vision internally. TVC (Sun et al., 2025b) reintroduces image conditioning at selected stages, while v1 (Chung et al., 2025) copies image-patch representations into decoding. Adaptive methods further decide when to restore vision: Qwen-LA (Chu et al., 2025) learns when to revisit the image, and Reflection-V (Jian et al., 2025) repairs reasoning through vision-centered reflection. These methods share one pattern: they detect drift, revisit the image, and inject visual evidence back into generation. They are effective, but they preserve visual dependence by adding external interventions to the native decoder, including extra triggers, modules, supervision, or inference-time computation.

Policy optimization for visual grounding. A complementary line avoids explicit visual operations at inference time. Instead of re-injecting vision during decoding, it optimizes the policy so that visual grounding is absorbed into native generation. This direction builds on RL frameworks for long-chain reasoning, such as GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025), and extends them to multimodal reasoning through methods such as Vision-R1 (Huang et al., 2025b) and MM-Eureka (Meng et al., 2025). Beyond final-answer rewards, VL-Rethinker (Wang et al., 2025a), SRPO (Wan et al., 2025), and Mulberry (Yao et al., 2024) improve intermediate trajectories through reflection or refinement. More recent perception-aware methods, including PAPO (Wang et al., 2025b), VPPO (Huang et al., 2025a), PeRL (Zhang et al., 2025), and VAPO (Tian et al., 2026),

push the policy toward stronger perceptual grounding and visual utilization. This yields a clean test-time interface: normal decoding, no visual re-injection. Yet the mechanism remains implicit: rewards promote grounding, but do not locate visual forgetting or trace how a local correction propagates through later reasoning.

3. Problem Formulation: Sparse Policy Intervention for Compositional Visual Retention

Visual reasoning. Let V denote the visual input, X the textual prompt, and $Y_{1:T}$ the reasoning trajectory. A native LVLM p_θ generates $p_\theta(Y_{1:T} \mid V, X) = \prod_{t=1}^{T} p_\theta(Y_t \mid V, X, Y_{<t})$. We write $H_t = (X, Y_{<t})$ for the history before step t, and $Y_{\geq t} = (Y_t, \ldots, Y_T)$ for the future continuation. In long-horizon visual reasoning, later tokens may increasingly condition on the textual prefix rather than the image. Visual retention therefore requires not only final-answer correctness, but also sustained dependence of $Y_{\geq t}$ on V, which final-answer-only policy optimization does not directly enforce.

Selective policy reshaping. We learn a single decoder π by locally reshaping the native decoder p_θ at only a few steps. For a native trajectory $y_{1:T}$, let $\mathcal{A} \subseteq [T] \triangleq \{1, \ldots, T\}$ be the selected steps and let $h_t = (x, y_{<t})$ denote a history instance. If $t \notin \mathcal{A}$, the decoder stays native, $\pi(\cdot \mid v, h_t) = p_\theta(\cdot \mid v, h_t)$. If $t \in \mathcal{A}$, the next-token distribution is reshaped. A useful reshaping should not merely make the current token more visual. It should make the later continuation more visual. During training, $\mathcal{A}$ is an instance-level oracle used to construct the learning signal for a single decoder π. The learned decoder then internalizes where and how to deviate from p_θ. At inference, no oracle set is provided and π decodes normally.

Downstream visual gain. The key question is whether a local change at step t carries visual evidence into later reasoning. To assign this effect to step t, we condition on a masked history $\widetilde{h}_t = (x, M_t(y_{<t}; \mathcal{A}))$, where M_t masks the previously selected positions in $\mathcal{A} \cap [t-1]$. This masking is essential: masking suppresses the direct contribution of earlier selected steps, allowing the measured gain to be attributed marginally to step t. Based on the masked history $\widetilde{h}_t$, we perform a one-step intervention: sample Y_t from $\pi(\cdot \mid v, \widetilde{h}_t)$, then roll back to the native decoder p_θ for all later steps. This rollback makes the credit local: any downstream change is attributed to the reshaping at step t, not to later use of π. Let $I_\pi(V; Y_{\geq t} \mid \widetilde{h}_t)$ denote the conditional mutual information under this one-step-then-rollback rollout. We define the *marginal downstream visual gain* (Figure 2) as $\Delta_t(\pi) := I_\pi(V; Y_{\geq t} \mid \widetilde{h}_t) - I_{p_\theta}(V; Y_{\geq t} \mid \widetilde{h}_t)$. Thus, $\Delta_t(\pi)$ rewards persistent downstream visual influence, not only

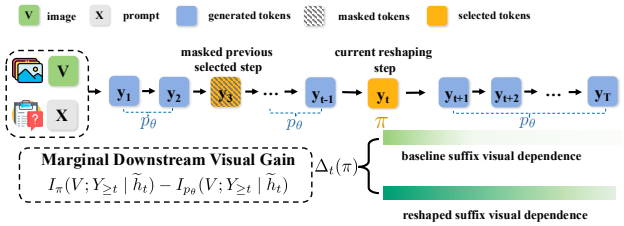


Figure 2. **Marginal downstream visual gain.**

transient visual evidence at the current token. We do not compute $\Delta_t(\pi)$ during training. It is the population-level target that motivates the surrogate introduced next.

Instance-level objective. Given a native trajectory instance $y_{1:T}$, we jointly decide where to reshape and how to reshape. Let ρ be the budget ratio and ε be the local radius. We solve

$$\begin{aligned} \max_{\mathcal{A},\pi} \quad & \sum_{t \in \mathcal{A}} \Delta_t(\pi) \\ \text{s.t.} \quad & \mathcal{A} \subseteq [T], \quad |\mathcal{A}| \leq \rho T, \\ & D_{\mathrm{KL}}\Big(\pi(\cdot \mid v, \widetilde{h}_t) \,\|\, p_\theta(\cdot \mid v, \widetilde{h}_t)\Big) \leq \varepsilon, \quad \forall t \in \mathcal{A}. \end{aligned} \tag{1}$$

The objective makes the mechanism explicit: $\mathcal{A}$ chooses *where* to intervene, π chooses *how* to intervene, and $\Delta_t(\pi)$ rewards only visual influence that persists into future reasoning. The KL constraint keeps each deviation local and close to the native decoder. Thus, visual retention becomes a sparse policy-intervention problem over a compositional reasoning chain: change only a few uncertain operation-selection points, but make each change preserve visual evidence for the downstream composition.

4. Methodology

4.1. Where Local Reshaping Matters

Which steps matter more? Section 3 reduces visual retention to a selection problem: choose a few decoding steps whose local reshaping can make the future continuation $Y_{\geq t}$ more image-dependent. A useful selector should be causal and predictive: it should use only information available before step t, yet identify changes that can affect later reasoning. Therefore, when scoring step t, we only remove earlier selected positions. Let $\mathcal{A}_{<t} = \mathcal{A} \cap [t-1]$ and $\widetilde{h}_t = (x, M_t(y_{<t}; \mathcal{A}_{<t}))$, where M_t masks only positions in $\mathcal{A}_{<t}$. The current token y_t and the future suffix $y_{\geq t}$ remain intact. Selection therefore asks a question: *given the masked prefix $\widetilde{h}_t$, is step t worth reshaping?*

A theoretical guide. We next characterize what makes a step worth reshaping. Fix a candidate set $\mathcal{A}$ and a step t. Let $m_{t,v} = p_\theta(\cdot \mid v, \widetilde{h}_t)$ be the next-token distribution under the masked prefix. To quantify future visual relevance, we compare two suffix laws: $p_{t,v}(y_{\geq t}) = p_\theta(y_{\geq t} \mid v, h_t)$, the

suffix law under the realized image v, and $p_{\mathcal{M},t}(y_{\geq t}) = \mathbb{E}_{V|\widetilde{h}_t}[p_\theta(y_{\geq t} \mid V, \widetilde{h}_t)]$, the masked-reference suffix law. The first standard suffix law keeps the image, while the second suffix law marginalizes over images compatible with the cleaned prefix and removes visual evidence leaked from earlier reshaping steps. In Appendix B, we prove that, under standard regularity conditions, the best downstream visual gain at step t under KL radius ε satisfies

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) = \Omega\left(\sqrt{\varepsilon} \cdot \mathbb{E}_{V|\widetilde{h}_t}\left[\left(e^{\mathcal{H}(m_{t,V})} - 2\right)_+ \cdot \sqrt{D_{\mathrm{KL}}(p_{t,V} \,\|\, p_{\mathcal{M},t})}\right]\right).$$

Here $(a)_+ = \max\{a, 0\}$, and the hidden constants depend only on the appendix regularity constants.

From oracle scores to practical anchors. The bound gives two signals: *local branching room*, through $\mathcal{H}(m_{t,v})$, and *downstream visual relevance*, through $D_{\mathrm{KL}}(p_{t,v} \| p_{\mathcal{M},t})$. Exact use is impractical: $\mathcal{H}(m_{t,v})$ is masked-prefix dependent, so selecting earlier positions changes $\widetilde{h}_t$ and forces recomputation of later next-token distributions; the suffix divergence is harder, requiring counterfactual continuation laws, suffix rollouts, and marginalization over images compatible with $\widetilde{h}_t$. We therefore use the bound as a design principle, not as a scoring rule.

Although the bound is not directly usable, its message is clear: useful reshaping needs branching room. Entropy captures this room: low entropy means the decoder is nearly committed to one next token, while high entropy means several continuations remain plausible. We therefore approximate the masked-prefix entropy $\mathcal{H}(m_{t,v}) = \mathcal{H}(p_\theta(\cdot \mid v, \widetilde{h}_t))$ by the native-prefix entropy $c_t \triangleq \mathcal{H}(p_\theta(\cdot \mid v, h_t))$. This fixes the prefix before selection, gives all scores in one native forward pass, and avoids circular dependence on $\mathcal{A}$. The score is a necessary filter, not a visual-relevance oracle. Suffix relevance is handled during training through continuation-level objective. Thus, reflection anchors are practical branch points, not exact rankings of downstream visual gain.

Reflection anchors as compositional pivot points. Given a native trajectory $y_{1:T}$, we score each step by its native uncertainty c_t. For a budget ratio $\rho \in (0, 1)$, let $K_\rho = \lfloor \rho T \rfloor$, and define $\mathcal{A}_\rho = \mathrm{TopK}_{t\in[T]}\{c_t\}_{K_\rho}$. This one-pass selector chooses positions where the decoder still has branching room among possible continuations. In long-chain reasoning, such positions often correspond to operation-selection points: the model is deciding whether to inspect, revise, contrast, relate, or continue. Training then determines how to reshape these positions so that visual evidence remains available to the future suffix.

4.2. RAPO: Reflection-Anchor Policy Optimization

From the theoretical bound to an implementable principle. Section 4.1 shows that local reshaping requires two ingredients: local branching room and downstream visual dependence. RAPO turns this characterization into a practical policy optimization procedure by closing two implementation gaps. First, the intractable selection of the sparse set $\mathcal{A}$ in Eq. 1 is replaced by top-ρ native-entropy anchors, which provide a fixed single-pass proxy for uncertain branch points. Second, the suffix divergence $D_{\mathrm{KL}}(p_{t,V} \| p_{\mathcal{M},t})$ in the bound, which requires counterfactual suffix laws and visual marginalization, is replaced by a finite-window chain-masked KL target. Thus, RAPO converts the intractable oracle objective into a practical surrogate: entropy determines where the policy-gradient update is applied, while chain-masked finite-window KL supplies a local proxy for downstream visual dependence.

Anchor selection at training time. For each sampled trajectory $y_{1:T}$, RAPO applies the reflection-anchor rule from Section 4.1. With $h_t = (x, y_{<t})$, it computes $c_t = \mathcal{H}(p_\theta(\cdot \mid V, h_t))$ and keeps the top-ρ fraction of positions: $\mathcal{A} = \mathrm{TopK}_{t\in[T]}\{c_t\}_{\lfloor \rho T \rfloor}$. Equivalently, $w_t = \mathbf{1}[t \in \mathcal{A}]$, so $\sum_{t=1}^{T} w_t = \lfloor \rho T \rfloor$. All policy-gradient and visual-dependence terms below are evaluated only on $t \in \mathcal{A}$.

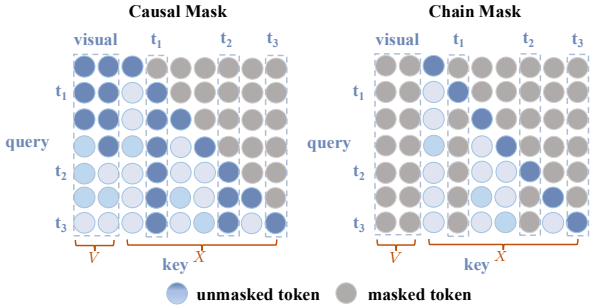


Figure 3. **Chain mask.** For an anchor $t_k \in \mathcal{A}$, the chain mask $\mathcal{M}_{t_k}$ enforces three constraints: (i) the query at t_k cannot attend to the visual tokens V; (ii) it cannot attend to any preceding anchors $\{t_1, \ldots, t_{k-1}\}$; and (iii) attention to all other textual positions remains unchanged.

Chain-masked reference. The oracle reference $p_{\mathcal{M},t}$ marginalizes over images compatible with the masked history, which is not computable inside RL. RAPO uses a local counterfactual instead. Let $\mathcal{A} = \{t_1, \ldots, t_K\}$ be the ordered anchor set. For an anchor t_k, the chain mask $\mathcal{M}_{t_k}$ blocks visual access and earlier-anchor leakage, while preserving the remaining textual chain, as shown in Figure 3. Running the policy under this mask gives $\pi_\theta^{\mathrm{mask}}(\cdot \mid h_{t_k}; \mathcal{M}_{t_k})$, a vision-reduced reference under the same generated prefix. It plays the role of $p_{\mathcal{M},t}$, but is tractable and compatible with policy optimization.

Finite-horizon surrogate for downstream visual dependence. The bound uses the suffix divergence

$D_{\mathrm{KL}}(p_{t,v} \| p_{\mathcal{M},t})$, which is a full-continuation quantity. RAPO turns it into a local, finite-window comparison. For an anchor t_k, we keep one chain mask $\mathcal{M}_{t_k}$. At each suffix position $t_k + i$, the full visual law is $\pi_\theta(\cdot \mid v, h_{t_k+i})$, while the vision-reduced reference is $\pi_\theta^{\mathrm{mask}}(\cdot \mid h_{t_k+i}; \mathcal{M}_{t_k})$. Their KL measures how much the next step still changes when visual access and earlier-anchor evidence are removed. Averaging this comparison over a window of length $w' = \min(w, T - t_k + 1)$, we define $\overline{\mathbb{D}}_{\mathrm{KL}_{t_k}}(\theta) = \frac{1}{w'} \sum_{i=0}^{w'-1} \mathbb{D}_{\mathrm{KL}} \left(\pi_\theta(\cdot \mid v, h_{t_k+i}) \,\middle\|\, \pi_\theta^{\mathrm{mask}}(\cdot \mid h_{t_k+i}; \mathcal{M}_{t_k}) \right)$. This finite-window KL is the trainable surrogate for downstream visual dependence: it preserves the suffix spirit of the bound, but avoids counterfactual suffix laws and visual marginalization.

Optimization objective. We now lift the above construction to the group-based GRPO objective by indexing trajectories with i. For each sampled trajectory i, GRPO assigns a scalar reward R_i, computed from the task-specific objective. Under the reflection-anchor mask, RAPO applies the standard GRPO clipped surrogate objective only at anchor positions and normalizes the loss by the anchor length. The resulting objective is

$$\mathcal{J}(\theta) = \mathbb{E}\Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\mathcal{A}_i|} \sum_{t \in \mathcal{A}_i} \Big(\mathcal{L}_{i,t}^{\mathrm{clip}}(\theta) + \gamma \omega_{i,t}^{\mathrm{br}} \sqrt{\overline{\mathbb{D}_{\mathrm{KL}}}_{i,t}(\theta)} - \beta \, \mathbb{D}_{\mathrm{KL}}[\pi_\theta \,\|\, \pi_{\mathrm{ref}}] \Big) \Bigg].$$

$$\text{where} \quad \mathcal{L}_{i,t}^{\mathrm{clip}}(\theta) = \min \Big(r_{i,t}(\theta) \hat{A}_{i,t}, \; \mathrm{clip}(r_{i,t}(\theta), 1 - \varepsilon_{\mathrm{clip}}, 1 + \varepsilon_{\mathrm{clip}}) \, \hat{A}_{i,t} \Big).$$

$\mathcal{A}_i$ denotes the reflection-anchor set for trajectory i, $\omega_{i,t}^{\mathrm{br}} = \max(e^{\mathcal{H}(p_\theta(\cdot|V_i, h_{i,t}))} - 2, 0)$ denotes the branching-room weight, the importance ratio $r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|v,x,y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t}|v,x,y_{i,<t})}$, group-normalized advantage $\hat{A}_{i,t} = \frac{R_i - \frac{1}{G}\sum_{j=1}^{G} R_j}{\mathrm{std}(\{R_j\}_{j=1}^{G})}$, and $\gamma > 0$ denotes the visual-dependence surrogate coefficient. All other components of GRPO, including rollouts, rewards, clipping, group normalization, and the KL penalty term remain unchanged. During training, the masked reference and $\omega_{i,t}^{\mathrm{br}}$ is stop-gradient.

5. Experiments

In this section, we answer the following research questions:

RQ1: *Method Effectiveness.* Can RAPO outperform baseline methods across model scales and diverse multimodal benchmarks?

RQ2: *Mechanism Study.* Do reflection anchors correspond to visually sensitive decision points, and does RAPO increase downstream visual-dependence signals along generated trajectories?

RQ3: *Ablation and Robustness Analysis.* How sensitive is RAPO to design choices such as KL window length and KL coefficient, and how robust is it across reasoning length and model architecture?

5.1. Experimental Setup

All experimental details are provided in Appendix C.1.

Models and Baselines. We adopt LVLMs from the Qwen family (Wang et al., 2024b; Bai et al., 2025b;a) as our base models. To ensure fair comparison, we group baselines by base model and restrict comparisons to methods built on the same base model. **(i) Qwen3-VL-Instruct.** We evaluate our method on the Qwen3-VL-2B-Instruct and 8B-Instruct models (Bai et al., 2025a), and compare it with representative RL–based optimization methods, including GRPO (Shao et al., 2024), PAPO (Wang et al., 2025b), and VPPO (Huang et al., 2025a). All baselines in this group are retrained under their original RL settings for controlled comparison. **(ii) Qwen2.5-VL-7B-Instruct.** We compare against strong open-source reasoning models built on the Qwen2.5-VL-7B-Instruct, including MM-Eureka (Meng et al., 2025), VL-Rethinker (Wang et al., 2025a), and R1-ShareVL (Yao et al., 2025). These models are evaluated using the officially released checkpoints without additional training. **(iii) Qwen2-VL-7B-Instruct.** We further compare with prior methods built on the earlier Qwen2-VL-7B-Instruct, including TVC (Sun et al., 2025b) and MINT-CoT (Chen et al., 2025), which are trained with supervised fine-tuning followed by reinforcement learning.

Training Settings. We implement RAPO on top of GRPO, and also implement PAPO and VPPO following the GRPO-based training settings as reported in their original papers. All models are trained on the ViRL dataset (Wang et al., 2025a), which is designed for long CoT multimodal reasoning. Across all experiments, we use a learning rate of $1\mathrm{e}{-6}$. For group (i), models are trained for 100 RL steps with a rollout batch size of 256 and a PPO minibatch size of 64, which is sufficient to reach stable reward values. For group (ii) and (iii), models are trained for 200 RL steps, respectively, using a rollout batch size of 384 and a PPO batch size of 128. For all RAPO training across three groups, we set the KL coefficient γ to 0.01, the window size w to 3 for Qwen3-VL-2B-Instruct and 1 for the others, and select reflection anchors as the top 20% highest-entropy tokens.

Evaluation Settings. We evaluate all models using the vLLM framework (Kwon et al., 2023) on six multimodal benchmarks spanning two categories. (1) **Reasoning-intensive (Reas.)** benchmarks assess multimodal reasoning ability and include MathVision (Wang et al., 2024a),

Table 1. **Main results (mean@8 acc %).** Performance comparison across reasoning-intensive and general-domain multimodal benchmarks. The best results in each group are highlighted in **bold**. The second best results in each group are highlighted in underlined. RAPO_G and RAPO_D denote RAPO built upon GRPO and DAPO, respectively.

Base Model	Model	Reasoning-Intensive				General-Domain		Avg.
		MathVision	MathVerse	EMMA	LogicVista	MMMU-Pro	RealWorldQA	
Qwen3-VL-8B Instruct	Base	41.74	50.21	38.34	62.28	37.59	67.20	49.56
	+ GRPO	48.03	55.23	41.34	59.51	48.50	67.66	53.38 (↑ 3.82)
	+ PAPO_G	43.75	54.57	39.00	64.29	50.00	69.41	53.50 (↑ 3.94)
	+ VPPO_G	48.03	56.35	36.00	**66.07**	49.16	67.48	53.85 (↑ 4.29)
	+ RAPO_G	**48.27**	**56.72**	**42.50**	62.50	**51.10**	**69.54**	**55.11** (↑ 5.55)
Qwen3-VL-2B Instruct	Base	27.30	36.75	26.87	47.13	24.44	60.47	37.16
	+ GRPO	28.29	44.35	28.50	48.71	30.61	64.28	40.79 (↑ 3.63)
	+ PAPO_G	28.29	45.18	27.25	46.43	31.45	64.05	40.44 (↑ 3.28)
	+ VPPO_G	30.59	45.43	28.75	45.76	31.73	63.14	40.90 (↑ 3.74)
	+ RAPO_G	**33.55**	**46.13**	**31.25**	**49.55**	**32.20**	**65.36**	**43.01** (↑ 5.85)
Qwen2.5-VL-7B Instruct	Base	21.05	29.44	23.25	42.19	21.22	54.51	31.94
	MM-Eureka	32.57	45.18	28.75	44.64	26.82	**61.96**	39.99 (↑ 8.05)
	VL-Rethinker	**34.21**	45.30	28.75	44.20	34.10	59.22	40.96 (↑ 9.02)
	R1-ShareVL	25.99	46.22	**30.50**	46.88	34.45	56.47	40.09 (↑ 8.15)
	Base+RAPO_G	27.96	47.34	29.00	45.31	36.01	60.26	40.98 (↑ 9.04)
	Base+RAPO_D	30.26	**47.59**	**30.50**	**47.13**	**36.04**	58.40	**41.65** (↑ 9.71)
Qwen2-VL-7B Instruct	Base	15.79	10.25	11.75	13.17	11.73	31.83	15.75
	TVC	18.75	21.98	20.75	**41.96**	23.34	52.88	29.94 (↑ 14.19)
	MINT-CoT	22.04	24.62	19.00	31.47	17.63	49.15	27.32 (↑ 11.57)
	Base+RAPO_G	22.37	32.99	25.25	37.72	28.22	64.67	35.21 (↑ 19.46)
	Base+RAPO_D	**23.36**	**36.31**	**28.75**	39.73	**29.39**	**65.01**	**37.09** (↑ 21.34)

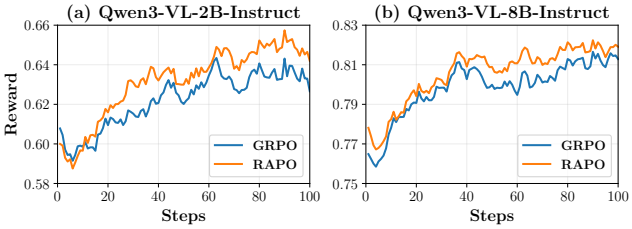


Figure 4. **Training dynamics.** Reward curves of GRPO and RAPO on Qwen3-VL-2B-Instruct and 8B during RL.

MathVerse (Zhang et al., 2024), EMMA (Hao et al., 2025) and LogicVista (Xiao et al., 2024). (2) **General-domain (Gen.)** benchmarks evaluate overall ability across diverse domains and include MMMU-Pro (Yue et al., 2025) and RealWorldQA[1]. Across all evaluations, we use a sampling temperature of 1.0, a maximum generation length of 8192 tokens and sample 8 responses per input, reporting final performance using the mean@8 accuracy metric following prior work (Wang et al., 2025b; Huang et al., 2025a).

5.2. Main Results (RQ1)

RAPO achieves the best overall performance in both controlled and same-backbone reference comparisons. We evaluate RAPO across three experimental settings covering different base models and benchmark types, as summarized in Table 1. Across all three experimental groups, RAPO achieves the best overall average performance across both reasoning-intensive and general-domain benchmarks. In the pure RL setting on Qwen3-VL-Instruct models, RAPO yields consistent gains over GRPO, PAPO, and VPPO; since Qwen3-VL is already a strong backbone, these gains are best viewed as controlled improvements under reduced headroom. Moreover, RAPO shows improvements across both reasoning-intensive and general-domain benchmarks, indicating that the gains are not driven by overfitting to specific tasks. Under a compute-matched setting on Qwen3-VL-2B-Instruct, extending GRPO from 100 to 120 steps improves the average score from 40.79 to 41.45, but remains below RAPO-100 at 43.01. This suggests that the improvement is not explained solely by additional wall-clock training compute. Detailed wall-clock comparisons are provided in Appendix F.

RAPO exhibits faster and more stable reward optimization during training. Figure 4 illustrates the accuracy reward trajectories on Qwen3-VL-2B-Instruct and Qwen3-VL-8B-Instruct. For both model scales, RAPO achieves faster reward growth and maintains higher reward values than GRPO across training steps, indicating improved optimization efficiency across different model sizes.

5.3. Empirical Analysis of RAPO's Mechanism (RQ2)

In this subsection, we empirically examine why RAPO improves long-chain visual reasoning by analyzing how visual

[1] https://huggingface.co/datasets/xai-org/RealworldQA

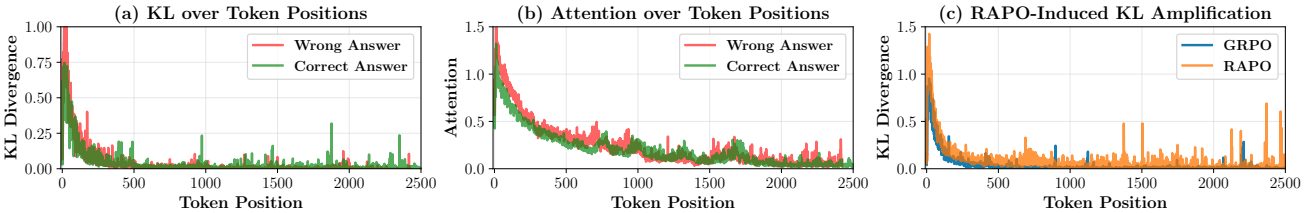
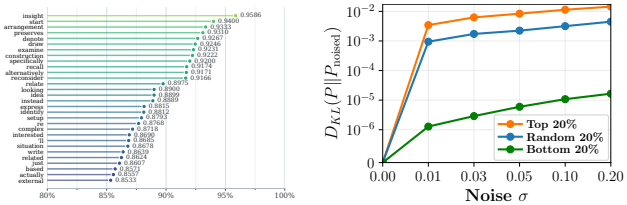


Figure 5. **Trajectory analysis. (a, b)** Visual dependence in correct vs. incorrect trajectories, measured by token-wise KL and attention; **(c)** Token-wise KL after GRPO and RAPO training. The KL at token t, $\overline{\mathbb{D}_{\mathrm{KL}t}}$, is the contrastive divergence between the distributions induced by vision-conditioned and vision-masked logits, with window length $w = 1$; higher values indicate stronger visual dependence.



(a) Anchor concentration. *(b)* D_{KL} vs. noise.

Figure 6. **Reflection anchor analysis. (a)** Fraction of each token type selected as top-entropy anchors. **(b)** KL divergence under visual noise across entropy groups.

information is used, where local intervention is effective, and how its influence propagates during generation. All analysis in this subsection are conducted using the same data and model configuration. We randomly sample 500 instances from the validation set of ViRL dataset and evaluate the Qwen3-VL-2B-Instruct model both before training and after GRPO and RAPO training from the main experiments. Additional experimental details are provided in Appendix D.

Sustained downstream visual dependence distinguishes successful reasoning and is reinforced by RAPO. We analyze token-wise contrastive KL along generated trajectories, defined as the divergence between vision-conditioned and vision-masked continuation distributions. Figure 5a shows that correct trajectories retain larger residual KL values and stronger late-stage fluctuations after the initial decay, whereas incorrect trajectories collapse more rapidly toward near-zero visual dependence. In contrast, visual-token attention, despite being a common proxy for visual use, follows similar positional trends for correct and incorrect outputs and fails to distinguish reasoning outcomes (Figure 5b). Figure 5c further shows that RAPO raises the token-wise KL profile over GRPO and maintains stronger late-stage KL signals. These results support the hypothesis that RAPO improves long-chain reasoning by maintaining stronger visual-dependence over the downstream continuation; unlike prior RL training where visual dependence remains an implicit property of the learned policy, RAPO explicitly reinforces this trajectory-level dependence beyond standard GRPO.

Reflection anchors provide interpretable compositional pivots. We analyze the positions selected by the entropy-based anchor rule. Figure 6a shows that high-entropy anchors concentrate on tokens that govern the next reasoning operation. These include explicit reasoning operators such as "examine," "reconsider," "identify," and "draw," as well as discourse and relational operators such as "actually," "instead," "based," and "specifically." These tokens are not merely frequent content words. They mark points where the model decides how the next operation should compose with the previous reasoning state: whether to inspect the image again, revise a hypothesis, introduce a contrast, or relate entities. This makes reflection anchors interpretable pivot points in the reasoning trajectory. Figure 6b further shows that these high-entropy positions are more sensitive to visual perturbations than random or low-entropy positions, indicating that they are not only linguistically procedural but visually leveraged. Consistently, restricting RAPO updates to the top-20% high-entropy anchors outperforms random selection, low-entropy selection, and alternative anchor budgets (Table 2(a)). These results support the central mechanism: RAPO improves multimodal reasoning by updating naturally occurring, interpretable operation-selection points rather than by injecting visual information at fixed positions.

Table 2. **Component ablation studies. (a)** Anchor-selection ablation. **(b)** Chain-mask ablation.

(a) Anchor-selection ablation.

Axis	Method	Reas.	Gen.	Avg.
–	GRPO	37.46	47.45	40.79
Quantity	top-10%	38.77	47.81	41.78
	top-20%	40.12	48.78	**43.01**
	top-50%	37.90	49.59	41.80
	top-100%	29.19	45.84	34.75
Position	top-20%	40.12	48.78	**43.01**
	bottom-20%	38.33	47.06	41.24
	random-20%	37.61	48.05	41.09

(b) Chain-mask ablation.

Mask	MathVis.	MathVer.	EMMA	LogicV.	MMMU-P.	RealWQ
Image	26.64	44.80	27.00	47.32	**32.79**	64.02
Chain	**33.55**	**46.13**	**31.25**	**49.55**	32.20	**65.36**

Anchor updates propagate visual evidence downstream. RAPO first amplifies visual dependence specifically at reflection anchors: the average contrastive KL at the top-20%

Table 3. **Ablation and robustness analysis. (a)** Ablation on the KL coefficient γ. **(b)** Ablation on KL window length w, i.e., the temporal range of visual-influence estimation. **(c)** Robustness across longer reasoning trajectories and different base model architectures.

(a) **Visual-Dependence Coefficient** γ

Method	Reas.	Gen.	Avg.
GRPO	37.46	47.45	40.79
RAPO$_G$ (γ=0)	38.59	47.19	41.46
RAPO$_G$ (γ=0.001)	39.66	47.23	42.18
RAPO$_G$ (γ=0.005)	38.61	47.70	41.64
RAPO$_G$ (γ=0.01)	40.12	48.78	43.01
RAPO$_G$ (γ=0.02)	37.68	48.14	41.16
RAPO$_G$ (γ=0.1)	Collapse		

(b) **Window Length** w

Method	Reas.	Gen.	Avg.
GRPO	37.46	47.45	40.79
RAPO$_G$ (w=1)	39.76	48.59	42.70
RAPO$_G$ (w=2)	38.52	47.93	41.66
RAPO$_G$ (w=3)	40.12	48.78	43.01
RAPO$_G$ (w=5)	38.24	48.93	41.80
RAPO$_G$ (w=10)	39.64	48.06	42.45

(c) **Base Model**

Model	Reas.	Gen.	Avg.
Thinking	38.48	41.41	39.45
+ GRPO	40.43	47.79	42.88
+ RAPO$_G$	44.04	47.82	45.30
Gemma3-4B-IT	15.86	26.99	19.57
+ GRPO	27.47	34.39	29.77
+ RAPO$_G$	28.62	33.67	30.30

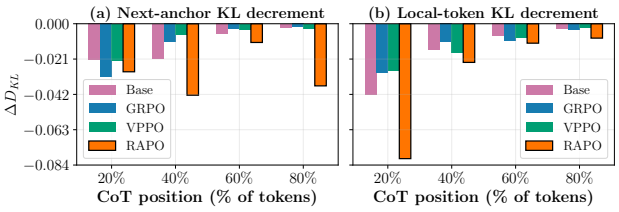


Figure 7. **Propagation of visual influence. (a)** KL change at the next anchor token. **(b)** Average KL change at subsequent local tokens under window length $w = 3$, measured by visual-token masking at different CoT positions.

entropy tokens increases from 0.30 to 0.55 after training, while the bottom-20% entropy tokens remain nearly unchanged. We then test whether this local amplification affects later reasoning by masking visual information at selected anchors and measuring the resulting KL decrement at the next anchor and nearby downstream tokens. Figure 7 shows that RAPO produces the largest downstream KL reductions, indicating that visual influence at an anchor persists beyond the local token and affects subsequent reasoning steps. The chain-mask ablation (Table 2(b)) further shows that replacing structured chain masking with blanket image-level masking lowers the average score from 43.01 to 40.43, suggesting that the propagation signal depends on preserving the structure of the reasoning chain. Together, these results turn RAPO from a black-box RL improvement into an interpretable mechanism for maintaining visual evidence through a composed reasoning trajectory.

5.4. Ablation and Robustness Analysis (RQ3)

We further evaluate RAPO through hyperparameter ablations and robustness checks. More experimental details and additional analysis are reported in Appendix G.

Hyperparameter Ablations. Table 3a shows that a small positive KL target coefficient γ improves over both GRPO and the $\gamma = 0$ variant, with the best result at $\gamma = 0.01$. Large γ over-optimizes contrastive visual divergence and can dominate the task reward, causing the policy to increase visual sensitivity without preserving answer correctness. Table 3b further studies the KL window length w, which controls the temporal range of measured visual influence. A moderate window $w = 3$ performs best: $w = 1$ captures only short-term effects, while larger windows $w \geq 5$ introduce noise from weakly coupled distant continuations.

Robustness. We evaluate RAPO under more capable reasoning models and different model architectures. On Qwen3-VL-2B-Thinking, which generates substantially longer CoT sequences than Qwen3-VL-2B-Instruct ($\sim$6k vs. $\sim$2k tokens) with a maximum generation length of 16,384 tokens, RAPO remains effective under long-CoT regimes and shows pronounced gains on reasoning-intensive benchmarks (Table 3c). We further evaluate RAPO on Gemma3-4B-IT (Team et al., 2025a), whose architecture differs from the Qwen series used in the main results. RAPO outperforms GRPO (Table 3c), indicating that its effectiveness is not tied to a specific reasoning length or model family.

6. Conclusion

We studied long-chain multimodal reasoning as a compositional process in which visual evidence must remain coupled to a sequence of intermediate reasoning operations. From this view, visual forgetting is a compositional grounding failure: the model continues to generate coherent reasoning while the later operations lose dependence on the image. We proposed **RAPO**, a GRPO-based method that performs sparse policy updates at high-entropy reflection anchors and optimizes a chain-masked finite-window objective for downstream visual dependence. RAPO requires no architectural change and no inference-time visual re-injection. Across multiple LVLM backbones and benchmarks, RAPO improves performance over strong baselines. More importantly for compositional learning, our mechanism analyses show that reflection anchors align with interpretable operation-selection tokens and that visual influence from these anchors propagates into downstream reasoning. These results suggest that sparse, interpretable intervention points can help make multimodal reasoning policies more compositionally grounded. Future work should extend this analysis to broader agentic settings, stronger safety stress tests, and more direct measurements of systematic out-of-distribution generalization.

Impact Statement

This work aims to improve the reasoning capability of large vision–language models by mitigating visual forgetting in long chain-of-thought generation. By enabling more reliable and sustained visual grounding, our approach has the potential to benefit applications that require accurate multimodal reasoning, such as scientific analysis, education, accessibility tools, and visual decision support systems.

At the same time, improvements in multimodal reasoning may amplify the impact of downstream misuse, including over-reliance on automated systems or the generation of convincing but incorrect multimodal explanations. While RAPO does not introduce new modalities, data sources, or architectural changes, and operates purely at the optimization level, careful evaluation and responsible deployment remain important, particularly in high-stakes or safety-critical settings.

We believe that making visual reasoning more faithful and interpretable contributes positively to the development of trustworthy multimodal AI systems.

References

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Chen, X., Zhang, R., Jiang, D., Zhou, A., Yan, S., Lin, W., and Li, H. MINT-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=vMpvtSmtXY.

Cheng, Z., Chen, Q., Xu, X., WANG, J., Wang, W., Fei, H., Wang, Y., Wang, A. J., Chen, Z., Che, W., and Qin, L. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=xPcKmKSEis.

Chu, X., Chen, X., Wang, G., Tan, Z., Huang, K., Lv, W., Mo, T., and Li, W. Qwen look again: Guiding vision-language reasoning models to re-attention visual information. *arXiv preprint arXiv:2505.23558*, 2025.

Chung, J., Kim, J., Kim, S., Lee, J., Kim, M. S., and Yu, Y. v1: Learning to point visual tokens for multimodal grounded reasoning, 2025. URL https://arxiv.org/abs/2505.18842.

Gao, J., Li, Y., Cao, Z., and Li, W. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19520–19529, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Hao, Y., Gu, J., Wang, H. W., Li, L., Yang, Z., Wang, L., and

Cheng, Y. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, 2025. URL https://arxiv.org/abs/2501.05444.

Huang, S., Qu, X., Li, Y., Luo, Y., He, Z., Liu, D., and Cheng, Y. Spotlight on token perception for multimodal reinforcement learning, 2025a. URL https://arxiv.org/abs/2510.09285.

Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., and Lin, S. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Jian, P., Wu, J., Sun, W., Wang, C., Ren, S., and Zhang, J. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9262–9281, 2025.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, W., Tang, R., Li, C., Zhang, C., Vulic, I., and Søgaard, A. Lost in embeddings: Information loss in vision-language models. *arXiv preprint arXiv:2509.11986*, 2, 2025.

Li, Y., Kuang, H., Zhang, H., Cao, J., Liu, Z., Hou, Q., and Cheng, M.-M. Rethinking token-level policy optimization for multimodal chain-of-thought, 2026. URL https://arxiv.org/abs/2603.22847.

Masry, A., Rodriguez, J. A., Zhang, T., Wang, S., Wang, C., Feizi, A., Suresh, A. K., Puri, A., Jian, X., Noël, P.-A., Madhusudhan, S. T., Pedersoli, M., Liu, B., Chapados, N., Bengio, Y., Hoque, E., Pal, C., Laradji, I. H., Vazquez, D., Taslakian, P., Gella, S., and Rajeswar, S. Alignvlm: Bridging vision and language latent spaces for multimodal document understanding, 2025. URL https://arxiv.org/abs/2502.01341.

Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., Zhang, K., Luo, P., Qiao, Y., Zhang, Q., and Shao, W. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

Qin, Y., Wei, B., Ge, J., Kallidromitis, K., Fu, S., Darrell, T., and Wang, X. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sun, G., Hua, H., Wang, J., Luo, J., Dianat, S., Rabbani, M., Rao, R., and Tao, Z. Latent chain-of-thought for visual reasoning, 2025a. URL https://arxiv.org/abs/2510.23925.

Sun, H.-L., Sun, Z., Peng, H., and Ye, H.-J. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. *arXiv preprint arXiv:2503.13360*, 2025b.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby,

R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025a. URL https://arxiv.org/abs/2503.19786.

Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., Chen, Z., Cui, J., Ding, H., Dong, M., Du, A., Du, C., Du, D., Du, Y., Fan, Y., Feng, Y., Fu, K., Gao, B., Gao, H., Gao, P., Gao, T., Gu, X., Guan, L., Guo, H., Guo, J., Hu, H., Hao, X., He, T., He, W., He, W., Hong, C., Hu, Y., Hu, Z., Huang, W., Huang, Z., Huang, Z., Jiang, T., Jiang, Z., Jin, X., Kang, Y., Lai, G., Li, C., Li, F., Li, H., Li, M., Li, W., Li, Y., Li, Y., Li, Z., Li, Z., Lin, H., Lin, X., Lin, Z., Liu, C., Liu, C., Liu, H., Liu, J., Liu, J., Liu, L., Liu, S., Liu, T. Y., Liu, T., Liu, W., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Z., Lu, E., Lu, L., Ma, S., Ma, X., Ma, Y., Mao, S., Mei, J., Men, X., Miao, Y., Pan, S., Peng, Y., Qin, R., Qu, B., Shang, Z., Shi, L., Shi, S., Song, F., Su, J., Su, Z., Sun, X., Sung, F., Tang, H., Tao, J., Teng, Q., Wang, C., Wang, D., Wang, F., Wang, H., Wang, J., Wang, J., Wang, J., Wang, S., Wang, S., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Z., Wang, Z., Wang, Z., Wei, C., Wei, Q., Wu, W., Wu, X., Wu, Y., Xiao, C., Xie, X., Xiong, W., Xu, B., Xu, J., Xu, J., Xu, L. H., Xu, L., Xu, S., Xu, W., Xu, X., Xu, Y., Xu, Z., Yan, J., Yan, Y., Yang, X., Yang, Y., Yang, Z., Yang, Z., Yang, Z., Yao, H., Yao, X., Ye, W., Ye, Z., Yin, B., Yu, L., Yuan, E., Yuan, H., Yuan, M., Zhan, H., Zhang, D., Zhang, H., Zhang, W., Zhang, X., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Z., Zhao, H., Zhao, Y., Zheng, H., Zheng, S., Zhou, J., Zhou, X., Zhou, Z., Zhu, Z., Zhuang, W., and Zu, X. Kimi k2: Open agentic intelligence, 2025b. URL https://arxiv.org/abs/2507.20534.

Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., Wang, D., Yuan, E., Lu, E., Li, F., Sung, F., Wei, G., Lai, G., Zhu, H., Ding, H., Hu, H., Yang, H., Zhang, H., Wu, H., Yao, H., Lu, H., Wang, H., Gao, H., Zheng, H., Li, J., Su, J., Wang, J., Deng, J., Qiu, J., Xie, J., Wang, J., Liu, J., Yan, J., Ouyang, K., Chen, L., Sui, L., Yu, L., Dong, M., Dong, M., Xu, N., Cheng, P., Gu, Q., Zhou, R., Liu, S., Cao, S., Yu, T., Song, T., Bai, T., Song, W., He, W., Huang, W., Xu, W., Yuan, X., Yao, X., Wu, X., Li, X., Zu, X., Zhou, X., Wang, X., Charles, Y., Zhong, Y., Li, Y., Hu, Y., Chen, Y., Wang, Y., Liu, Y., Miao, Y., Qin, Y., Chen, Y., Bao, Y., Wang, Y., Kang, Y., Liu, Y., Dong, Y., Du, Y., Wu, Y., Wang, Y., Yan, Y., Zhou, Z., Li, Z., Jiang, Z., Zhang, Z., Yang, Z., Huang, Z., Huang, Z., Zhao, Z., Chen, Z., and Lin, Z. Kimi-vl technical report, 2025c. URL https://arxiv.org/abs/2504.07491.

Tian, X., Zou, S., Yang, Z., He, M., Waschkowski, F., Wesemann, L., Tu, P. H., and Zhang, J. More thought, less accuracy? on the dual nature of reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XpL5eqjCjF.

Wan, Z., Dou, Z., Liu, C., Zhang, Y., Cui, D., Zhao, Q., Shen, H., Xiong, J., Xin, Y., Jiang, Y., et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*, 2025.

Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., and Chen, W. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning, 2025a. URL https://arxiv.org/abs/2504.08837.

Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., and Li, H. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024a.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Wang, Z., Guo, X., Stoica, S., Xu, H., Wang, H., Ha, H., Chen, X., Chen, Y., Yan, M., Huang, F., and Ji, H. Perception-aware policy optimization for multimodal reasoning, 2025b. URL https://arxiv.org/abs/2507.06448.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Xiao, Y., Sun, E., Liu, T., and Wang, W. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024. URL https://arxiv.org/abs/2407.04973.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

Yao, H., Yin, Q., Zhang, J., Yang, M., Wang, Y., Wu, W., Su, F., Shen, L., Qiu, M., Tao, D., and Huang, J. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo, 2025. URL https://arxiv.org/abs/2505.16673.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., YuYue, Dai, W., Fan, T., Liu, G., Liu, J., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, R., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.-Y., Zhang, Y.-Q., Yan, L., Wu, Y., and Wang, M. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., Su, Y., Chen, W., and Neubig, G. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. URL https://arxiv.org/abs/2409.02813.

Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.-W., Gao, P., and Li, H. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. URL https://arxiv.org/abs/2403.14624.

Zhang, Y., Ding, Y., Zhang, S., Zhang, X., Li, H., Li, Z.-z., Wang, P., Wu, J., Ji, L., Shen, Y., et al. Perl: Permutation-enhanced reinforcement learning for interleaved vision-language reasoning. *arXiv preprint arXiv:2506.14907*, 2025.

A. Related Works

A.1. Mitigating Visual Forgetting in long CoT Reasoning

A key challenge in applying Chain-of-Thought (CoT) reasoning to MLLMs is *visual forgetting*, where reliance on initial visual inputs progressively decays as reasoning chains lengthen, often leading to ungrounded hallucinations. To address this, recent work explores modality interleaving, explicit visual re-attention during inference, and visual self-reflection mechanisms. One line of work seeks to prevent visual attention decay by structurally embedding visual evidence throughout the reasoning process. Interleaved-Modal Chain-of-Thought (ICoT) (Gao et al., 2025) interleaves paired visual-textual rationales with intermediate reasoning steps, maintaining access to visual evidence during multi-hop inference. Similarly, MINT-CoT (Chen et al., 2025) employs interleaved image-text sequences with dynamic visual token selection to support long-context reasoning. From a representation perspective, Chain-of-Visual-Thought (CoVT) (Qin et al., 2025) enables reasoning in a continuous visual token space trained with dense perceptual supervision, thereby strengthening spatial and geometric grounding over extended reasoning horizons. Visual Thoughts (Cheng et al., 2026) provides a unified view of multimodal CoT by showing that both textual and interleaved MCoT can benefit from intermediate "visual thoughts," which convey image information through different expression forms and act as intermediaries between the input image and deeper reasoning layers. Latent Chain-of-Thought (Sun et al., 2025a) instead treats visual reasoning as inference over latent CoT variables, using amortized variational inference and sparse token-level rewards to encourage diverse, high-likelihood latent rationales. Another line of approaches emphasize inference-time re-attention to visual inputs. TVC (Sun et al., 2025b) mitigates visual forgetting by reintroducing image conditioning at critical reasoning stages and dynamically pruning redundant visual tokens, effectively preserving visual context over long chains. In parallel, v1 (Chung et al., 2025) proposes a *point-and-copy* mechanism that selects relevant image patches and injects their representations into the reasoning sequence, ensuring continued access to visual evidence. Recent studies further incorporate explicit visual self-reflection and reinforcement learning to address hallucinations induced by visual forgetting. Qwen-LA (Chu et al., 2025) introduces a vision-text reflection process and optimizes the timing of visual re-examination via reinforcement learning, complemented by visual token routing. Reflection-V (Jian et al., 2025) constructs vision-centered reasoning data for cold-start training and employs a visual-attention-based reward model to encourage reasoning trajectories that remain grounded in image evidence.

A.2. Multimodal Reinforcement Learning

Recent work has extended reinforcement learning (RL) techniques from large language models (LLMs) to multimodal large language models (MLLMs) to improve complex reasoning. Foundational policy optimization relies on algorithms such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which establishes a baseline for group-based advantage estimation without value functions. Building on this, Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) (Yu et al., 2025) addresses training instability and entropy collapse in long-chain reasoning through decoupled clipping and dynamic sampling mechanisms. Building on these frameworks, several studies aim to elicit robust reasoning behaviors in MLLMs. Vision-R1 (Huang et al., 2025b) proposes *progressive thinking suppression training* to reduce overthinking during cold starts, encouraging efficient reasoning trajectories. Similarly, MM-Eureka (Meng et al., 2025) extends rule-based RL to the multimodal domain, inducing emergent self-correction and reasoning patterns comparable to text-only models. To further improve reasoning quality, recent methods incorporate explicit self-reflection into the RL loop. VL-Rethinker (Wang et al., 2025a) addresses the vanishing advantage problem via *selective sample replay* and a F*orced Rethinking* trigger. SRPO (Self-Reflection enhanced reasoning with Group Relative Policy Optimization) (Wan et al., 2025) applies a two-stage framework that trains models to generate and refine their own reflections, rewarding instructive self-corrections. Adopting a tree-search perspective, Mulberry (Yao et al., 2024) leverages *Collective Monte Carlo Tree Search* to construct high-quality reasoning paths for training reasoning-intensive MLLMs. Beyond general reasoning, increasing focus has been placed on visual grounding in RL. PAPO (Perception-Aware Policy Optimization) (Wang et al., 2025b) introduces an *Implicit Perception Loss* to anchor reasoning steps to visual inputs, thereby reducing perception errors. VPPO (Visually-Perceptive Policy Optimization) (Huang et al., 2025a) further refines this by analyzing token-level visual dependency to prioritize perceptually pivotal tokens. Extending to more complex inputs, PeRL (Zhang et al., 2025) utilizes a permutation-enhanced RL framework for interleaved vision–language tasks, improving multi-image reasoning through rollout filtering and position-aware exploration.

B. Theoretical Results

Throughout this appendix, $\mathbb{D}_{\mathrm{KL}}(\cdot\|\cdot)$ denotes $D_{\mathrm{KL}}(\cdot\|\cdot)$, and

$$[a]_+ = \max\{a, 0\}.$$

All discrete entropies are Shannon entropies and are denoted by $\mathcal{H}(\cdot)$. We follow the notation convention in the main text: uppercase letters denote random variables and lowercase letters denote realizations.

This appendix justifies the sparse one-step-then-rollback objective in Eq. (1). The selected set $\mathcal{A}$ affects the masked history at step t only through its earlier elements

$$\mathcal{A}_{<t} = \mathcal{A} \cap [t-1].$$

Thus, the ideal local gain at step t is set-dependent: changing earlier selected positions changes the masked context $\widetilde{h}_t(\mathcal{A}_{<t})$, and hence changes both the local next-token marginal and the masked-reference suffix law.

The analysis is local in the following sense. For each selected step t, we evaluate a one-step intervention that changes only the next-token marginal and then rolls back to the native decoder for the future suffix. The resulting quantity is a marginal downstream visual gain. The joint sparse objective sums these marginal gains over selected steps; it is not the mutual information of a full rollout in which all selected steps are intervened on simultaneously.

When an optimizer need not exist, all occurrences of $\max$ can be read as $\sup$. The proofs below are unchanged under this convention.

B.1. Problem Setup

Let V be the visual input, X the textual prompt, and $Y_{1:T}$ the reasoning trajectory. Let $[T] = \{1, \ldots, T\}$. The native decoder p_θ generates

$$p_\theta(Y_{1:T} \mid V, X) = \prod_{s=1}^{T} p_\theta(Y_s \mid V, X, Y_{<s}).$$

For a realized native trajectory, write

$$h_t = (x, y_{<t}), \qquad Y_{\geq t} = (Y_t, \ldots, Y_T), \qquad Y_{>t} = (Y_{t+1}, \ldots, Y_T).$$

Fix a candidate selected set $\mathcal{A} \subseteq [T]$. Let

$$\mathcal{A}_{<t} = \mathcal{A} \cap [t-1],$$

and define the realized masked history

$$\widetilde{h}_t(\mathcal{A}_{<t}) = \big(x, M_t(y_{<t}; \mathcal{A}_{<t})\big).$$

When $\mathcal{A}$ and t are clear from context, we write $\widetilde{h}_t$ instead of $\widetilde{h}_t(\mathcal{A}_{<t})$. All conditional quantities below are evaluated conditioned on this realized masked history. Thus, $V \mid \widetilde{h}_t$ denotes the population distribution of visual inputs compatible with the masked context $\widetilde{h}_t$.

For a visual realization v, define the masked-context native next-token marginal

$$m_{t,v}(y) := p_\theta(y \mid V = v, \widetilde{h}_t).$$

Define the native masked-history rollback continuation by

$$K_{t,v}(y_{>t} \mid y) := p_\theta(y_{>t} \mid V = v, \widetilde{h}_t, Y_t = y).$$

The corresponding native masked-history suffix law is

$$\widetilde{p}_{t,v}(y, y_{>t}) = m_{t,v}(y) K_{t,v}(y_{>t} \mid y). \tag{2}$$

A local reshaping replaces only the next-token marginal $m_{t,v}$ by

$$q_t(\cdot \mid V = v, \widetilde{h}_t),$$

while keeping the native rollback continuation $K_{t,v}(\cdot \mid y)$ unchanged. The intervened suffix law is

$$\widetilde{p}^{q_t}_{t,v}(y, y_{>t}) = q_t(y \mid V = v, \widetilde{h}_t) K_{t,v}(y_{>t} \mid y). \tag{3}$$

The downstream visual gain of this one-step intervention is

$$\Delta_t(q_t; \mathcal{A}_{<t}) = I_{q_t}(V; Y_{\geq t} \mid \widetilde{h}_t) - I_{p_\theta}(V; Y_{\geq t} \mid \widetilde{h}_t), \tag{4}$$

where I_{q_t} is computed under the one-step-then-rollback rollout in Eq. (3), and the native term is computed under Eq. (2).

The local feasible set is

$$\mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon) = \left\{ q_t : \mathbb{D}_{\mathrm{KL}}\Big(q_t(\cdot \mid V, \widetilde{h}_t) \,\Big\|\, m_{t,V} \Big) \leq \varepsilon \quad \text{for } V \mid \widetilde{h}_t\text{-a.e.} \right\}. \tag{5}$$

The local optimal downstream gain is

$$\Delta^\star_t(\mathcal{A}_{<t}, \varepsilon) = \sup_{q_t \in \mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon)} \Delta_t(q_t; \mathcal{A}_{<t}). \tag{6}$$

We also define the full-history native next-token distribution, evaluated at the same realized full prefix $h_t = (x, y_{<t})$, by

$$\pi^{\mathrm{full}}_{t,v}(y) := p_\theta(y \mid V = v, h_t),$$

and the full-history native continuation by

$$K^{\mathrm{full}}_{t,v}(y_{>t} \mid y) := p_\theta(y_{>t} \mid V = v, h_t, Y_t = y).$$

The full-history suffix law is

$$p_{t,v}(y, y_{>t}) = \pi^{\mathrm{full}}_{t,v}(y) K^{\mathrm{full}}_{t,v}(y_{>t} \mid y). \tag{7}$$

The asymmetry between $p_{t,v}$ and the masked-reference suffix law below is intentional: $p_{t,v}$ measures the visual separability of the realized full-history continuation, while the masked-reference law removes visual evidence directly leaked through previously selected positions.

Let

$$K_\rho = \lfloor \rho T \rfloor.$$

For a selected set $\mathcal{A}$, define the decoder-feasible class

$$\Pi(\mathcal{A}, \varepsilon) = \left\{ \pi : \pi(\cdot \mid V, \widetilde{h}_t(\mathcal{A}_{<t})) \in \mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon) \quad \forall t \in \mathcal{A} \right\}.$$

For $\pi \in \Pi(\mathcal{A}, \varepsilon)$, write

$$\Delta_t(\pi; \mathcal{A}_{<t}) = \Delta_t(q^\pi_t; \mathcal{A}_{<t}), \qquad q^\pi_t(\cdot \mid V, \widetilde{h}_t) = \pi(\cdot \mid V, \widetilde{h}_t).$$

The joint sparse oracle objective corresponding to Eq. (1) is

$$\mathcal{J}^\star(\varepsilon) = \max_{\mathcal{A} \subseteq [T],\, |\mathcal{A}| \leq K_\rho} \; \sup_{\pi \in \Pi(\mathcal{A}, \varepsilon)} \sum_{t \in \mathcal{A}} \Delta_t(\pi; \mathcal{A}_{<t}). \tag{8}$$

B.2. Additional Necessary Notations

Fix a candidate set $\mathcal{A}$, a step t, and the induced masked history

$$\widetilde{h}_t = \widetilde{h}_t(\mathcal{A}_{<t}).$$

Masked-reference suffix law. Define the masked-reference suffix distribution by

$$p_{\mathcal{M},t}(y_{\geq t}) = \mathbb{E}_{V|\widetilde{h}_t}\left[\widetilde{p}_{t,V}(y_{\geq t})\right]. \tag{9}$$

Equivalently,

$$p_{\mathcal{M},t}(y, y_{>t}) = \mathbb{E}_{V|\widetilde{h}_t}\left[\widetilde{p}_{t,V}(y, y_{>t})\right].$$

Its next-token marginal is denoted by

$$m_{\mathcal{M},t}(y) = \sum_{y_{>t}} p_{\mathcal{M},t}(y, y_{>t}).$$

For an intervened next-token distribution q_t, define the corresponding intervened masked-reference suffix law by

$$p^{q_t}_{\mathcal{M},t}(y, y_{>t}) = \mathbb{E}_{V|\widetilde{h}_t}\left[\widetilde{p}^{q_t}_{t,V}(y, y_{>t})\right].$$

Downstream visual score. Define the pointwise masked-history-to-reference log-likelihood ratio

$$\ell_t(v, y_{\geq t}) = \log \frac{\widetilde{p}_{t,v}(y_{\geq t})}{p_{\mathcal{M},t}(y_{\geq t})}.$$

The downstream visual score of choosing token y at step t is

$$\psi_t(v, y) = \mathbb{E}_{Y_{>t}\sim K_{t,v}(\cdot|y)}\left[\ell_t(v, (y, Y_{>t}))\right]. \tag{10}$$

This score measures how distinguishable the future suffix induced by choosing y remains from the masked-reference suffix distribution.

Claim B.1. For any y with $m_{t,v}(y) > 0$,

$$\psi_t(v, y) = \log \frac{m_{t,v}(y)}{m_{\mathcal{M},t}(y)} + \mathbb{D}_{\mathrm{KL}}(K_{t,v}(\cdot \mid y) \,\|\, p_{\mathcal{M},t}(\cdot \mid y)). \tag{11}$$

Claim B.2. Under the native masked-history one-step-then-rollback suffix law $\widetilde{p}_{t,V}$,

$$I_{p_\theta}(V; Y_{\geq t} \mid \widetilde{h}_t) = \mathbb{E}_{V|\widetilde{h}_t}\mathbb{E}_{Y_t\sim m_{t,V}}\left[\psi_t(V, Y_t)\right]. \tag{12}$$

B.3. Assumptions

Assumption B.1 (Finite common support and absolute continuity)**.** For each step t, all next-token distributions are supported on a finite set $\mathcal{Y}_t$. Since the horizon T is finite, the corresponding suffix space is also finite. For every candidate set $\mathcal{A}$, every step t, and $V \mid \widetilde{h}_t(\mathcal{A}_{<t})$-almost every v, the distributions $\widetilde{p}_{t,v}$, $\widetilde{p}^{q_t}_{t,v}$, $p_{t,v}$, and $p_{\mathcal{M},t}$ are mutually compatible in the following sense: whenever a likelihood ratio or KL divergence appears below, the numerator is absolutely continuous with respect to the denominator. In particular,

$$q_t(\cdot \mid v, \widetilde{h}_t) \ll m_{t,v}$$

for every feasible q_t.

Remark B.1 (Role of Assumption B.1). This assumption avoids support pathologies and ensures that the KL divergences, likelihood ratios, and exponential-tilting perturbations are well-defined. For a finite-vocabulary decoder with strictly positive softmax probabilities, the next-token part of this assumption is automatic. For suffix distributions, it can be enforced by restricting the analysis to the common support used for scoring.

Assumption B.2 (Bounded reference ratios and bounded full-history gap)**.** There exists $0 < B < \infty$ such that, for every candidate set $\mathcal{A}$, every step t, and every $(v, y, y_{>t})$ with $\widetilde{p}_{t,v}(y, y_{>t}) > 0$,

$$\left|\log \frac{\widetilde{p}_{t,v}(y, y_{>t})}{p_{\mathcal{M},t}(y, y_{>t})}\right| \leq B.$$

Moreover, the full-history suffix gap is bounded by the same constant:

$$0 \leq \mathbb{D}_{\mathrm{KL}}(p_{t,v} \| p_{\mathcal{M},t}) \leq B \qquad \text{for } V \mid \widetilde{h}_t(\mathcal{A}_{<t})\text{-a.e. } v.$$

Remark B.2 (Role of Assumption B.2). The first bound ensures that the downstream score ψ_t is uniformly bounded, which is needed for the KL trust-region expansion. The second bound ensures that the full-history reference gap $\mathbb{D}_{\mathrm{KL}}(p_{t,V} \| p_{\mathcal{M},t})$ is finite and can be used as a multiplicative reference-gap factor in the entropy lower bound. Both are standard bounded-likelihood-ratio conditions. In practice, the same condition can be imposed on the truncated candidate support used for scoring.

Assumption B.3 (Generic non-coincidence)**.** There exists $\gamma > 0$ such that, for every candidate set $\mathcal{A}$, every step t, and $V \mid \widetilde{h}_t(\mathcal{A}_{<t})$-almost surely, the score map

$$y \mapsto \psi_t(V, y)$$

is injective on the support of $m_{t,V}$ and γ-separated. That is, for any $y \neq y'$ with $m_{t,V}(y)m_{t,V}(y') > 0$,

$$|\psi_t(V, y) - \psi_t(V, y')| \geq \gamma.$$

Remark B.3 (Role of Assumption B.3). Exact score ties are a measure-zero degeneracy. In our setting, $\psi_t(V, y)$ is a real-valued functional of logits and continuation likelihoods. Under non-degenerate perturbations, distinct candidate tokens almost surely have distinct downstream scores. The separation constant γ formalizes finite numerical resolution. Since Assumption B.2 also bounds $|\psi_t|$, this assumption is best interpreted on the finite candidate set used by the reshaping or scoring procedure, rather than as a claim about an arbitrarily large full vocabulary.

Assumption B.4 (Local target realizability)**.** For every candidate set $\mathcal{A}$ and every collection of feasible local reshaping targets

$$\{q_t : t \in \mathcal{A},\ q_t \in \mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon)\},$$

there exists a decoder $\pi \in \Pi(\mathcal{A}, \varepsilon)$ such that, for every $t \in \mathcal{A}$,

$$\pi(\cdot \mid V, \widetilde{h}_t(\mathcal{A}_{<t})) = q_t(\cdot \mid V, \widetilde{h}_t(\mathcal{A}_{<t})) \quad \text{for } V \mid \widetilde{h}_t(\mathcal{A}_{<t})\text{-a.e.}$$

Remark B.4 (Role of Assumption B.4). This is an oracle target-construction assumption. The theory identifies local reshaping targets at selected masked histories, and training a single decoder π attempts to approximate those targets. If exact realizability is replaced by approximation up to error δ_{real} in downstream gain per selected step, the final joint lower bound loses at most $K_\rho \delta_{\mathrm{real}}$.

B.4. Local Technical Lemmas

Lemma B.1 (Exact first-order decomposition and one-sided lower bound)**.** *Fix a candidate set $\mathcal{A}$, a step t, and the induced masked history $\widetilde{h}_t = \widetilde{h}_t(\mathcal{A}_{<t})$. Assume Assumption B.1. For any $q_t \in \mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon)$, the downstream visual gain admits the exact decomposition*

$$\Delta_t(q_t; \mathcal{A}_{<t}) = G_{1,t}(q_t; \mathcal{A}_{<t}) + \mathbb{E}_{V|\widetilde{h}_t}\left[\mathbb{D}_{\mathrm{KL}}\Big(q_t(\cdot \mid V, \widetilde{h}_t) \,\Big\|\, m_{t,V}\Big)\right] - \mathbb{D}_{\mathrm{KL}}\Big(p^{q_t}_{\mathcal{M},t} \,\Big\|\, p_{\mathcal{M},t}\Big), \tag{13}$$

where

$$G_{1,t}(q_t; \mathcal{A}_{<t}) = \mathbb{E}_{V|\widetilde{h}_t}\left[\sum_y \Big(q_t(y \mid V, \widetilde{h}_t) - m_{t,V}(y)\Big)\, \psi_t(V, y)\right]. \tag{14}$$

Moreover,

$$\Delta_t(q_t; \mathcal{A}_{<t}) \geq G_{1,t}(q_t; \mathcal{A}_{<t}). \tag{15}$$

Proof. Throughout the proof, write

$$\mathbb{E}[\cdot] = \mathbb{E}_{V|\widetilde{h}_t}[\cdot], \qquad q_V(y) = q_t(y \mid V, \widetilde{h}_t).$$

Under the one-step intervention,

$$\widetilde{p}^{q_t}_{t,V}(y, y_{>t}) = q_V(y) K_{t,V}(y_{>t} \mid y),$$

and

$$p^{q_t}_{\mathcal{M},t}(y, y_{>t}) = \mathbb{E}\left[\widetilde{p}^{q_t}_{t,V}(y, y_{>t})\right].$$

Thus

$$I_{q_t}(V; Y_{\geq t} \mid \widetilde{h}_t) = \mathbb{E}\left[\sum_{y, y_{>t}} \widetilde{p}^{q_t}_{t,V}(y, y_{>t}) \log \frac{\widetilde{p}^{q_t}_{t,V}(y, y_{>t})}{p^{q_t}_{\mathcal{M},t}(y, y_{>t})}\right].$$

Insert the native masked-history ratio:

$$\frac{\widetilde{p}^{q_t}_{t,V}(y, y_{>t})}{p^{q_t}_{\mathcal{M},t}(y, y_{>t})} = \frac{\widetilde{p}_{t,V}(y, y_{>t})}{p_{\mathcal{M},t}(y, y_{>t})} \cdot \frac{q_V(y)}{m_{t,V}(y)} \cdot \frac{p_{\mathcal{M},t}(y, y_{>t})}{p^{q_t}_{\mathcal{M},t}(y, y_{>t})}.$$

Splitting the logarithm gives three terms. The first term equals

$$\mathbb{E}\left[\sum_y q_V(y)\psi_t(V, y)\right].$$

The second term equals

$$\mathbb{E}\left[\mathbb{D}_{\mathrm{KL}}(q_V \| m_{t,V})\right].$$

The third term equals

$$-\,\mathbb{D}_{\mathrm{KL}}(p^{q_t}_{\mathcal{M},t} \| p_{\mathcal{M},t}).$$

Therefore,

$$I_{q_t}(V; Y_{\geq t} \mid \widetilde{h}_t) = \mathbb{E}\left[\sum_y q_V(y)\psi_t(V, y)\right] + \mathbb{E}\left[\mathbb{D}_{\mathrm{KL}}(q_V \| m_{t,V})\right] - \mathbb{D}_{\mathrm{KL}}(p^{q_t}_{\mathcal{M},t} \| p_{\mathcal{M},t}).$$

Under the native choice, $q_V = m_{t,V}$ and $p^{q_t}_{\mathcal{M},t} = p_{\mathcal{M},t}$. By Claim B.2,

$$I_{p_\theta}(V; Y_{\geq t} \mid \widetilde{h}_t) = \mathbb{E}\left[\sum_y m_{t,V}(y)\psi_t(V, y)\right].$$

Subtracting proves Eq. (13).

It remains to show the one-sided bound. By joint convexity of KL divergence,

$$\mathbb{D}_{\mathrm{KL}}(p^{q_t}_{\mathcal{M},t} \| p_{\mathcal{M},t}) = \mathbb{D}_{\mathrm{KL}}\Big(\mathbb{E}[\widetilde{p}^{q_t}_{t,V}] \,\Big\|\, \mathbb{E}[\widetilde{p}_{t,V}]\Big) \leq \mathbb{E}\left[\mathbb{D}_{\mathrm{KL}}(\widetilde{p}^{q_t}_{t,V} \| \widetilde{p}_{t,V})\right].$$

Because the rollback continuation $K_{t,V}(\cdot \mid y)$ is the same in $\widetilde{p}^{q_t}_{t,V}$ and $\widetilde{p}_{t,V}$, the continuation cancels in the inner KL:

$$\mathbb{D}_{\mathrm{KL}}(\widetilde{p}^{q_t}_{t,V} \| \widetilde{p}_{t,V}) = \mathbb{D}_{\mathrm{KL}}(q_V \| m_{t,V}).$$

Therefore,

$$\mathbb{D}_{\mathrm{KL}}(p^{q_t}_{\mathcal{M},t} \| p_{\mathcal{M},t}) \leq \mathbb{E}\left[\mathbb{D}_{\mathrm{KL}}(q_V \| m_{t,V})\right].$$

Substituting this inequality into the exact decomposition yields

$$\Delta_t(q_t; \mathcal{A}_{<t}) \geq G_{1,t}(q_t; \mathcal{A}_{<t}).$$

□

Lemma B.2 (Optimal first-order perturbation and variance lower bound)**.** *Fix a candidate set $\mathcal{A}$, a step t, and the induced masked history $\widetilde{h}_t = \widetilde{h}_t(\mathcal{A}_{<t})$. Assume Assumptions B.1 and B.2. For each visual realization v, define*

$$\mu_t(v) = \sum_y m_{t,v}(y)\psi_t(v, y), \qquad \widetilde{\psi}_t(v, y) = \psi_t(v, y) - \mu_t(v),$$

and

$$\sigma_t(v) = \sqrt{\mathrm{Var}_{Y \sim m_{t,v}}\big(\psi_t(v, Y)\big)}.$$

Define the averaged conditional score standard deviation

$$\overline{\Sigma}_t(\mathcal{A}_{<t}) = \mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})}\left[\sigma_t(V)\right]. \tag{16}$$

Let

$$G^\star_{1,t}(\mathcal{A}_{<t}, \varepsilon) = \sup_{q_t \in \mathcal{Q}_t(\mathcal{A}_{<t},\varepsilon)} G_{1,t}(q_t; \mathcal{A}_{<t}).$$

Then the following statements hold.

(i) Pointwise first-order optimizer. *For $V \mid \widetilde{h}_t$-almost every v, any finite interior optimizer of the pointwise first-order problem has the exponential-tilting form*

$$q^\star_t(y \mid v, \widetilde{h}_t) = \frac{m_{t,v}(y)\exp\big(\eta^\star_t(v)\psi_t(v,y)\big)}{\sum_{y'} m_{t,v}(y')\exp\big(\eta^\star_t(v)\psi_t(v,y')\big)} = \frac{m_{t,v}(y)\exp\big(\eta^\star_t(v)\widetilde{\psi}_t(v,y)\big)}{\sum_{y'} m_{t,v}(y')\exp\big(\eta^\star_t(v)\widetilde{\psi}_t(v,y')\big)}, \tag{17}$$

where $\eta^\star_t(v) \geq 0$. If $\sigma_t(v) = 0$, the native choice $q^\star_t(\cdot \mid v, \widetilde{h}_t) = m_{t,v}(\cdot)$ is optimal.

(ii) Lower bound for the optimal first-order gain. *There exists a constant $C_B > 0$, depending only on the bounded-ratio constant B, such that for every $\varepsilon > 0$,*

$$G^\star_{1,t}(\mathcal{A}_{<t}, \varepsilon) \geq \sqrt{2\varepsilon}\,\overline{\Sigma}_t(\mathcal{A}_{<t}) - C_B\varepsilon. \tag{18}$$

Proof. For a fixed visual realization v, define the pointwise first-order objective

$$g_v(q) = \sum_y \left(q(y) - m_{t,v}(y)\right)\psi_t(v,y)$$

under the constraint

$$\mathbb{D}_{\mathrm{KL}}(q \| m_{t,v}) \leq \varepsilon.$$

Because the KL constraint is pointwise in v, the first-order optimization separates across visual realizations.

For fixed v, maximizing $g_v(q)$ is equivalent to maximizing

$$\sum_y q(y)\psi_t(v,y)$$

subject to the KL constraint. The Lagrangian calculation gives, for some $\lambda_v > 0$,

$$q^\star(y) \propto m_{t,v}(y)\exp\left(\frac{\psi_t(v,y)}{\lambda_v}\right).$$

Writing $\eta^\star_t(v) = 1/\lambda_v$ gives the exponential-tilting form in Eq. (17). Adding the v-dependent constant $-\mu_t(v)$ to the score does not change the normalized distribution, so the centered form is equivalent. If $\sigma_t(v) = 0$, then $g_v(q) = 0$ for all feasible q, and the native choice is optimal. This proves Part (i).

We now prove Part (ii). By Assumption B.2, for every $y \in \mathrm{Supp}(m_{t,v})$,

$$|\psi_t(v,y)| \leq B.$$

Therefore the centered score satisfies

$$|\widetilde{\psi}_t(v,y)| \leq 2B.$$

Let

$$M = 2B.$$

For a fixed v, consider the centered exponential family

$$q_{v,\eta}(y) = \frac{m_{t,v}(y)\exp\big(\eta\widetilde{\psi}_t(v,y)\big)}{\sum_{y'} m_{t,v}(y')\exp\big(\eta\widetilde{\psi}_t(v,y')\big)}, \qquad \eta \geq 0.$$

Let

$$\Lambda_v(\eta) = \log \sum_y m_{t,v}(y) \exp\big(\eta \widetilde{\psi}_t(v,y)\big).$$

Then

$$\Lambda_v'(0) = 0, \qquad \Lambda_v''(0) = \sigma_t^2(v),$$

and

$$\Lambda_v'(\eta) = \mathbb{E}_{Y\sim q_{v,\eta}}\left[\widetilde{\psi}_t(v,Y)\right], \qquad \Lambda_v''(\eta) = \mathrm{Var}_{Y\sim q_{v,\eta}}\left[\widetilde{\psi}_t(v,Y)\right].$$

Since $|\widetilde{\psi}_t| \le M$, the density ratio between $q_{v,\eta}$ and $m_{t,v}$ is bounded between $e^{-2M\eta}$ and $e^{2M\eta}$. Hence

$$e^{-2M\eta}\sigma_t^2(v) \le \Lambda_v''(\eta) \le e^{2M\eta}\sigma_t^2(v).$$

For $0 \le \eta \le 1/(2M)$, using $e^{-x} \ge 1 - x$, we obtain

$$\Lambda_v'(\eta) = \int_0^\eta \Lambda_v''(u)\,du \ge \sigma_t^2(v)\int_0^\eta (1-2Mu)\,du = \sigma_t^2(v)\big(\eta - M\eta^2\big).$$

Moreover,

$$\mathbb{D}_{\mathrm{KL}}(q_{v,\eta}\|m_{t,v}) = \eta\Lambda_v'(\eta) - \Lambda_v(\eta) = \int_0^\eta u\,\Lambda_v''(u)\,du.$$

For $0 \le u \le 1/(2M)$, we have $e^{2Mu} \le 1 + 4Mu$. Therefore,

$$\mathbb{D}_{\mathrm{KL}}(q_{v,\eta}\|m_{t,v}) \le \sigma_t^2(v)\int_0^\eta u(1+4Mu)\,du = \sigma_t^2(v)\left(\frac{\eta^2}{2} + \frac{4M\eta^3}{3}\right).$$

Equivalently,

$$\mathbb{D}_{\mathrm{KL}}(q_{v,\eta}\|m_{t,v}) \le \frac{\sigma_t^2(v)\eta^2}{2}\left(1 + \frac{8M\eta}{3}\right).$$

If $\sigma_t(v) = 0$, the desired lower bound holds because the native choice gives first-order gain zero. Assume $\sigma_t(v) > 0$. Define

$$a_v = \frac{\sqrt{2\varepsilon}}{\sigma_t(v)}, \qquad s_v = Ma_v, \qquad d_0 = \frac{8}{3}.$$

First consider the small-radius regime $s_v \le 1/d_0$. Choose

$$\eta_v = a_v(1 - d_0 s_v).$$

Then $0 \le \eta_v \le a_v$ and $M\eta_v \le s_v \le 1/d_0 < 1/2$. Hence the preceding KL bound applies. Since $M\eta_v \le s_v$,

$$\mathbb{D}_{\mathrm{KL}}(q_{v,\eta_v}\|m_{t,v}) \le \frac{\sigma_t^2(v)a_v^2}{2}(1-d_0s_v)^2(1+d_0s_v).$$

Because $\sigma_t^2(v)a_v^2/2 = \varepsilon$ and

$$(1-z)^2(1+z) \le 1 \qquad \text{for } 0 \le z \le 1,$$

with $z = d_0 s_v$, we have

$$\mathbb{D}_{\mathrm{KL}}(q_{v,\eta_v}\|m_{t,v}) \le \varepsilon.$$

Thus q_{v,η_v} is feasible. Its first-order gain satisfies

$$g_v(q_{v,\eta_v}) = \Lambda_v'(\eta_v) \ge \sigma_t^2(v)(\eta_v - M\eta_v^2).$$

Using $\eta_v = a_v(1 - d_0 s_v)$ and $s_v = Ma_v$, we obtain

$$g_v(q_{v,\eta_v}) \ge \sqrt{2\varepsilon}\,\sigma_t(v) - 2M(d_0+1)\varepsilon.$$

Now consider the complementary regime $s_v > 1/d_0$. Since $a_v = \sqrt{2\varepsilon}/\sigma_t(v)$, this implies

$$\sqrt{2\varepsilon}\,\sigma_t(v) = \frac{2\varepsilon}{a_v} < 2Md_0\,\varepsilon.$$

Choosing $C_B \geq 2M\max\{d_0+1, d_0\}$, the right-hand side

$$\sqrt{2\varepsilon}\,\sigma_t(v) - C_B\varepsilon$$

is nonpositive. Since the native choice $q = m_{t,v}$ is feasible and gives $g_v(q) = 0$, the desired lower bound also holds in this regime.

Combining both regimes, for all v,

$$g_v^\star(\varepsilon) \geq \sqrt{2\varepsilon}\,\sigma_t(v) - C_B\varepsilon,$$

where C_B depends only on B, since $M = 2B$. Taking expectation over $V \mid \widetilde{h}_t(\mathcal{A}_{<t})$ yields

$$G_{1,t}^\star(\mathcal{A}_{<t}, \varepsilon) \geq \sqrt{2\varepsilon}\,\mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})}[\sigma_t(V)] - C_B\varepsilon = \sqrt{2\varepsilon}\,\overline{\Sigma}_t(\mathcal{A}_{<t}) - C_B\varepsilon.$$

□

Lemma B.3 (Entropy–reference-gap lower bound)**.** *Fix a candidate set $\mathcal{A}$, a step t, and the induced masked history $\widetilde{h}_t = \widetilde{h}_t(\mathcal{A}_{<t})$. Assume Assumptions B.1, B.2, and B.3. Then*

$$\overline{\Sigma}_t(\mathcal{A}_{<t}) \geq \frac{\gamma}{\sqrt{2\pi e\,B}}\mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})}\left[\left(\exp\big(\mathcal{H}(m_{t,V})\big) - 2\right)_+ \sqrt{\mathbb{D}_{\mathrm{KL}}(p_{t,V}\,\|\,p_{\mathcal{M},t})}\right]. \tag{19}$$

Remark B.5. Lemma B.3 expresses $\overline{\Sigma}_t(\mathcal{A}_{<t})$ as a product of an entropy-based factor and a reference-gap factor. The entropy factor arises from a variance–entropy inequality applied to $\psi_t(V, Y)$. The reference-gap factor is introduced through Assumption B.2: since

$$\mathbb{D}_{\mathrm{KL}}(p_{t,V}\|p_{\mathcal{M},t})/B \in [0,1],$$

multiplying any nonnegative lower bound on $\sigma_t^2(V)$ by this ratio yields a possibly looser lower bound on the same quantity. The resulting factored form is therefore conservative: it is tight in the entropy factor but loose in the reference-gap factor, and reflects the operational principle that branching room is useful only when the corresponding continuation remains distinguishable from the vision-marginalized reference.

Proof. Fix a candidate set $\mathcal{A}$, a step t, the induced masked history $\widetilde{h}_t$, and a visual realization v. Let $Y \sim m_{t,v}$, and define

$$Z = \psi_t(v, Y).$$

By Assumption B.3, distinct values of Z are separated by at least γ. Let $U \sim \mathrm{Unif}[0,\gamma)$ be independent of Z, and define

$$W = Z + U.$$

Since the support points of Z are γ-separated, the intervals $[z, z+\gamma)$ are disjoint up to endpoints. Hence

$$h(W) = \mathcal{H}(Z) + \log\gamma,$$

where $h(\cdot)$ denotes differential entropy. By the maximum-entropy property of Gaussians,

$$\mathrm{Var}(W) \geq \frac{1}{2\pi e}\exp(2h(W)).$$

Since Z and U are independent,

$$\mathrm{Var}(W) = \mathrm{Var}(Z) + \frac{\gamma^2}{12}.$$

Therefore,

$$\mathrm{Var}(Z) + \frac{\gamma^2}{12} \geq \frac{\gamma^2}{2\pi e}\exp(2\mathcal{H}(Z)).$$

Because $Z = \psi_t(v, Y)$ is a deterministic function of Y, and Assumption B.3 makes this function injective on $\mathrm{Supp}(m_{t,v})$,

$$\mathcal{H}(Z) = \mathcal{H}(Y) = \mathcal{H}(m_{t,v}).$$

Thus

$$\mathrm{Var}_{Y\sim m_{t,v}}\big(\psi_t(v, Y)\big) \geq \left[\frac{\gamma^2}{2\pi e}\exp(2\mathcal{H}(m_{t,v})) - \frac{\gamma^2}{12}\right]_+.$$

Equivalently,

$$\mathrm{Var}_{Y\sim m_{t,v}}\big(\psi_t(v, Y)\big) \geq \frac{\gamma^2}{2\pi e}\left[\exp(2\mathcal{H}(m_{t,v})) - \frac{\pi e}{6}\right]_+.$$

By Assumption B.2,

$$0 \leq \mathbb{D}_{\mathrm{KL}}(p_{t,v}\|p_{\mathcal{M},t}) \leq B.$$

Therefore, for any nonnegative A_0,

$$A_0 \geq \frac{\mathbb{D}_{\mathrm{KL}}(p_{t,v}\|p_{\mathcal{M},t})}{B}A_0.$$

Applying this with

$$A_0 = \frac{\gamma^2}{2\pi e}\left[\exp(2\mathcal{H}(m_{t,v})) - \frac{\pi e}{6}\right]_+$$

gives

$$\mathrm{Var}_{Y\sim m_{t,v}}\big(\psi_t(v, Y)\big) \geq \frac{\gamma^2}{2\pi e\, B}\left[\exp(2\mathcal{H}(m_{t,v})) - \frac{\pi e}{6}\right]_+ \mathbb{D}_{\mathrm{KL}}(p_{t,v}\|p_{\mathcal{M},t}).$$

Taking square roots yields

$$\sigma_t(v) \geq \frac{\gamma}{\sqrt{2\pi e\, B}}\sqrt{\left[\exp(2\mathcal{H}(m_{t,v})) - \frac{\pi e}{6}\right]_+ \mathbb{D}_{\mathrm{KL}}(p_{t,v}\|p_{\mathcal{M},t})}.$$

Using

$$\sqrt{(x-c)_+} \geq (\sqrt{x} - \sqrt{c})_+, \qquad x \geq 0,\ c > 0,$$

with

$$x = \exp(2\mathcal{H}(m_{t,v})), \qquad c = \frac{\pi e}{6},$$

and using $\sqrt{\pi e/6} < 2$, we obtain

$$\sigma_t(v) \geq \frac{\gamma}{\sqrt{2\pi e\, B}}\left(\exp(\mathcal{H}(m_{t,v})) - 2\right)_+ \sqrt{\mathbb{D}_{\mathrm{KL}}(p_{t,v}\|p_{\mathcal{M},t})}.$$

Averaging over $V \mid \widetilde{h}_t(\mathcal{A}_{<t})$ proves Eq. (19). □

Lemma B.4 (Local downstream visual-gain lower bound)**.** *Fix a candidate set $\mathcal{A}$, a step t, and the induced masked history $\widetilde{h}_t = \widetilde{h}_t(\mathcal{A}_{<t})$. Assume Assumptions B.1, B.2, and B.3. Then*

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \geq \frac{\gamma\sqrt{\varepsilon}}{\sqrt{\pi e\, B}}\mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})}\left[\big(\exp\big(\mathcal{H}(m_{t,V})\big) - 2\big)_+ \sqrt{\mathbb{D}_{\mathrm{KL}}(p_{t,V}\,\|\,p_{\mathcal{M},t})}\right] - C_B\varepsilon. \tag{20}$$

Proof. By Lemma B.1, for any $q_t \in \mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon)$,

$$\Delta_t(q_t; \mathcal{A}_{<t}) \geq G_{1,t}(q_t; \mathcal{A}_{<t}).$$

Taking the supremum over $q_t \in \mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon)$ gives

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \geq G_{1,t}^\star(\mathcal{A}_{<t}, \varepsilon).$$

By Lemma B.2(ii),

$$G_{1,t}^\star(\mathcal{A}_{<t}, \varepsilon) \geq \sqrt{2\varepsilon}\,\overline{\Sigma}_t(\mathcal{A}_{<t}) - C_B\varepsilon.$$

Combining the two inequalities,

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \geq \sqrt{2\varepsilon}\,\overline{\Sigma}_t(\mathcal{A}_{<t}) - C_B\varepsilon.$$

By Lemma B.3,

$$\overline{\Sigma}_t(\mathcal{A}_{<t}) \geq \frac{\gamma}{\sqrt{2\pi e\, B}} \mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})} \left[\left(\exp\big(\mathcal{H}(m_{t,V})\big) - 2\right)_+ \sqrt{\mathbb{D}_{\mathrm{KL}}(p_{t,V} \| p_{\mathcal{M},t})} \right].$$

Substituting this into the previous display yields

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \geq \frac{\gamma\sqrt{\varepsilon}}{\sqrt{\pi e\, B}} \mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})} \left[\left(\exp\big(\mathcal{H}(m_{t,V})\big) - 2\right)_+ \sqrt{\mathbb{D}_{\mathrm{KL}}(p_{t,V} \| p_{\mathcal{M},t})} \right] - C_B\varepsilon,$$

which proves the claim. □

B.5. Main Theorem: Joint Sparse Selection

For each candidate set $\mathcal{A}$ and each step t, define the ideal local lower-bound score

$$\mathcal{L}_t(\mathcal{A}, \varepsilon) = \left[\frac{\gamma\sqrt{\varepsilon}}{\sqrt{\pi e\, B}} \mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})} \left[\left(\exp\big(\mathcal{H}(m_{t,V})\big) - 2\right)_+ \sqrt{\mathbb{D}_{\mathrm{KL}}(p_{t,V} \,\|\, p_{\mathcal{M},t})} \right] - C_B\varepsilon \right]_+. \tag{21}$$

Here $p_{\mathcal{M},t}$ and $m_{t,V}$ are defined using the masked history $\widetilde{h}_t(\mathcal{A}_{<t})$. The positive part is valid because the native choice is feasible and gives zero downstream gain.

Theorem B.1 (Joint sparse lower bound for one-step rollback gains)**.** *Let $K_\rho = \lfloor \rho T \rfloor$. Assume Assumptions B.1, B.2, B.3, and B.4. Then the following statements hold.*

(1) Lower bound for every candidate selected set. *For every $\mathcal{A} \subseteq [T]$ with $|\mathcal{A}| \leq K_\rho$,*

$$\sup_{\pi \in \Pi(\mathcal{A},\varepsilon)} \sum_{t\in\mathcal{A}} \Delta_t(\pi; \mathcal{A}_{<t}) \geq \sum_{t\in\mathcal{A}} \mathcal{L}_t(\mathcal{A}, \varepsilon). \tag{22}$$

(2) Joint lower bound for the sparse oracle. *The joint sparse optimum in Eq. (8) satisfies*

$$\mathcal{J}^\star(\varepsilon) \geq \max_{\mathcal{A}\subseteq[T],\, |\mathcal{A}|\leq K_\rho} \sum_{t\in\mathcal{A}} \mathcal{L}_t(\mathcal{A}, \varepsilon). \tag{23}$$

(3) Ideal lower-bound selection. *The ideal lower-bound selected set is*

$$\mathcal{A}_{\mathrm{LB}}^\star \in \arg\max_{\mathcal{A}\subseteq[T],\, |\mathcal{A}|\leq K_\rho} \sum_{t\in\mathcal{A}} \mathcal{L}_t(\mathcal{A}, \varepsilon). \tag{24}$$

(4) Top-K_ρ rule as a special case. *If the local scores are fixed independently of $\mathcal{A}$, that is,*

$$\mathcal{L}_t(\mathcal{A}, \varepsilon) = \mathcal{L}_t(\varepsilon) \quad \textit{for all } \mathcal{A},$$

then the ideal lower-bound selection reduces to

$$\mathcal{A}_{\mathrm{LB}}^\star \in \mathrm{TopK}_{t\in[T]} \{\mathcal{L}_t(\varepsilon)\}_{K_\rho}.$$

Equivalently, it selects the largest $K_\rho = \lfloor \rho T \rfloor$ ideal local lower-bound scores, with ties broken arbitrarily.

Remark B.6 (Set-dependent scores). Because $\widetilde{h}_t(\mathcal{A}_{<t})$ depends on the earlier selected positions, the score $\mathcal{L}_t(\mathcal{A}, \varepsilon)$ can change when $\mathcal{A}$ changes. Thus the general lower-bound selection problem in Eq. (24) is a genuine joint set optimization. The simple top-K_ρ rule is rigorously valid only in the fixed-score special case.

Remark B.7 (Scope of the lower bound). The theorem lower bounds the sparse objective used in Eq. (1), namely the sum of one-step-then-rollback marginal downstream gains. It does not claim that this sum is equal to the conditional mutual information of a policy that applies all selected interventions simultaneously during a full rollout. The masking construction and rollback evaluation intentionally isolate marginal credit for each selected step.

Remark B.8 (Connection to practical reflection anchors). The lower-bound score $\mathcal{L}_t(\mathcal{A}, \varepsilon)$ contains two signals: local branching room, through $\mathcal{H}(m_{t,V})$, and downstream visual relevance, through $\mathbb{D}_{\mathrm{KL}}(p_{t,V} \| p_{\mathcal{M},t})$. Computing the full score would require set-dependent masking, counterfactual suffix laws, and visual marginalization. The practical reflection-anchor rule in the main text therefore uses native-history entropy as a tractable proxy for the branching-room factor, while the continuation-level training objective supplies the downstream visual-relevance pressure.

Proof. Fix any candidate set $\mathcal{A} \subseteq [T]$ with $|\mathcal{A}| \le K_\rho$. For each $t \in \mathcal{A}$, Lemma B.4 gives

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \ge \frac{\gamma\sqrt{\varepsilon}}{\sqrt{\pi e\, B}} \mathbb{E}_{V|\widetilde{h}_t(\mathcal{A}_{<t})} \left[\big(\exp\big(\mathcal{H}(m_{t,V})\big) - 2\big)_+ \sqrt{\mathbb{D}_{\mathrm{KL}}(p_{t,V} \| p_{\mathcal{M},t})} \right] - C_B \varepsilon.$$

Since the native choice $q_t = m_{t,V}$ is feasible and gives zero gain,

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \ge 0.$$

Therefore,

$$\Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \ge \mathcal{L}_t(\mathcal{A}, \varepsilon).$$

Summing over $t \in \mathcal{A}$ gives

$$\sum_{t \in \mathcal{A}} \Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) \ge \sum_{t \in \mathcal{A}} \mathcal{L}_t(\mathcal{A}, \varepsilon).$$

We next relate the local optima to a single decoder. Fix any $\eta > 0$. For each $t \in \mathcal{A}$, choose a feasible local target

$$q_t^\eta \in \mathcal{Q}_t(\mathcal{A}_{<t}, \varepsilon)$$

such that

$$\Delta_t(q_t^\eta; \mathcal{A}_{<t}) \ge \Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) - \frac{\eta}{\max\{|\mathcal{A}|, 1\}}.$$

By Assumption B.4, there exists $\pi^\eta \in \Pi(\mathcal{A}, \varepsilon)$ realizing all these local targets at the selected masked histories. Hence

$$\sup_{\pi \in \Pi(\mathcal{A}, \varepsilon)} \sum_{t \in \mathcal{A}} \Delta_t(\pi; \mathcal{A}_{<t}) \ge \sum_{t \in \mathcal{A}} \Delta_t(q_t^\eta; \mathcal{A}_{<t}) \ge \sum_{t \in \mathcal{A}} \Delta_t^\star(\mathcal{A}_{<t}, \varepsilon) - \eta.$$

Letting $\eta \downarrow 0$ gives

$$\sup_{\pi \in \Pi(\mathcal{A}, \varepsilon)} \sum_{t \in \mathcal{A}} \Delta_t(\pi; \mathcal{A}_{<t}) \ge \sum_{t \in \mathcal{A}} \Delta_t^\star(\mathcal{A}_{<t}, \varepsilon).$$

Combining this with the lower bound on $\sum_{t \in \mathcal{A}} \Delta_t^\star(\mathcal{A}_{<t}, \varepsilon)$ proves Eq. (22).

Taking the maximum over all $\mathcal{A} \subseteq [T]$ with $|\mathcal{A}| \le K_\rho$, and using the definition of $\mathcal{J}^\star(\varepsilon)$, gives

$$\mathcal{J}^\star(\varepsilon) \ge \max_{\mathcal{A} \subseteq [T],\, |\mathcal{A}| \le K_\rho} \sum_{t \in \mathcal{A}} \mathcal{L}_t(\mathcal{A}, \varepsilon),$$

which proves Eq. (23). The selection rule in Eq. (24) follows directly from maximizing this lower-bound surrogate over $\mathcal{A}$.

Finally, suppose the local scores are fixed independently of $\mathcal{A}$, so that

$$\mathcal{L}_t(\mathcal{A}, \varepsilon) = \mathcal{L}_t(\varepsilon).$$

Then the surrogate becomes modular:

$$\max_{\mathcal{A} \subseteq [T],\, |\mathcal{A}| \le K_\rho} \sum_{t \in \mathcal{A}} \mathcal{L}_t(\varepsilon).$$

Let

$$\mathcal{L}_{(1)}(\varepsilon) \ge \mathcal{L}_{(2)}(\varepsilon) \ge \cdots \ge \mathcal{L}_{(T)}(\varepsilon)$$

be the sorted scores. Since all scores are nonnegative by construction, any set of size at most K_ρ has value at most $\sum_{i=1}^{K_\rho} \mathcal{L}_{(i)}(\varepsilon)$, and this value is achieved by selecting the K_ρ indices with the largest scores. Hence, in the fixed-score special case, the lower-bound oracle selects the largest $K_\rho = \lfloor \rho T \rfloor$ ideal local lower-bound scores. □

B.6. Proofs of Claim B.1 and Claim B.2

Proof of Claim B.1. Fix a candidate set $\mathcal{A}$, a step t, the induced masked history $\widetilde{h}_t$, and a visual realization v. For any y with $m_{t,v}(y) > 0$, by the definition of ψ_t,

$$\psi_t(v, y) = \mathbb{E}_{Y_{>t} \sim K_{t,v}(\cdot|y)} \left[\log \frac{\widetilde{p}_{t,v}(y, Y_{>t})}{p_{\mathcal{M},t}(y, Y_{>t})} \right].$$

Using the factorizations

$$\widetilde{p}_{t,v}(y, y_{>t}) = m_{t,v}(y) K_{t,v}(y_{>t} \mid y),$$

and

$$p_{\mathcal{M},t}(y, y_{>t}) = m_{\mathcal{M},t}(y) p_{\mathcal{M},t}(y_{>t} \mid y),$$

we obtain

$$\log \frac{\widetilde{p}_{t,v}(y, Y_{>t})}{p_{\mathcal{M},t}(y, Y_{>t})} = \log \frac{m_{t,v}(y)}{m_{\mathcal{M},t}(y)} + \log \frac{K_{t,v}(Y_{>t} \mid y)}{p_{\mathcal{M},t}(Y_{>t} \mid y)}.$$

Taking expectation over $Y_{>t} \sim K_{t,v}(\cdot \mid y)$ yields

$$\psi_t(v, y) = \log \frac{m_{t,v}(y)}{m_{\mathcal{M},t}(y)} + \mathbb{D}_{\mathrm{KL}}(K_{t,v}(\cdot \mid y) \,\|\, p_{\mathcal{M},t}(\cdot \mid y)).$$

□

Proof of Claim B.2. By the definition of conditional mutual information under the native masked-history one-step-then-rollback suffix law,

$$I_{p_\theta}(V; Y_{\geq t} \mid \widetilde{h}_t) = \mathbb{E}_{V|\widetilde{h}_t} \left[\sum_{y_{\geq t}} \widetilde{p}_{t,V}(y_{\geq t}) \log \frac{\widetilde{p}_{t,V}(y_{\geq t})}{p_{\mathcal{M},t}(y_{\geq t})} \right].$$

Using

$$\widetilde{p}_{t,V}(y_{\geq t}) = m_{t,V}(y_t) K_{t,V}(y_{>t} \mid y_t),$$

we rewrite this as

$$I_{p_\theta}(V; Y_{\geq t} \mid \widetilde{h}_t) = \mathbb{E}_{V|\widetilde{h}_t} \mathbb{E}_{Y_t \sim m_{t,V}} \mathbb{E}_{Y_{>t} \sim K_{t,V}(\cdot|Y_t)} \left[\log \frac{\widetilde{p}_{t,V}(Y_t, Y_{>t})}{p_{\mathcal{M},t}(Y_t, Y_{>t})} \right].$$

The inner conditional expectation is exactly $\psi_t(V, Y_t)$. Therefore,

$$I_{p_\theta}(V; Y_{\geq t} \mid \widetilde{h}_t) = \mathbb{E}_{V|\widetilde{h}_t} \mathbb{E}_{Y_t \sim m_{t,V}} \left[\psi_t(V, Y_t) \right].$$

□

C. Experimental Details of Main Results

Table 4. **Main results with standard deviation (mean@8 acc %).**

Base Model	Model	Reasoning-Intensive				General-Domain		Avg.
		MathVision	MathVerse	EMMA	LogicVista	MMMU-Pro	RealWorldQA	
Qwen3-VL-8B Instruct	Base	41.74 ± 1.21	50.21 ± 0.87	38.34 ± 0.25	62.28 ± 0.04	37.59 ± 0.20	67.20 ± 0.55	49.56
	+ GRPO	48.03 ± 0.14	55.23 ± 0.49	41.34 ± 0.41	59.51 ± 0.96	48.50 ± 0.60	67.66 ± 0.62	53.38 (↑ 3.82)
	+ PAPO$_G$	43.75 ± 0.29	54.57 ± 0.15	39.00 ± 0.11	64.29 ± 0.32	50.00 ± 0.50	69.41 ± 0.74	53.50 (↑ 3.94)
	+ VPPO$_G$	48.03 ± 0.27	56.35 ± 0.38	36.00 ± 0.23	**66.07** ± 0.38	49.16 ± 0.52	67.48 ± 0.17	53.85 (↑ 4.29)
	+ RAPO$_G$	**48.27** ± 0.19	**56.72** ± 0.09	**42.50** ± 0.44	62.50 ± 0.86	**51.10** ± 0.25	**69.54** ± 0.18	**55.11** (↑ 5.55)
Qwen3-VL-2B Instruct	Base	27.30 ± 1.67	36.75 ± 0.75	26.87 ± 1.72	47.13 ± 1.16	24.44 ± 0.49	60.47 ± 0.38	37.16
	+ GRPO	28.29 ± 1.71	44.35 ± 0.87	28.50 ± 0.30	48.71 ± 0.99	30.61 ± 0.60	64.28 ± 0.85	40.79 (↑ 3.63)
	+ PAPO$_G$	28.29 ± 0.33	45.18 ± 0.67	27.25 ± 0.33	46.43 ± 0.66	31.45 ± 0.14	64.05 ± 0.57	40.44 (↑ 3.28)
	+ VPPO$_G$	30.59 ± 0.25	45.43 ± 0.32	28.75 ± 0.65	45.76 ± 1.06	31.73 ± 0.32	63.14 ± 0.44	40.90 (↑ 3.74)
	+ RAPO$_G$	**33.55** ± 1.21	**46.13** ± 0.56	**31.25** ± 0.71	**49.55** ± 0.48	**32.20** ± 0.63	**65.36** ± 0.40	**43.01** (↑ 5.85)
Qwen2.5-VL-7B Instruct	Base	21.05 ± 0.52	29.44 ± 0.44	23.25 ± 0.42	42.19 ± 0.57	21.22 ± 0.42	54.51 ± 0.46	31.94
	MM-Eureka	32.57 ± 0.27	45.18 ± 0.45	28.75 ± 0.46	44.64 ± 1.03	26.82 ± 0.33	**61.96** ± 0.65	39.99 (↑ 8.05)
	VL-Rethinker	**34.21** ± 0.33	45.30 ± 0.89	28.75 ± 0.77	44.20 ± 0.34	34.10 ± 0.23	59.22 ± 0.92	40.96 (↑ 9.02)
	R1-ShareVL	25.99 ± 0.11	46.22 ± 0.34	**30.50** ± 0.56	46.88 ± 0.25	34.45 ± 0.28	56.47 ± 0.10	40.09 (↑ 8.15)
	Base+RAPO$_G$	27.96 ± 0.33	47.34 ± 0.71	29.00 ± 0.67	45.31 ± 0.55	36.01 ± 0.48	60.26 ± 0.13	40.98 (↑ 9.04)
	Base+RAPO$_D$	30.26 ± 0.50	**47.59** ± 0.42	**30.50** ± 0.99	**47.13** ± 0.13	**36.04** ± 0.48	58.40 ± 0.77	**41.65** (↑ 9.71)
Qwen2-VL-7B Instruct	Base	15.79 ± 1.16	10.25 ± 0.65	11.75 ± 0.46	13.17 ± 0.94	11.73 ± 0.45	31.83 ± 0.51	15.75
	TVC	18.75 ± 0.82	21.98 ± 1.10	20.75 ± 0.45	**41.96** ± 0.77	23.34 ± 0.87	52.88 ± 0.38	29.94 (↑ 14.19)
	MINT-CoT	22.04 ± 0.19	24.62 ± 0.23	19.00 ± 0.66	31.47 ± 0.71	17.63 ± 0.86	49.15 ± 0.12	27.32 (↑ 11.57)
	Base+RAPO$_G$	22.37 ± 0.33	32.99 ± 0.71	25.25 ± 0.40	37.72 ± 0.55	28.22 ± 0.48	64.67 ± 0.13	35.21 (↑ 19.46)
	Base+RAPO$_D$	**23.36** ± 0.87	**36.31** ± 0.66	**28.75** ± 0.56	39.73 ± 1.52	**29.39** ± 0.86	**65.01** ± 0.35	**37.09** (↑ 21.34)

C.1. Training Details

We adopt ViRL39K[2] as the RL training dataset, consisting of 38,870 curated and verifiable multimodal QA pairs. All training samples are formatted using the prompt template illustrated in the box below.

CoT Training Data Template:

```
<image>
{Question}
```

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think> </think>` tags. The final answer MUST BE put in `\boxed{}`.

We implement GRPO and RAPO using the VeRL framework[3], and PAPO and VPPO using Easy-R1[4], following the training settings reported in their original papers. In the main experiments, among the baselines we only conduct RL training for PAPO and VPPO. Both methods follow the original training settings reported in their respective papers, with minor adjustments to the `batch_size`, `training steps`, `max_prompt_length`, and `max_response_length` to ensure consistency with RAPO when training on Qwen3-VL-2B-Instruct and Qwen3-VL-8B-Instruct. For VPPO, we adopt the same training settings as provided in its official code repository to train Qwen3-VL-8B-Instruct. For PAPO, we follow the settings used for Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B to train Qwen3-VL-2B-Instruct and Qwen3-VL-8B-Instruct, respectively. The reward signal used in the main experiments, as well as in all additional experiments, is a binary accuracy score, where a correct answer receives a reward of 1 and an incorrect answer receives a reward of 0. In addition, the

[2] https://huggingface.co/datasets/TIGER-Lab/ViRL39K
[3] https://github.com/volcengine/verl
[4] https://github.com/hiyouga/EasyR1

Table 5. RL hyperparameters for different experimental groups.

Hyperparameter	Group (i)	Group (ii) & (iii)
General Training		
Learning rate	$1e{-}6$	$1e{-}6$
RL steps	100	200
Optimizer	AdamW	AdamW
LR scheduler	Constant	Constant
Freeze vision tower	True	True
Basic GRPO Settings		
Rollout batch size	256	384
Rollouts per prompt	5	5
PPO minibatch size	64	128
Rollout temperature	1.0	1.0
Rollout top-p	1.0	1.0
Max prompt length	2048	4096
Max response length	8192	2048
KL penalty coefficient	0.02	0.02
Loss aggregation mode	Token-level	Token-level
DAPO Recipe		
Dynamic sampling	True	True
Clip ratio (low)	0.20	0.20
Clip ratio (high)	0.28	0.28
Overlong reward shaping	True	False
Overlong buffer length	1024	–
Overlong penalty factor	1.0	–
Loss aggregation mode	Token-level	Token-level
RAPO-Specific Settings		
MI enhancement factor γ	0.01	0.01
Token selection threshold	Top 20%	Top 20%
Window length	3 (1 for 8B)	1

vision tower is frozen during training for all RL experiments. For computational resources, Qwen3-VL-2B-Instruct and Qwen3-VL-8B-Instruct are trained on 8× NVIDIA RTX PRO 6000 Blackwell GPUs, while Qwen2-VL-7B-Instruct and Qwen2.5-VL-7B-Instruct are trained on 4 GPUs. All key hyperparameters for RL training are summarized in Table 5.

C.2. Evaluation Details

Benchmarks We evaluate our method on two categories of benchmarks: reasoning-intensive and general-domain, enabling comprehensive assessment of VLM capabilities. All evaluations are based on directly verifiable answers rather than LLM-as-a-judge evaluation. Detailed benchmark statistics are summarized in Table 6.

Table 6. Evaluation benchmarks used in our experiments.

Benchmark	Size	Domain	Task Type	Subset	Source
MathVision	304	Math	Free-form QA	Test-mini	HuggingFace
MathVerse	788	Math	Multiple-choice & Free-form QA	Visual-dominant	HuggingFace
EMMA-mini	400	STEM	Multiple-choice QA	Full	HuggingFace
LogicVista	448	Logic	Multiple-choice QA	Full	HuggingFace
MMMU-Pro	1,730	General	Multiple-choice QA	Vision	HuggingFace
RealWorldQA	765	General	Free-form QA	Full	HuggingFace

Evaluation Settings. Benchmark evaluation is conducted using vLLM v0.11.0[5]. Multiple-choice questions are evaluated via exact-match accuracy, while free-form questions are verified using math-verify[6]. The prompt template used during

[5] https://github.com/vllm-project/vllm
[6] https://github.com/huggingface/Math-Verify

evaluation is identical to that employed for training. During evaluation, we set the sampling temperature to 1.0 with top-$p = 1.0$) and top-k set to -1. Inference is performed with a batch size of 512. We enable the vLLM V1 execution mode (`VLLM_USE_V1=True`) and the maximum number of newly generated tokens is set to 8,192. We report mean and standard deviation over 8 sampled evaluation responses per input in Table 4.

D. Experimental Details of Empirical Analysis of RAPO's Mechanism

We randomly sample 500 instances from the ViRL dataset and evaluate the Qwen3-VL-2B-Instruct model before training and after GRPO and RAPO training, using the same settings as in the main experiments. Responses are generated via greedy decoding with a maximum length of 4096 tokens. We conduct three groups of experiments.

D.1. Trajectory Analysis

In this subsection, we visualize the KL divergence and attention distributions. The KL divergence is used to quantify *token-level visual dependency* by measuring the Kullback–Leibler (KL) divergence between the policy's predictive distributions under vision-conditioned and vision-masked settings. Formally, let V denote the visual input, X the textual context, and $Y_{1:T} = (Y_1, \ldots, Y_T)$ a length-T reasoning trajectory. We define the history at decoding step t as $H_t = (X, Y_{<t})$. The token-level visual dependency at step t is defined as the KL divergence between the policy output distributions conditioned on (V, H_t) and on H_t alone:

$$\mathcal{S}(H_t, V) = \mathbb{D}_{\mathrm{KL}}(\pi_\theta(\cdot \mid H_t, V) \parallel \pi_\theta(\cdot \mid H_t)) \,.$$

A high value of $\mathcal{S}(H_t, V)$ indicates that visual information plays a critical role in predicting the token at step t, identifying a key moment of visually grounded reasoning. We further visualize the attention distribution from the next generated token to visual tokens. Denoting the attention function as $\mathrm{Attn}(\cdot)$, the attention weights for token Y_t over visual tokens are computed as

$$\mathrm{Attn}(Y_t, V) = \mathrm{softmax}\left(\frac{q_t K_V^\top}{\sqrt{d}}\right),$$

where q_t is the query vector corresponding to Y_t, K_V denotes the key vectors of all visual tokens, and d is the key dimension.

We use Qwen3-VL-2B-Instruct to generate responses for 500 questions, which are divided into two groups based on correctness: incorrect (322) and correct (178). For each group, we compute the token-wise average of the KL divergence and attention distributions. For training comparison, we average the KL distributions over all 500 instances to compare RAPO and GRPO. This analysis shows that RAPO effectively increases visual dependency in the later stages of the chain-of-thought (CoT), whereas GRPO does not. As shown in Figures 8, GRPO fails to improve the KL divergence relative to the base model, while RAPO leads to a clear enhancement.

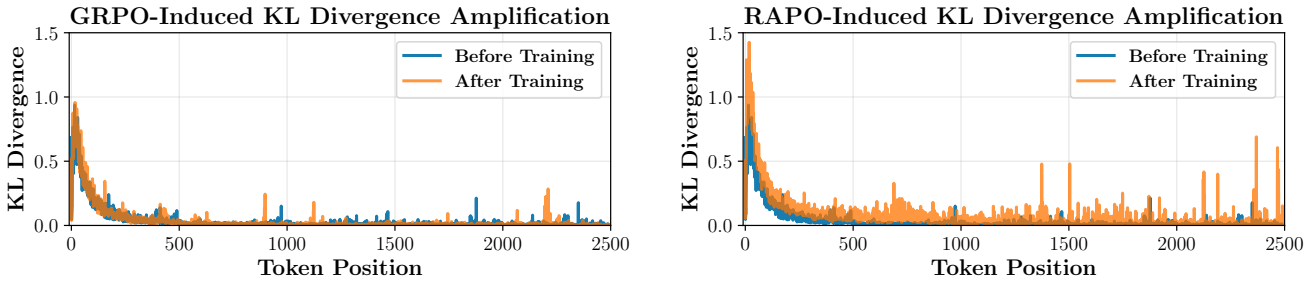


Figure 8. **KL Distributions along CoT.** (a) KL distribution after GRPO training; (b) KL distribution after RAPO training.

D.2. Anchor Token Analysis

For the anchor-token concentration visualization, we identify the top-20% highest-entropy tokens across all 500 responses generated by Qwen3-VL-2B-Instruct. To avoid rare-token artifacts, we restrict the analysis to token types appearing at least N times in the corpus ($N = 10$ in our experiments), and report for each retained type the fraction of its occurrences selected as a top-entropy anchor.

We construct a noised visual input $\tilde{V}$ by injecting additive Gaussian noise:

$$\tilde{V} = V + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where σ controls the noise magnitude and I denotes the identity covariance matrix with the same dimensionality as V. We then analyze the distributional shifts of different token categories under varying noise levels σ. Specifically, we use

$$\mathbb{D}_{\mathrm{KL}}\Big(\pi_\theta(\cdot \mid H_t, V) \parallel \pi_\theta(\cdot \mid H_t, \tilde{V})\Big)$$

to measure token sensitivity to visual perturbations, where larger KL values indicate stronger dependence on visual information.

D.3. Propagation of Visual Influence Analysis

We design a targeted experiment to measure *visual information propagation from anchor token positions*. Given a reflection anchor $t_k \in \mathcal{A}$, we evaluate how blocking its visual information, implemented by masking the corresponding visual attention, affects the visual dependency of other tokens. Specifically, we examine (i) the average impact on the subsequent three tokens $Y_{t_k:t_k+3}$, and (ii) the impact on the visual dependency of the next anchor token t_{k+1}. To quantify this effect, we define

$$\Delta\, \mathbb{D}_{\mathrm{KL}} = \mathcal{S}(H'_t, V) - \mathcal{S}(H_t, V),$$

where $\mathcal{S}(H_t, V) = \mathbb{D}_{\mathrm{KL}}(\pi_\theta(\cdot|H_t, V) \parallel \pi_\theta(\cdot|H_t))$ measures the visual dependence at step t under the original history, and $\mathcal{S}(H'_t, V)$ is the analogous quantity after masking the visual information associated with the selected anchor. By construction, $\Delta D_{\mathrm{KL}} \leq 0$: more negative values indicate that masking the anchor produces a larger drop in downstream visual dependence, i.e., stronger propagation of visual influence from the anchor.

We evaluate this effect across anchor tokens at different positions and compare RAPO with baseline methods. The results demonstrate that RAPO significantly enhances both the visual influence at anchor tokens and the propagation of visual information along the reasoning trajectory.

E. Intervention-Based Diagnostics for Reflection Anchors

We provide an intervention-based diagnostic to examine whether reflection anchors are associated with task-relevant visual evidence. This diagnostic does not directly estimate the oracle gain $\Delta_t(\pi)$, but tests whether the visual regions attended at reflection anchors have measurable influence on final prediction.

Attention-guided patch masking. We test whether early anchors attend to task-relevant visual evidence. On 100 randomly sampled ViRL validation examples, we select early reflection anchors and mask 20% of image patches under two strategies: attention-guided masking, which removes the highest-attended patches at the selected anchor, and random masking, which removes the same number of randomly sampled patches. We report the accuracy drop $\Delta\mathrm{Acc}$.

Table 7. **Attention-guided patch masking.** More negative values indicate larger accuracy degradation.

Model	Attention-guided	Random	Gap
GRPO	-24.3	-19.7	4.6
RAPO	-27.7	-21.3	6.4

As shown in Table 7, attention-guided masking causes larger accuracy degradation than random masking for both GRPO and RAPO. This suggests that early reflection anchors attend to image regions that are more relevant to final prediction than randomly selected patches. The larger gap under RAPO is consistent with the intended effect of strengthening the coupling between reflection anchors and task-relevant visual evidence. Together with the KL propagation analysis in Figure 7, this diagnostic provides intervention-based support for the proposed mechanism.

F. Compute-Matched Comparison

RAPO introduces an additional chain-masked contrastive KL signal during training, which increases per-step training cost relative to GRPO. To verify that the performance gain is not merely due to additional computation, we first report wall-clock

Table 8. **Training wall-clock cost.** RAPO incurs additional training overhead due to the chain-masked contrastive KL computation, but it uses the standard autoregressive forward pass at inference and therefore adds no inference-time overhead.

Model	GRPO Total	RAPO Total	GRPO / Step	RAPO / Step	Overhead
Qwen3-VL-2B	10h38m	12h31m	383s	451s	$\sim$18%
Qwen3-VL-8B	28h04m	35h58m	1010s	1295s	$\sim$28%

training cost and then compare RAPO against a compute-matched GRPO baseline trained for additional steps under a similar wall-clock budget. All measurements are conducted on $8\times$ NVIDIA RTX PRO 6000 Blackwell GPUs.

We then train GRPO for 120 steps to approximately match RAPO's 100-step wall-clock budget on Qwen3-VL-2B-Instruct. The GRPO-120 run takes 12h43m, comparable to RAPO-100 at 12h31m. As shown in Table 9, increasing GRPO from 100 to 120 steps improves the average score from 40.79 to 41.45, but still remains below RAPO's 43.01. This indicates that RAPO's gain is not explained by additional training computation, but by the proposed anchor-based propagation objective.

Table 9. **Compute-matched comparison.** GRPO is trained for additional steps to match RAPO's wall-clock budget. RAPO still achieves higher average performance.

Method	Steps	MathVis.	MathVer.	EMMA	LogicV.	MMMU-P.	RealWQ	Avg.
GRPO	100	28.29	44.35	28.50	48.71	30.61	64.28	40.79
GRPO	120	30.92	44.73	31.25	49.11	31.20	61.50	41.45
RAPO_G	100	**33.55**	**46.13**	**31.25**	**49.55**	**32.20**	**65.36**	**43.01**

RAPO's additional cost is training-only: at inference, the learned policy runs with the same standard autoregressive decoding procedure as the base model and does not require anchor search, chain masking, or additional visual re-attention modules.

G. Ablation and Robustness Study

G.1. Experimental Details

We additionally conduct experiments on two other models: Qwen3-VL-2B-Thinking and Gemma3-4B-IT. Qwen3-VL-2B-Thinking is designed to generate substantially longer CoT, which allows us to further validate the effectiveness of our method under long-CoT reasoning settings. For this model, we adopt the same training configuration as Qwen3-VL-2B-Instruct used in the main experiments, except that the maximum response length is increased to 16,384 tokens. The evaluation protocol is also kept identical, with the maximum number of newly generated tokens set to 16,384. Gemma3-4B-IT is a vision–language model with a substantially different architecture from the Qwen3-VL family. To assess the generality of our method across model architectures, we evaluate it under the same experimental settings as Qwen3-VL-Instruct, except that the maximum response length is set to 4,096 tokens.

Unless otherwise specified, all remaining ablation studies are conducted under the same training and evaluation configuration as the Qwen3-VL-2B-Instruct experiments reported in the main results.

G.2. Ablation on Anchor Selection

We further analyze how the anchor set $\mathcal{A}$ should be selected. The results are shown in Table 10.

Table 10. Ablation on anchor-selection strategies.

Strategy	Frac.	MathVis.	MathVer.	EMMA	LogicV.	MMMU-P.	RealWQ	Avg.
No anchors (GRPO)	100%	28.29	44.35	28.50	48.71	30.61	64.28	40.79
Random 20%	20.0%	29.61	42.93	26.56	48.88	30.81	60.98	39.96
Noun	28.7%	27.30	43.94	28.56	46.88	31.36	62.09	40.02
Fixed 100	–	29.93	45.11	29.75	**50.39**	31.33	64.58	41.85
Fixed 200	–	30.26	43.50	25.75	45.76	30.48	63.14	39.82
Outlier ($\mu + \sigma$)	16.0%	**34.21**	45.69	26.50	47.54	30.14	63.01	41.18
Top-20% entropy	20.0%	33.55	**46.13**	**31.25**	49.55	**32.20**	**65.36**	**43.01**

Table 10 shows that anchor selection is crucial. Unrestricted GRPO-style updating without anchors achieves 40.79 average, while applying the objective to Top-20% entropy anchors improves the score to 43.01. Random and noun-based anchors perform poorly (39.96 and 40.02), even though noun selection uses a larger token fraction than Top-20% entropy selection (28.7% vs. 20.0%). This suggests that effective anchors are not simply visually named content words, but positions with high structural uncertainty and visual leverage.

Top-20% entropy selection is also more robust than fixed or threshold-based alternatives. It outperforms Fixed-100 and Fixed-200 anchors (43.01 vs. 41.85 and 39.82), indicating that a fraction-based rule better adapts to variable sequence lengths. It also outperforms the fixed outlier criterion $c_t > \mu + \sigma$ (43.01 vs. 41.18), supporting our use of a per-trajectory Top-ρ criterion for sparse and adaptive anchor selection.

G.3. Robustness on Decoding Temperature

We further evaluate whether RAPO relies primarily on stochastic sampling or also improves deterministic decoding. Specifically, we compare greedy decoding with temperature 0 and pass@1 against stochastic decoding with temperature 1 and mean@8 on Qwen3-VL-2B-Instruct.

Table 11. Robustness to decoding temperature.

Method	Decoding	MathVision	MathVerse	EMMA	LogicVista	MMMU-Pro	RealWorldQA	Avg
Base	greedy	13.16	26.78	14.50	31.25	18.90	58.82	27.24
Base	temp=1	27.30	36.75	26.87	47.13	24.44	60.47	37.16
GRPO	greedy	11.51	27.41	13.75	32.81	22.49	59.22	27.87
GRPO	temp=1	28.29	44.35	28.50	48.71	30.61	64.28	40.79
RAPO$_G$	greedy	**16.12**	**33.63**	**15.50**	**33.71**	**25.55**	**63.92**	**31.41**
RAPO$_G$	temp=1	**33.55**	**46.13**	**31.25**	**49.55**	**32.20**	**65.36**	**43.01**

GRPO only marginally improves greedy decoding over the base model, from 27.24 to 27.87, while its larger gain appears under stochastic decoding. In contrast, RAPO improves both settings, reaching 31.41 under greedy decoding and 43.01 under temperature-1 sampling. Moreover, the greedy–stochastic gap is smaller for RAPO, 11.46, than for GRPO, 12.92, suggesting that RAPO learns a more stable policy rather than depending mainly on stochastic sampling.

G.4. Robustness on Context Length

Finally, we evaluate robustness under a longer context budget. We set the maximum context length to 16,384 tokens and compare GRPO with RAPO. Under the longer-context setting, RAPO improves the average score from 41.4 to 43.8. This result shows that RAPO remains effective when the context budget is increased and that its gains are not an artifact of a particular maximum generation length.

Table 12. Robustness to longer context length.

Model	MathVision	MathVerse	EMMA	LogicVista	MMMU-Pro	RealWorldQA	Avg
GRPO	31.6	44.0	28.0	51.3	31.2	62.2	41.4
RAPO	**34.5**	**46.2**	**31.5**	**51.6**	**33.0**	**65.8**	**43.8**

H. RAPO based on DAPO

H.1. RAPO$_D$ Optimization Objective

To extend DAPO with RAPO, we incorporate the same $\overline{\mathbb{D}_{\mathrm{KL}}}_{i,t}(\theta)$ used in RAPO$_G$ (Section 4).The complete RAPO$_D$ objective is shown as follows:

$$\mathcal{J}_{\mathrm{RAPO}_D}(\theta) = \mathbb{E}_{\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(O|q,I),t\in\mathcal{A}_i,\mathcal{A}_i\subset o_i}\Bigg[\frac{1}{\sum_{i=1}^{G}|\mathcal{A}_i|}\sum_{i=1}^{G}\sum_{t\in\mathcal{A}_i}\Big\{ \min\Big[r_{i,t}(\theta)\,\hat{A}_{i,t},\ \mathrm{clip}\big(r_{i,t}(\theta),\,1-\epsilon_l,\,1+\epsilon_h\big)\,\hat{A}_{i,t}\Big] + \gamma\,\omega_{i,t}^{\mathrm{br}}\,\overline{\mathbb{D}_{\mathrm{KL}}}_{i,t}(\theta)\Big\}\Bigg], \tag{5}$$

with

$$0 < |\{\, o_i \mid \texttt{is_equivalent}(a, o_i)\,\}| < G.$$

Here, γ is the weighting coefficient for $\omega_{i,t}^{\mathrm{br}}\,\overline{\mathbb{D}_{\mathrm{KL}}}_{i,t}(\theta)$ and $\mathcal{A}_i$ is the set of reflection anchors.

H.2. Results of DAPO Experiments

The DAPO results for the pure RL experiment group are reported in Table 13. The evaluation settings follow the main experiments, except that the temperature is set to 0.6. Results show that RAPO$_D$ still outperforms the baselines.

Table 13. **DAPO Results of pure RL Group.** (mean@8 acc %). Performance comparison across reasoning-intensive and general-domain multimodal benchmarks. The best results in each group are highlighted in **bold**. The second-best results are underlined. RAPO$_D$, PAPO$_D$ and VPPO$_D$ denote RAPO, PAPO and VPPO built upon DAPO, respectively.

Base Model	Model	Reasoning-Intensive				General-Domain		Avg.
		MathVision	MathVerse	EMMA	LogicVista	MMMU-Pro	RealWorldQA	
Qwen3-VL-8B Instruct	Base	41.74	50.21	38.34	62.28	37.59	67.20	49.56
	+ DAPO	50.33	**60.95**	38.75	60.94	51.85	68.30	55.19 (↑ 5.63)
	+ PAPO$_D$	49.67	59.64	42.25	62.95	52.54	71.37	56.40 (↑ 6.84)
	+ VPPO$_D$	50.99	59.64	43.75	61.61	52.60	66.80	55.90 (↑ 6.34)
	+ RAPO$_D$	**51.36**	58.50	**45.50**	**64.51**	**52.83**	**71.63**	**57.39** (↑ 7.83)

I. Licenses of Existing Assets

We summarize the existing external assets used in this work, including codebases, datasets, benchmarks, and models, together with their publicly stated licenses or access terms. We cite the original papers and official repositories, dataset cards, or model cards throughout the paper. All assets are used only for research training, evaluation, and reproducibility purposes. When an asset is distributed under a model-specific license or usage policy, we use it in accordance with the corresponding license agreement, acceptable use policy, and redistribution terms. The results are presented in Table 14.

J. Limitations and Broader Impacts

J.1. Limitations

RAPO provides a practical training-time route to visual-retention optimization in long-chain multimodal reasoning. The current implementation uses native-entropy anchors and a chain-masked finite-window KL target, which keeps the method simple, scalable, and architecture-free at inference, while leaving richer anchor criteria and counterfactual references to future work. Future extensions could incorporate semantic or causal anchor signals, adaptive counterfactual references, and dynamically selected window lengths. Also, RAPO tends to provide larger gains when the base policy exhibits substantial visual forgetting or weak long-chain visual retention. On stronger or architecturally different backbones, the gains may be smaller, suggesting that propagation-aware shaping is most useful when visual-dependence signals are not already well preserved. While our experiments span multiple LVLM backbones and benchmark families, applying propagation-aware visual-retention objectives to broader real-world visual reasoning tasks and more fine-grained faithfulness diagnostics remains an important next step.

Table 14. Existing assets used in this work and their publicly stated licenses or access terms.

Asset	Type	License / Terms of Use
VeRL Github	Code	Apache-2.0
EasyR1 Github	Code	Apache-2.0
vLLM (Kwon et al., 2023)	Code	Apache-2.0
Math-Verify Github	Code	Apache-2.0
ViRL39K (Wang et al., 2025a)	Dataset	MIT
MathVision (Wang et al., 2024a)	Benchmark	MIT
MathVerse (Zhang et al., 2024)	Benchmark	MIT
EMMA (Hao et al., 2025)	Benchmark	No explicit standard dataset license found; research evaluation only
LogicVista (Xiao et al., 2024)	Benchmark	Apache-2.0
MMMU-Pro (Yue et al., 2025)	Benchmark	Apache-2.0
RealWorldQA Huggingface	Benchmark	CC BY-ND 4.0
Qwen3-VL-2B-Instruct (Bai et al., 2025a)	Model	Apache-2.0
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	Model	Apache-2.0
Qwen3-VL-2B-Thinking (Bai et al., 2025a)	Model	Apache-2.0
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	Model	Apache-2.0
Qwen2-VL-7B-Instruct (Wang et al., 2024b)	Model	Apache-2.0
Gemma3-4B-IT (Team et al., 2025a)	Model	Gemma Terms of Use
MM-Eureka (Meng et al., 2025)	Model	Apache-2.0
VL-Rethinker (Wang et al., 2025a)	Model	Apache-2.0
R1-ShareVL (Yao et al., 2025)	Model	Apache-2.0
TVC (Sun et al., 2025b)	Model	Apache-2.0
MINT-CoT (Chen et al., 2025)	Model	No explicit standard license found; research baseline only

J.2. Broader Impacts

This work aims to improve visual grounding in long-chain multimodal reasoning. Positive impacts include more reliable reasoning over diagrams, charts, documents, and real-world visual inputs, especially in settings where models must maintain visual evidence across long reasoning chains. Potential negative impacts follow from improving the capability of general-purpose LVLMs: stronger multimodal reasoning could be misused for more persuasive automated content generation, surveillance-oriented visual analysis, or incorrect high-confidence decisions in sensitive domains. RAPO is evaluated as a research training method and is not intended for deployment in safety-critical settings without task-specific risk assessment, calibration, and human oversight. We do not release a new dataset or high-risk model checkpoint; released materials are intended for research reproduction.

K. Case Study

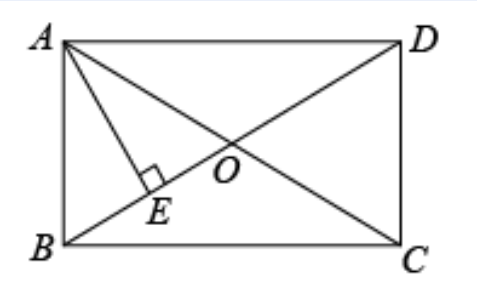
K.1. Visualization of Entropy Distribution on Tokens

Figure 9a and Figure 9b show the distribution of token entropy. We observe that high-entropy tokens are more concentrated around key reasoning steps, indicating increased uncertainty at decision points where multiple continuations remain plausible.

K.2. Visualization of KL Distribution before and after RAPO training

Figure 10 and Figure 11 show the KL distributions before and after RAPO training. Compared to the base model, RAPO consistently amplifies KL values in the later stages of generation, suggesting enhanced and sustained utilization of visual information throughout the reasoning process.

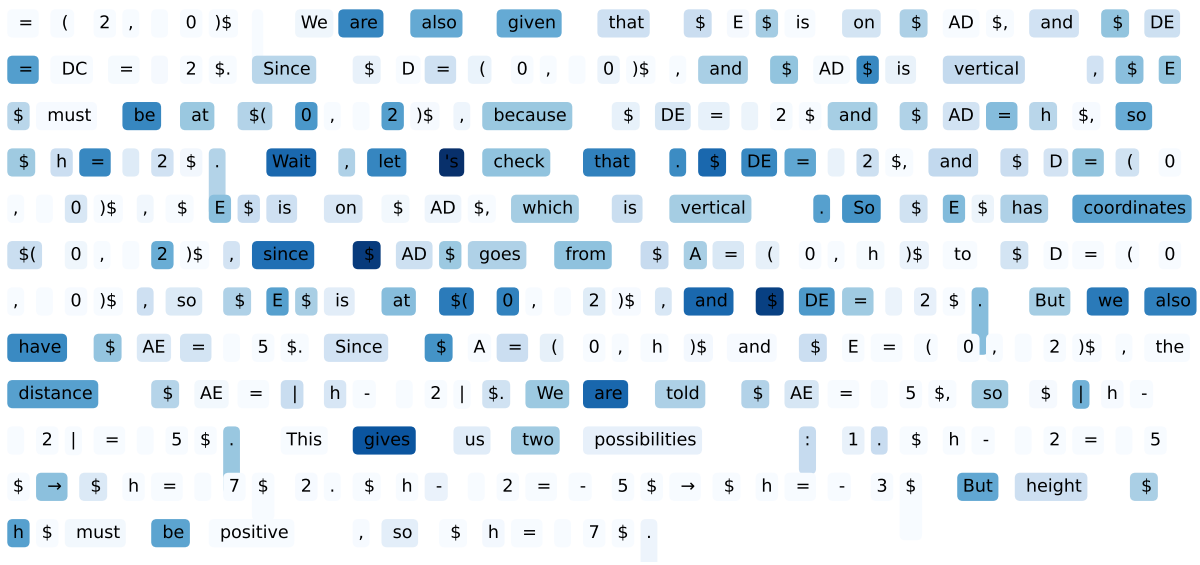
Case Study



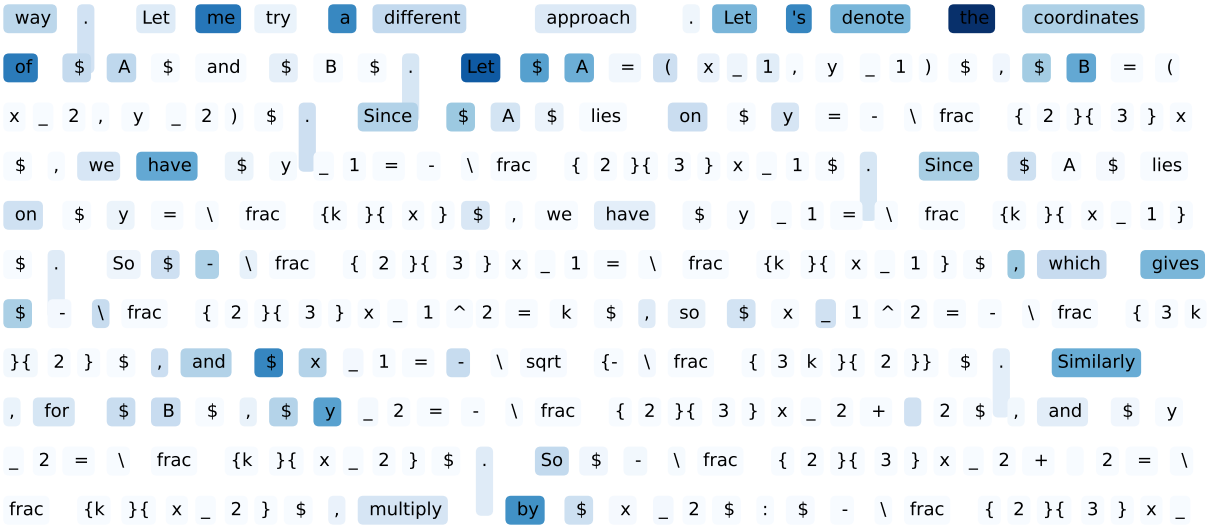
Question

As shown in the figure, within rectangle $ABCD$, diagonals AC and BD intersect at point O, AE perpendicular bisects BO, with $AE = \sqrt{3}$, cm, what is the length of OD?

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think>` `</think>` tags. The final answer MUST BE put in `\boxed{}`.



(a) **Case 1.**



(b) **Case 2.**

Figure 9. **Case Study of Entropy Distributions along CoT.**

Internal Mon ologue : ** First , let 's understand the given information . We have a rectangle $ AB CD $ with diagon als $ AC $ and $ BD $ intersect ing at point $ O $. Since it 's a rectangle , the diagon als are equal in length and bis ect each other . So , $ AO = OC $ and $ BO = OD $. We 're told that $ AE $ is perpendicular to $ BO $ and that $ AE $ bis ect s $ BO $. This means that $ AE $ is the perpendicular bis ector of $ BO $. Since $ AE $ is perpendicular to $ BO $ and passes through the midpoint of $ BO $, it implies that $ E $ is the midpoint of $ BO $. Given that $ AE = \ sqrt { 3 }$ cm , and since $ AE $ is the perpendicular bis ector of $ BO $, we can use the properties of right triangles and the Py thag orean theorem . Let 's denote $ BO = 2 x $. Then , since $ E $ is the midpoint of $ BO $, $ BE = EO = x $. Since $ AE $ is perpendicular to $ BO $, triangle $ A EB $ is a right triangle with right angle at $ E $. Now , in right triangle $ A EB $, we have : - $ AE = \ sqrt { 3 } $ - $ BE = x $ - $ AB $ is the hyp oten use of triangle $ A EB $ But we don 't know $ AB $ yet . However , we can use the fact that $ AO $ is half of the diagonal $ AC $, and since $ AC = BD $, and $ O $ is the midpoint , $ AO = \ frac { 1 }{ 2 } AC $. But we can also consider triangle $ AO B $. Since $ AO = BO $ (because diagon als of a rectangle bis ect each other and are equal), we have $ AO = BO = 2 x $. So , $ AO = 2 x $. Now , in triangle $ AO B $, we can use the Py thag orean theorem in triangle $ AO B $: $$ AB ^ 2 = AO ^ 2 + BO ^ 2 = (2 x)^ 2 + (2 x)^ 2 = 4 x ^ 2 + 4 x ^ 2 = 8 x ^ 2 $$ So , $ AB = \ sqrt { 8 x ^ 2 } = 2 \ sqrt { 2 } x $ Now , in triangle $ A EB $, we have : - $ AE = \ sqrt { 3 } $ - $ BE = x $ - $ AB = 2 \ sqrt { 2 } x $ Using the Py thag orean theorem in triangle $ A EB $: $$ AB ^ 2 = AE ^ 2 + BE ^ 2 $$ $$ (2 \ sqrt { 2 } x)^ 2 = (\ sqrt { 3 }) ^ 2 + x ^ 2 $$ $$ 8 x ^ 2 = 3 + x ^ 2 $$ $$ 8 x ^ 2 - x ^ 2 = 3 $$ $$ 7 x ^ 2 = 3 $$ $$ x ^ 2 = \ frac { 3 }{ 7 } $$ $$ x = \ sqrt {\ frac { 3 }{ 7 }} $$ Now , we want to find $ OD $. Since $ OD = BO = 2 x $, we have : $$ OD = 2 x = 2 \ sqrt {\ frac { 3 }{ 7 }} = \ sqrt {\ frac { 4 \ cdot 3 }{ 7 }} = \ sqrt {\ frac { 1 2 }{ 7 }} $$ So , the length of $ OD $ is $\ boxed {\ sqrt {\ frac { 1 2 }{ 7 }} }$ cm . Wait , let me double -check this . I think I made a mistake in the reasoning . Let me re -ex amine . Actually , I think I made a mistake in assuming $ AO = BO $. That 's not correct . In a rectangle , the diagon als are equal and bis ect each other , so $ AO = OC $ and $ BO = OD $. But $ AO $ and $ BO $ are not necessarily equal . They are equal only if the rectangle is a square . Wait , no . In a rectangle , diagon als are equal and bis ect each other , so $ AO = OC $ and $ BO = OD $. So $ AO = OC $ and $ BO = OD $. But $ AO $ and $ BO $ are not necessarily equal unless the rectangle is a square . So , let 's correct that . Let me re -ex press . Let ’s denote $ BO = 2 x $. Then $ OD = 2 x $ because $ O $ is the midpoint of $ BD $. So $ BO = OD = 2 x $. Now , $ AE $ is perpendicular to $ BO $ and $ E $ is the midpoint of $ BO $, so $ BE = EO = x $. Now , in triangle $ A EB $, we have : - $ AE = \ sqrt { 3 } $ - $ BE = x $ - $ AB $ is the hyp oten use But we also know that $ AB $ is a side of the rectangle , and $ AB $ is related to $ AO $ and $ BO $. Wait , in triangle $ AO B $, we have : - $ AO = \ frac { 1 }{ 2 } AC $ - $ BO = \ frac { 1 }{ 2 } BD $ - Since $ AC = BD $, $ AO = BO $ So $ AO = BO $. So triangle $ AO B $ is is os ce les with $ AO = BO $. So $ AO = BO = 2 x $. So $ AO = 2 x $, $ BO = 2 x $. Now , in triangle $ A EB $, we have : - $ AE = \ sqrt { 3 } $ - $ BE = x $ - $ AB $ is the hyp oten use But $ AB $ is the side of the rectangle , and $ AB $ is related to $ AO $ and $ BO $. Wait , in triangle $ AO B $, we have : - $ AO = 2 x $ - $ BO = 2 x $ - $ AB $ is the hyp oten use So $ AB ^ 2 = AO ^ 2 + BO ^ 2 = (2 x)^ 2 + (2 x)^ 2 = 8 x ^ 2 $ So $ AB = \ sqrt { 8 x ^ 2 } = 2 \ sqrt { 2 } x $ Now , in triangle $ A EB $, we have : - $ AE = \ sqrt { 3 } $ - $ BE = x $ - $ AB = 2 \ sqrt { 2 } x $ Using the Py thag orean theorem : $$ AB ^ 2 = AE ^ 2 + BE ^ 2 $$ $$ (2 \ sqrt { 2 } x)^ 2 = (\ sqrt { 3 }) ^ 2 + x ^ 2 $$ $$ 8 x ^ 2 = 3 + x ^ 2 $$ $$ 7 x ^ 2 = 3 $$ $$ x ^ 2 = \ frac { 3 }{ 7 } $$ $$ x = \ sqrt {\ frac { 3 }{ 7 }} $$ Now , $ OD = BO = 2 x = 2 \ sqrt {\ frac { 3 }{ 7 }} = \ sqrt {\ frac { 4 \ cdot 3 }{ 7 }} = \ sqrt {\ frac { 1 2 }{ 7 }} $ So $ OD = \ sqrt {\ frac { 1 2 }{ 7 }} $ cm . But let me check if this makes sense . Is $ AE $ perpendicular to $ BO $? Yes , because $ AE $ is the perpendicular bis ector of $ BO $, so it is perpendicular to $ BO $. And $ E $ is the midpoint of $ BO $, so $ BE = EO = x $. So the calculation seems correct . Therefore , the length of $ OD $ is $\ boxed {\ sqrt {\ frac { 1 2 }{ 7 }} }$ cm .

Figure 10. **Token KL Distribution of Base Model (Wrong Answer).**

Internal Mon ologue : ** First , let 's understand the given information . We have a rectangle $ AB CD $ with diagon als $ AC $ and $ BD $ intersect ing at point $ O $. This means $ O $ is the midpoint of both diagon als . In a rectangle , the diagon als are equal in length and bis ect each other . So , $ AO = OC $ and $ BO = OD $. We 're told that $ AE $ is perpendicular to $ BO $ and that $ AE $ bis ect s $ BO $. This means that $ E $ is the midpoint of $ BO $, and $ AE \ per p BO $. Also , $ AE = \ sqrt { 3 }\ ,\ text { cm }$. Let 's consider triangle $ AO B $. Since $ O $ is the midpoint of $ BD $, and $ AE $ is perpendicular to $ BO $, and $ E $ is the midpoint of $ BO $, we can use coordinate geometry to solve this . Let ’s place the rectangle on the coordinate plane . Let ’s assume : - $ B $ is at $(0 , 0)$ - $ C $ is at $(c , 0)$ - $ D $ is at (c , d) - $ A $ is at $(0 , d)$ Then , the diagon als are : - $ AC $: from $(0 , d)$ to $(c , 0)$ - $ BD $: from $(0 , 0)$ to (c , d) The intersection point $ O $ is the midpoint of both diagon als . So , $$ O = \ left (\ frac { 0 +c }{ 2 }, \ frac { d + 0 }{ 2 }\ right) = \ left (\ frac { c }{ 2 }, \ frac { d }{ 2 }\ right) $$ Now , let 's find the coordinates of $ E $. Since $ AE $ is perpendicular to $ BO $ and $ E $ is the midpoint of $ BO $, we can find $ E $. First , find the vector $ BO $: $$ \ vec { BO } = \ left (\ frac { c }{ 2 }, \ frac { d }{ 2 }\ right) $$ Since $ E $ is the midpoint of $ BO $, its coordinates are : $$ E = \ left (\ frac { 0 + \ frac { c }{ 2 }} { 2 }, \ frac { 0 + \ frac { d }{ 2 }} { 2 }\ right) = \ left (\ frac { c }{ 4 }, \ frac { d }{ 4 }\ right) $$ Now , we can find the vector $ AE $: $$ \ vec { AE } = E - A = \ left (\ frac { c }{ 4 } - 0 , \ frac { d }{ 4 } - d \ right) = \ left (\ frac { c }{ 4 }, - \ frac { 3 d }{ 4 }\ right) $$ We are told that $ AE \ per p BO $. So , the dot product of $\ vec { AE }$ and $\ vec { BO }$ is zero . $$ \ vec { AE } \ cdot \ vec { BO } = \ left (\ frac { c }{ 4 }\ right)\ left (\ frac { c }{ 2 }\ right) + \ left (- \ frac { 3 d }{ 4 }\ right)\ left (\ frac { d }{ 2 }\ right) = 0 $$ $$ \ frac { c ^ 2 }{ 8 } - \ frac { 3 d ^ 2 }{ 8 } = 0 $$ $$ \ frac { c ^ 2 - 3 d ^ 2 }{ 8 } = 0 $$ $$ c ^ 2 = 3 d ^ 2 $$ Now , we know that $ AE = \ sqrt { 3 }\ ,\ text { cm } $. Let 's find the length of $ AE $ using the coordinates : $$ AE = \ sqrt {\ left (\ frac { c }{ 4 }\ right)^ 2 + \ left (- \ frac { 3 d }{ 4 }\ right)^ 2 } = \ sqrt {\ frac { c ^ 2 }{ 1 6 } + \ frac { 9 d ^ 2 }{ 1 6 }} = \ frac { 1 }{ 4 }\ sqrt { c ^ 2 + 9 d ^ 2 } $$ We already have $ c ^ 2 = 3 d ^ 2 $, so substitute : $$ AE = \ frac { 1 }{ 4 }\ sqrt { 3 d ^ 2 + 9 d ^ 2 } = \ frac { 1 }{ 4 }\ sqrt { 1 2 d ^ 2 } = \ frac { 1 }{ 4 } \ cdot 2 \ sqrt { 3 } d = \ frac {\ sqrt { 3 }} { 2 } d $$ We are given $ AE = \ sqrt { 3 }$, so : $$ \ frac {\ sqrt { 3 }} { 2 } d = \ sqrt { 3 } $$ Div ide both sides by $\ sqrt { 3 }$: $$ \ frac { 1 }{ 2 } d = 1 $$ So , $ d = 2 $ Now , we want to find $ OD $. Since $ O $ is at $\ left (\ frac { c }{ 2 }, \ frac { d }{ 2 }\ right)$ and $ D $ is at (c , d) , the distance $ OD $ is : $$ OD = \ sqrt {\ left (c - \ frac { c }{ 2 }\ right)^ 2 + \ left (d - \ frac { d }{ 2 }\ right)^ 2 } = \ sqrt {\ left (\ frac { c }{ 2 }\ right)^ 2 + \ left (\ frac { d }{ 2 }\ right)^ 2 } $$ We know $ d = 2 $, so : $$ OD = \ sqrt {\ left (\ frac { c }{ 2 }\ right)^ 2 + \ left (1 \ right)^ 2 } $$ But we also have $ c ^ 2 = 3 d ^ 2 = 3 (4) = 1 2 $, so $ c = 2 \ sqrt { 3 }$ Sub stitute : $$ OD = \ sqrt {\ left (\ frac { 2 \ sqrt { 3 }} { 2 }\ right)^ 2 + 1 ^ 2 } = \ sqrt { (\ sqrt { 3 }) ^ 2 + 1 } = \ sqrt { 3 + 1 } = \ sqrt { 4 } = 2 $$ So , the length of $ OD $ is $ 2 \ ,\ text { cm }$. ** Final Answer : ** \ boxed { 2 }

Figure 11. **Token KL Distribution after RAPO Training (Right Answer).**