Reflection Anchors for Propagation-Aware Visual Retention in Long-Chain Multimodal Reasoning

Xuan Gong[1,2] **Hanbo Huang**[1] **Hao Zheng**[1] **Yiran Zhang**[1] **Wenbin Dai**[1,2]
Weishu Zhao[1,2] **Shiyu Liang**[1,2,†]
[1]Shanghai Jiao Tong University [2]Shanghai AI Laboratory
Code Project Page

Abstract

Long chain-of-thought (CoT) reasoning improves large vision–language models, but visual information often fades during generation, limiting long-horizon multimodal reasoning. Existing methods either re-inject vision at inference or train policies for stronger grounding, but *where* to intervene relies on perception heuristics rather than principled gain analysis, and *how* local visual influence propagates remains implicit. We study this problem from an information-theoretic standpoint and derive a lower bound on the downstream visual gain of a one-step intervention, which suggests two factors: local branching room (token entropy) and downstream visual propagation potential (suffix divergence from a vision-marginalized reference). Guided by this analysis, we propose reflection-anchor policy optimization **(RAPO)**, a GRPO-based policy optimization method that selects high-entropy reflection anchors and optimizes a chain-masked finite-window KL surrogate for downstream visual dependence. Experiments on reasoning-intensive and general-domain benchmarks show that RAPO delivers substantial gains over strong baselines across multiple LVLM backbones. Mechanism analyses further indicate that reflection anchors are enriched for visually sensitive decision points and that RAPO increases contrastive visual-dependence signals along generated trajectories.

1 Introduction

Test-time scaling via long Chain-of-Thought (CoT) [1, 2] has emerged as a cornerstone of recent advancements in Large Language Models (LLMs) [3–5]. Inspired by this, recent works have adapted CoT paradigms to Large Vision-Language Models (LVLMs) [6–9]. However, unlike text-only LLMs that effectively maintain context through linguistic recurrence, LVLMs struggle to sustain visual information across extended reasoning sequences. This "visual forgetting" reflects a mismatch between prefix-based visual embeddings and visual reasoning demands [10–12].

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 [11, 13–20]. 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 [21–23], 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:

[†]Corresponding author. Correspondence to lsy18602808513@sjtu.edu.cn.
Technical inquiries: Xuan Gong (gongxuan0610@sjtu.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

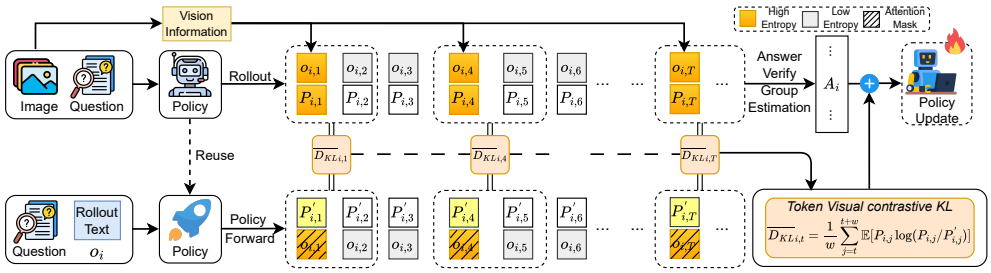


Figure 1: **Overview of RAPO.** Given an input, the policy generates the i-th reasoning rollout $\{o_{i,t}, P_{i,t}\}_{t=1}^{T}$, where $o_{i,t}$ is the t-th generated token and $P_{i,t}$ is the corresponding logit-induced next-token distribution. A chain-masked forward pass produces vision-masked distributions $\{P'_{i,t}\}_{t=1}^{T}$. 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.

Can visual dependence be preserved by minimally reshaping native decoding at sparse steps, each selected to maximize downstream influence on future reasoning?

We formulate this as optimizing the downstream visual-dependence gain induced by in-place intervention of the next-token distribution. We derive a lower bound showing that nontrivial visual gains arise at tokens where (1) the next-token distribution has high entropy, leaving room for effective local intervention, and (2) the future continuation remains distributionally separated from its visual-marginalized counterpart, allowing visual influence to propagate. This suggests a simple algorithmic principle: select intervention sites by uncertainty, and train interventions by propagation. We call the selected positions *reflection anchors*—a name that reflects an empirical observation: high-entropy positions coincide with reasoning-procedural tokens such as "examine" and "reconsider." Based on this principle, we propose **RAPO** (Figure 1), a GRPO-based policy optimization algorithm [24] that selects reflection anchors by reference-policy entropy and optimizes a branching-room-weighted contrastive KL between normal continuations and their chain-masked counterfactuals. Thus, RAPO realizes *propagation-aware sparse policy reshaping* without modifying the inference-time decoder.

Our contributions are as follows:

- We provide an information-theoretic view on the downstream visual-dependence gain in long CoT reasoning, characterizing where and how in-place intervention can shape future reasoning.
- Based on the principle, we propose **RAPO**, a GRPO-based reinforcement learning algorithm that targets reflection anchors and reinforces propagation of visual influence across reasoning steps.
- Extensive experiments on various benchmarks show that RAPO outperforms baselines across multiple base models in long-chain multimodal reasoning.
- Empirical studies highlight the contribution of each component and provide guidance for effective downstream visual-dependence gain.

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 [16] inserts image–text rationales, and MINT-CoT [15] selects visual tokens on the fly. CoVT [18] keeps the same idea but moves the evidence into a continuous visual-token space. Decoder-side methods inject vision internally. TVC [11] reintroduces image conditioning at selected stages, while v1 [17] copies image-patch representations into decoding. Adaptive methods further decide when to restore vision: Qwen-LA [14] learns when to revisit the image, and Reflection-V [13] 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 [24] and DAPO [25], and extends them to multimodal reasoning through methods such as Vision-R1 [8] and MM-Eureka [9]. Beyond final-answer rewards, VL-Rethinker [26], SRPO [27], and Mulberry [28] improve intermediate trajectories through reflection or refinement. More recent perception-aware methods, including PAPO [21], VPPO [22], PeRL [29], and VAPO [30], 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: Selective Policy Reshaping for Visual Retention

Visual reasoning. Let V denote the visual input, X the textual prompt, and $Y_{1:T}$ the reasoning trajectory[1]. 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 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.

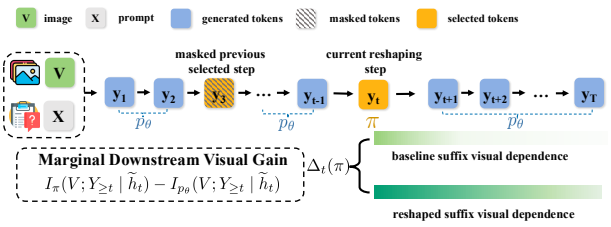


Figure 2: **Marginal downstream visual gain.**

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 KL radius. We solve

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

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

[1] In this paper, uppercase and lowercase letters denote the random variables and their realizations, respectively.

constraint keeps each deviation local and close to the native decoder. Thus, visual retention is cast as sparse instance-level reshaping: change only a few steps, but make each change propagate.

4 Methodology

4.1 Reflection Anchors: 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 local intervention 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 establish a lower bound on the optimal downstream visual gain under a local KL constraint:

$$\Delta_t^\star(A_{<t}, \varepsilon) \geq \left[c_1 \sqrt{\varepsilon}\, \mathbb{E}_{V|\tilde{h}_t} \left[\left(e^{H(m_{t,V})} - 2 \right)_+ \cdot \sqrt{D_{\mathrm{KL}}(p_{t,V} \| p_{M,t})} \right] - c_2 \varepsilon \right]_+ ,$$

where $(a)_+ = \max\{a, 0\}$ and $c_1, c_2 > 0$ depend only on the regularity constants. The bound captures a leading-order $\sqrt{\varepsilon}$ contribution under sufficiently small KL perturbations.

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 $\mathcal{A}$. Selecting the highest-entropy positions maximizes the aggregate branching-room score under a fixed intervention budget. This yields an efficient and principled localization strategy, with the subsequent optimization determining how visual information propagates from the selected anchors.

Reflection anchors. 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 $\mathcal{A}_\rho = \mathrm{TopK}_{t \in [T]} \{c_t\}_{K_\rho}$, Here $\mathrm{TopK}\{\cdot\}_K$ returns any set of K indices with the largest scores. This gives a one-pass selector: choose the steps where the native decoder still has branching room, then let training decide how to reshape them.

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 $t \in \mathcal{A}$.

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.

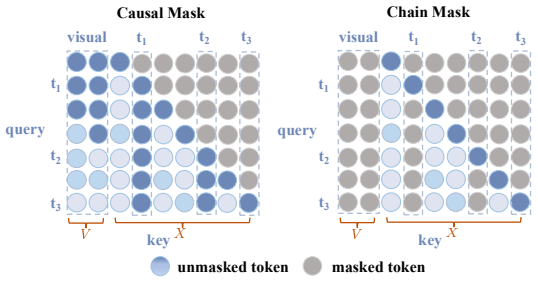


Figure 3: **Chain mask.** For an anchor $t_k \in \mathcal{A}$, the chain mask $\mathcal{M}_{t_k}$ enforces three constraints: (i) the query at t_k cannot attend to the visual tokens V; (ii) it cannot attend to any preceding anchors $\{t_1, \ldots, t_{k-1}\}$; and (iii) attention to all other textual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

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 $\overline{\mathbb{D}}_{\mathrm{KL}_{t_k}}(\theta) = \frac{1}{w'} \sum_{i=0}^{w'-1} \mathbb{D}_{\mathrm{KL}} \left(\pi_\theta(\cdot \mid v, h_{t_k+i}) \,\|\, \pi_\theta^{\mathrm{mask}}(\cdot \mid h_{t_k+i}; \mathcal{M}_{t_k}) \right)$. 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

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

where $\mathcal{L}_{i,t}^{\mathrm{clip}}(\theta) = \min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \mathrm{clip}\left(r_{i,t}(\theta), 1 - \varepsilon_{\mathrm{clip}}, 1 + \varepsilon_{\mathrm{clip}}\right)\hat{A}_{i,t} \right)$, $\mathcal{A}_i$ 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 $\omega_{i,t}^{\mathrm{br}}$ are stop-gradient.

Connection to prior work. RAPO combines the explicitness of visual restoration with the clean interface of policy optimization. Visual-restoration methods repair forgetting by revisiting or re-injecting vision during decoding [16, 15, 13], while policy-optimization methods internalize grounding into normal generation [24, 25, 21, 22]. RAPO keeps the latter interface but adds a local retention mechanism: entropy-selected anchors specify where a KL-bounded update has room to matter, and the branching-room-weighted chain-masked KL rewards whether visual dependence persists over downstream continuations. Thus, visual retention is trained as selective policy reshaping inside the native decoder, rather than added as inference-time repair or left as a global grounding reward.

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
	+ GRPO	25.33	43.40	28.00	44.42	32.83	56.21	38.37 (↑ 6.43)
	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
	+ GRPO	21.05	31.85	24.25	35.27	27.11	61.44	33.50 (↑ 17.75)
	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)

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 [31, 32, 7] 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 [7], and compare it with representative RL–based optimization methods, including GRPO [24], PAPO [21], and VPPO [22]. 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 [9], VL-Rethinker [26], and R1-ShareVL [33]. 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 [11] and MINT-CoT [15], 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 [26], which is designed for long CoT multimodal reasoning. Across all

experiments, we use a learning rate of $1e{-}6$. 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 [34] on six multimodal benchmarks spanning two categories. (1) **Reasoning-intensive (Reas.)** benchmarks assess multimodal reasoning ability and include MathVision [35], MathVerse [36], EMMA [37] and LogicVista [38]. (2) **General-domain (Gen.)** benchmarks evaluate overall ability across diverse domains and include MMMU-Pro [39] and RealWorldQA[2]. 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 [21, 22].

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.

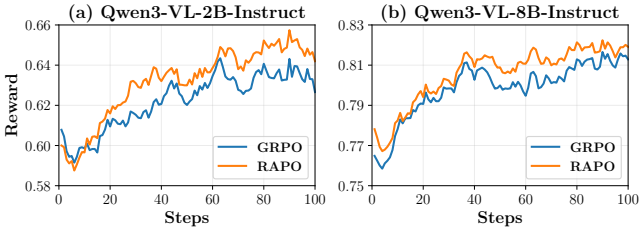


Figure 4: **Training dynamics.** Reward curves of GRPO and RAPO on Qwen3-VL-2B-Instruct and 8B during RL.

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 information is used, where local intervention is effective, and how its influence propagates during generation. All analyses 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.

Our mechanism study combines trajectory-level visual-dependence diagnostics with targeted interventions. The trajectory analysis measures token-wise contrastive KL throughout generation, complementing RAPO's anchor-focused, finite-window training objective.

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

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

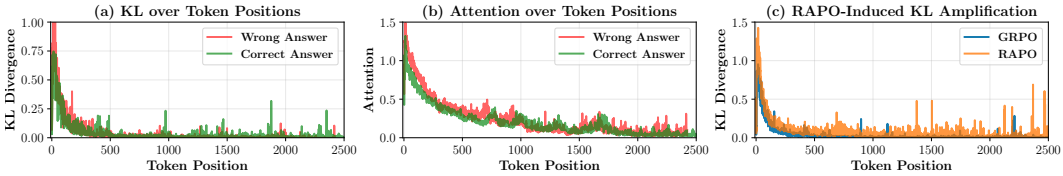
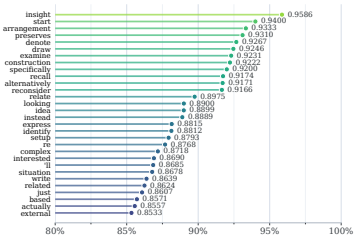
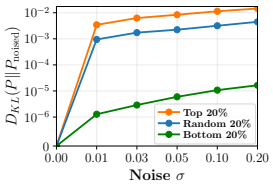


Figure 5: **Trajectory analysis. (a, b)** Visual dependence in correct vs. incorrect trajectories, measured by token-wise KL and attention; **(c)** Token-wise KL after GRPO and RAPO training. The KL at token t, $\overline{\mathbb{D}_{\mathrm{KL}t}}$, is the contrastive divergence between the distributions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



(a) Anchor Token Concentration.



(b) D_{KL} vs. Visual Noise.

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%	37.00	45.90	39.96

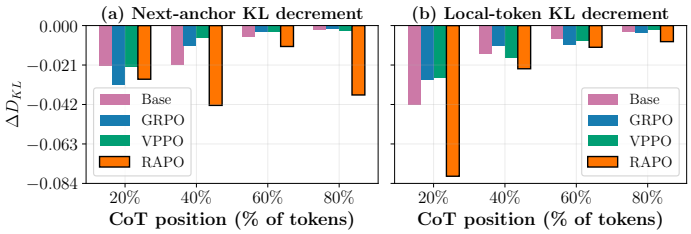
(c) Anchor-Selection Ablation.

Figure 6: **Reflection anchor analysis. (a)** Reflection-anchor concentration, measured as the fraction of each token type's occurrences selected as top-entropy anchors. **(b)** KL divergence between original and noise-perturbed token distributions across entropy groups. **(c)** Analysis of anchor token selection strategies based on entropy ranking and token position.

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 contrastive visual-dependence signals 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 locate the positions whose local reshaping most plausibly carries visual evidence into future reasoning. We analyze the positions selected by the entropy-based reflection-anchor rule. Figure 6a shows that high-entropy anchors are concentrated on tokens that control the next reasoning operation, including explicit inspection or revision verbs such as "examine," "reconsider," "identify," and "draw," but also discourse and relational operators such as "actually," "instead," "based," and "specifically." Such tokens often mark points where the model decides whether to inspect, revise, contrast, relate, or reparameterize the current reasoning state. This supports the view that reflection anchors occur at operation-selecting positions, where local reshaping can alter how visual evidence is carried forward. Figure 6b further shows that high-entropy tokens consistently exhibit larger KL divergence under visual perturbations than random or low-entropy tokens. Correspondingly, restricting RAPO's updates to the top 20% high-entropy tokens outperforms random selection, low-entropy tokens, and alternative anchor set sizes (Table 6c). Beyond these, Appendix G.2 evaluates noun-based, fixed-count, and outlier-threshold selectors. Together, these results indicate that RAPO intervenes at naturally occurring, decision-uncertain steps that are effective for local reshaping, in contrast to visual-injection methods that rely on fixed or predefined positions.

RAPO strengthens downstream contrastive visual-dependence signals beyond reflection anchors. RAPO first selectively amplifies visual dependence at reflection anchors: after training, the average contrastive KL at the top 20% entropy tokens rises from 0.30 to 0.55, while the bottom 20% entropy tokens slightly decrease from 0.06 to 0.05. We then study whether this local amplification propagates by masking visual tokens at selected anchors and measuring the induced KL decrement at the next anchor and over subsequent local tokens. Figure 7 shows that RAPO yields the largest downstream KL reductions in both cases, indicating that visual influence reinforced at reflection anchors persists



(c) Chain-mask ablation.

Mask	MathVis.	MathVer.	EMMA
Image	26.64	44.80	27.00
Chain	**33.55**	**46.13**	**31.25**

Mask	LogicV.	MMMU-P.	RealWQ
Image	47.32	**32.79**	64.02
Chain	**49.55**	32.20	**65.36**

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. **(c)** Comparison between image-level masking and the proposed chain mask.

into the future suffix rather than remaining a local token effect. A chain-mask ablation further shows that replacing chain masking with image-level masking lowers the average score from 43.01 to 40.43, indicating that structured anchor-chain masking provides a more effective propagation signal than blanket image removal. Together with intervention-based causal diagnostics in Appendix E, these results support the interpretation that RAPO reinforces not only local visual dependence at anchor positions, but also its propagation into the future suffix, yielding a more explicit propagation mechanism than prior RL methods whose visual influence remains localized or implicit.

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 2a 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 2b 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. We further evaluate two adaptive window strategies based on entropy-gated stopping and inter-anchor boundaries. The fixed three-token window achieves the highest average accuracy (43.01), outperforming both entropy-gated (41.32) and inter-anchor (42.04) variants. This validates a compact propagation horizon as an effective design choice for RAPO. Detailed comparisons are provided in Appendix G.5.

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 (∼6k vs. ∼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 2c). We further evaluate RAPO on Gemma3-4B-IT [40], whose architecture differs from the Qwen series used in the main results. RAPO outperforms GRPO (Table 2c), indicating that its effectiveness is not tied to a specific reasoning length or model family.

Table 2: **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

6 Conclusion

We frame visual retention in long-chain multimodal reasoning as selective policy reshaping: local deviations from the native decoder should make future continuations remain visually dependent. This led to RAPO, an RL method that updates high-entropy reflection anchors and optimizes a chain-masked finite-window target for downstream visual dependence, without architectural changes or explicit visual-token injection at inference. Experiments across multiple LVLM backbones and multimodal benchmarks show that RAPO improves average performance over strong baselines, while mechanism analyses indicate that entropy anchors correspond to visually sensitive decision points and that RAPO strengthens downstream visual-dependence signals. The results point to propagation-aware visual-retention optimization as a useful direction for long-CoT LVLM reasoning. We discuss the current scope and future extensions of RAPO in Appendix J.

Acknowledgments

This work was supported by the National Key R&D Program of China (2025YFF0516900 & 2025YFF0516904), NSFC U25B2039 and National Natural Science Foundation of China (No. 62306179).

References

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[2] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 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.

[4] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming

Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2025. URL https://arxiv.org/abs/2507.20534.

[5] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[6] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report, 2025. URL https://arxiv.org/abs/2504.07491.

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

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

[9] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

[10] Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulic, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. *arXiv preprint arXiv:2509.11986*, 2, 2025.

[11] Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. *arXiv preprint arXiv:2503.13360*, 2025.

[12] Ahmed Masry, Juan A. Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-André Noël, Sathwik Tejaswi Madhusudhan, Marco Pedersoli, Bang Liu, Nicolas Chapados, Yoshua Bengio, Enamul Hoque, Christopher Pal, Issam H. Laradji, David Vazquez, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. Alignvlm: Bridging vision and language latent spaces for multimodal document understanding, 2025. URL https://arxiv.org/abs/2502.01341.

[13] Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9262–9281, 2025.

[14] Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information. *arXiv preprint arXiv:2505.23558*, 2025.

[15] Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. 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.

[16] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025.

[17] Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. v1: Learning to point visual tokens for multimodal grounded reasoning, 2025. URL https://arxiv.org/abs/2505.18842.

[18] Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

[19] Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi WANG, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, and Libo Qin. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=xPcKmKSEis.

[20] Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning, 2025. URL https://arxiv.org/abs/2510.23925.

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

[22] Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning, 2025. URL https://arxiv.org/abs/2510.09285.

[23] Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. Rethinking token-level policy optimization for multimodal chain-of-thought, 2026. URL https://arxiv.org/abs/2603.22847.

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

[25] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu,

and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

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

[27] Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*, 2025.

[28] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

[29] Yizhen Zhang, Yang Ding, Shuoshuo Zhang, Xinchen Zhang, Haoling Li, Zhong-zhi Li, Peijie Wang, Jie Wu, Lei Ji, Yelong Shen, et al. Perl: Permutation-enhanced reinforcement learning for interleaved vision-language reasoning. *arXiv preprint arXiv:2506.14907*, 2025.

[30] Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Henry Tu, and Jing Zhang. 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.

[31] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

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

[33] Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and Jiaxing Huang. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo, 2025. URL https://arxiv.org/abs/2505.16673.

[34] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

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

[36] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. URL https://arxiv.org/abs/2403.14624.

[37] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, 2025. URL https://arxiv.org/abs/2501.05444.

[38] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024. URL https://arxiv.org/abs/2407.04973.

[39] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. URL https://arxiv.org/abs/2409.02813.

[40] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

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) [16] interleaves paired visual-textual rationales with intermediate reasoning steps, maintaining access to visual evidence during multi-hop inference. Similarly, MINT-CoT [15] employs interleaved image-text sequences with dynamic visual token selection to support long-context reasoning. From a representation perspective, Chain-of-Visual-Thought (CoVT) [18] 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 [19] 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 [20] 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 [11] 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 [17] 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 [14] introduces a vision-text reflection process and optimizes the timing of visual re-examination via reinforcement learning, complemented by visual token routing. Reflection-V [13] 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) [24], which establishes a baseline for group-based advantage estimation without value functions. Building on this, Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) [25] 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 [8] proposes *progressive thinking suppression training* to reduce overthinking during cold starts, encouraging efficient reasoning trajectories. Similarly, MM-Eureka [9] 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 [26] 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) [27] 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 [28] 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) [21] introduces an *Implicit Perception Loss* to anchor reasoning steps to visual inputs, thereby reducing perception errors. VPPO (Visually-Perceptive Policy Optimization) [22] further refines this by analyzing token-level visual dependency to prioritize perceptually pivotal tokens. Extending to more complex inputs, PeRL [29] 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

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

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 $[T] = \{1, \dots, T\}$. 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

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

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

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

Equivalently,

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

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

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

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

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

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*

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

where

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

Moreover,

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

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

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

Define the averaged conditional score standard deviation

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

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 $\eta_t^\star(v) = 1/\lambda_v$ 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 $y \in \mathrm{Supp}(m_{t,v})$,

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

Therefore the centered score satisfies

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

Let

$$M = 2B.$$

For a fixed v, consider the centered exponential family

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

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 $M\eta_v \le s_v$,

$$\mathbb{D}_{\mathrm{KL}}(q_{v,\eta_v} \| m_{t,v}) \le \frac{\sigma_t^2(v) a_v^2}{2}(1 - d_0 s_v)^2(1 + d_0 s_v).$$

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 $C_B \ge 2M \max\{d_0 + 1, d_0\}$, 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,

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

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

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

□

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*

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

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

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

Equivalently,

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

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

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

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*

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

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,

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

Substituting this into the previous display yields

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

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

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

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 $|\mathcal{A}| \leq K_\rho$,*

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

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

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

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

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

(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.

Connection to finite-window propagation. We further connect the continuation-level divergence to RAPO's finite-horizon objective through the autoregressive KL chain rule. For two continuation distributions $P(Y_{t:T})$ and $Q(Y_{t:T})$, we have

$$\begin{aligned} &D_{\mathrm{KL}}(P(Y_{t:T}) \| Q(Y_{t:T})) \\ &= \sum_{i=t}^{T} \mathbb{E}_{Y_{t:i-1} \sim P} \left[D_{\mathrm{KL}} \left(P(Y_i \mid Y_{t:i-1}) \| Q(Y_i \mid Y_{t:i-1}) \right) \right]. \end{aligned} \tag{25}$$

For a fixed chain mask associated with an anchor, this decomposition expresses continuation-level visual contrast as the accumulated conditional next-token divergences. RAPO uses a finite-horizon, length-normalized trajectory-level estimate of these terms, concentrating optimization on the local continuation following each anchor while reducing overlap between different anchor windows.

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

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

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,

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

Using

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

we rewrite this as

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

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

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

□

C Experimental Details of Main Results

Table 3: **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[3] 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[4], and PAPO and VPPO using Easy-R1[5], 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,

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

Table 4: 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

the 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$\times$ 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 4.

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 5.

Table 5: 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[6]. Multiple-choice questions are evaluated via exact-match accuracy, while free-form questions are verified using math-verify[7]. The prompt template used during 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 3.

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.

$$a_t^{\mathrm{vis}} = \sum_{j \in \mathcal{V}} \left[\mathrm{softmax}\left(\frac{q_t K_{\mathrm{all}}^\top}{\sqrt{d}} \right) \right]_j ,$$

where $\mathcal{V}$ denotes the set of visual-token positions and K_{all} contains all accessible visual and textual keys.

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.

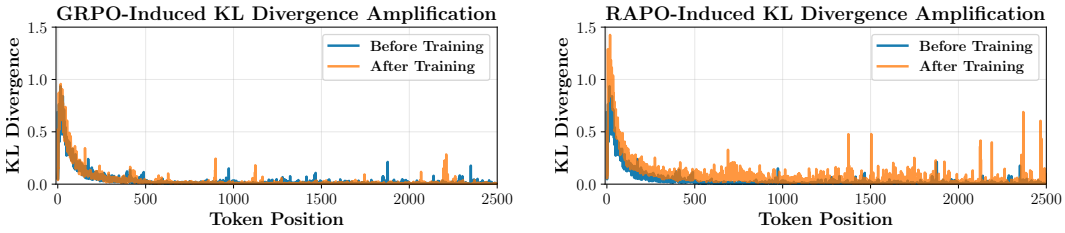


Figure 8: **KL Distributions along CoT.** (a) KL distribution after GRPO training; (b) KL distribution after RAPO training.

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

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.

E.1 Outcome-Level Patch Masking

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 $\Delta\mathrm{Acc}$.

As shown in Table 6, 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.

Table 6: **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

E.2 Visual-Specific Anchor Interventions

We further examine whether entropy-selected reflection anchors correspond to visually consequential reasoning positions.Using the frozen native policy, we conduct a paired intervention study on 100 matched samples from each of MathVerse's Text Dominant and Vision Only subsets. For each selected anchor, we replay the same prefix, suppress visual access at the anchor position, restore normal access afterward, and generate the remaining response. We compare the resulting accuracy with a position-matched control intervention. The additional anchor-specific effect is defined as $G = \Delta_{\text{anchor}} - \Delta_{\text{control}}$, where each Δ denotes the accuracy degradation relative to clean decoding.

Table 7 shows that the additional effect of anchor intervention is substantially stronger on Vision Only questions than on Text Dominant questions. This provides functional evidence that entropy-selected anchors identify branching points with visual-specific influence.

Table 7: **Anchor-specific intervention effects on MathVerse.** Accuracy and intervention effects are reported in percentage points.

Subset	Clean	Anchor	Control	Δ_a	Δ_c	G
Text Dominant	34.0	31.5	31.0	2.5	3.0	-0.5
Vision Only	24.0	18.6	20.8	5.4	3.2	2.2

E.3 Robustness Across Visual Counterfactuals

We evaluate whether the visual-dependence patterns identified by chain masking also appear under alternative counterfactual interventions in image space. Using frozen Base, GRPO, and RAPO checkpoints, we conduct teacher-forced evaluations on the same 500-example ViRL validation subset. We consider four visual interventions: chain masking, full-image masking, Gaussian image noise, and question-guided region removal.

Table 8 shows that all four interventions reproduce the same qualitative ordering: correct trajectories exhibit stronger visual contrast than incorrect trajectories, and RAPO retains stronger visual-dependence signals than GRPO. These consistent patterns across attention-level and pixel-space interventions demonstrate the robustness of the observed visual-retention effect.

Table 8: **Visual contrast under alternative counterfactual interventions.**

Intervention	Correct / Incorrect	RAPO / GRPO
Chain mask	0.7612 / 0.6329	1.0250 / 0.7648
Full-image mask	0.1000 / 0.0770	0.1940 / 0.1047
Gaussian noise ($\sigma = 0.5$)	0.0292 / 0.0258	0.0589 / 0.0317
Question-guided region removal	0.0219 / 0.0147	0.0476 / 0.0228

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 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.

Table 9: **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%

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 10, 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 10: **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 11.

Table 11 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 $c_t > \mu+\sigma$

Table 11: 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**

(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 12: 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

Table 13: 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**

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.

G.5 Adaptive Window Strategies

We compare the fixed-window propagation objective with two adaptive window selection strategies (Table 14). **Entropy-gated window.** For each anchor, the window extends until the first subsequent token whose entropy falls below the trajectory median. **Inter-anchor window.** Each window extends from the current anchor up to, but excluding, the next selected anchor. Both adaptive variants

Table 14: Comparison of fixed and adaptive propagation windows on Qwen3-VL-2B-Instruct.

Window	Mean Length	Reas.	Gen.	Avg.
Fixed ($w = 3$)	3.00	40.12	48.78	**43.01**
Entropy-gated	4.85	38.09	47.79	41.32
Inter-anchor	4.99	38.67	48.76	42.04

normalize the KL objective by their resulting window lengths. The fixed three-token window achieves the best overall accuracy, particularly on reasoning-intensive benchmarks. This result supports the use of a compact propagation horizon for retaining locally relevant visual influence.

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:

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

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 15. 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 15: **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 16.

Table 16: 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 [34]	Code	Apache-2.0
Math-Verify Github	Code	Apache-2.0
ViRL39K [26]	Dataset	MIT
MathVision [35]	Benchmark	MIT
MathVerse [36]	Benchmark	MIT
EMMA [37]	Benchmark	No explicit standard dataset license found; research evaluation only
LogicVista [38]	Benchmark	Apache-2.0
MMMU-Pro [39]	Benchmark	Apache-2.0
RealWorldQA Huggingface	Benchmark	CC BY-ND 4.0
Qwen3-VL-2B-Instruct [7]	Model	Apache-2.0
Qwen3-VL-8B-Instruct [7]	Model	Apache-2.0
Qwen3-VL-2B-Thinking [7]	Model	Apache-2.0
Qwen2.5-VL-7B-Instruct [32]	Model	Apache-2.0
Qwen2-VL-7B-Instruct [31]	Model	Apache-2.0
Gemma3-4B-IT [40]	Model	Gemma Terms of Use
MM-Eureka [9]	Model	Apache-2.0
VL-Rethinker [26]	Model	Apache-2.0
R1-ShareVL [33]	Model	Apache-2.0
TVC [11]	Model	Apache-2.0
MINT-CoT [15]	Model	No explicit standard license found; research baseline only

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.

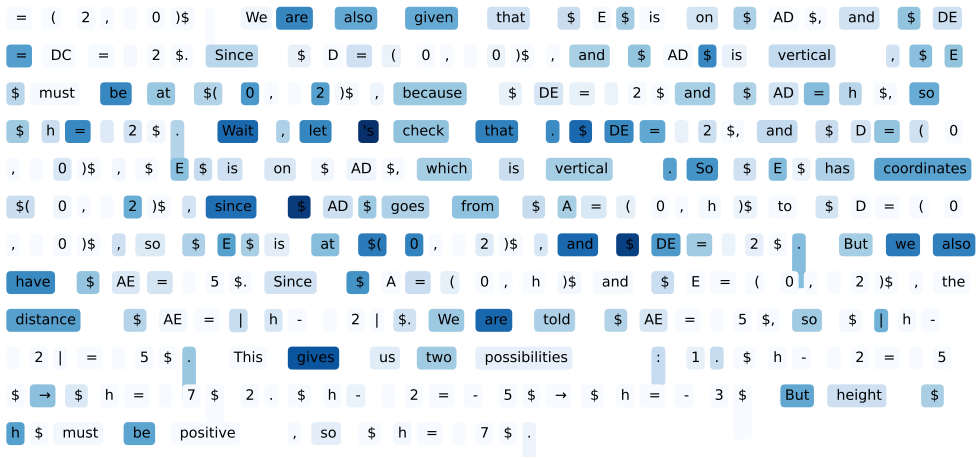
J.2 Broader Impacts

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.

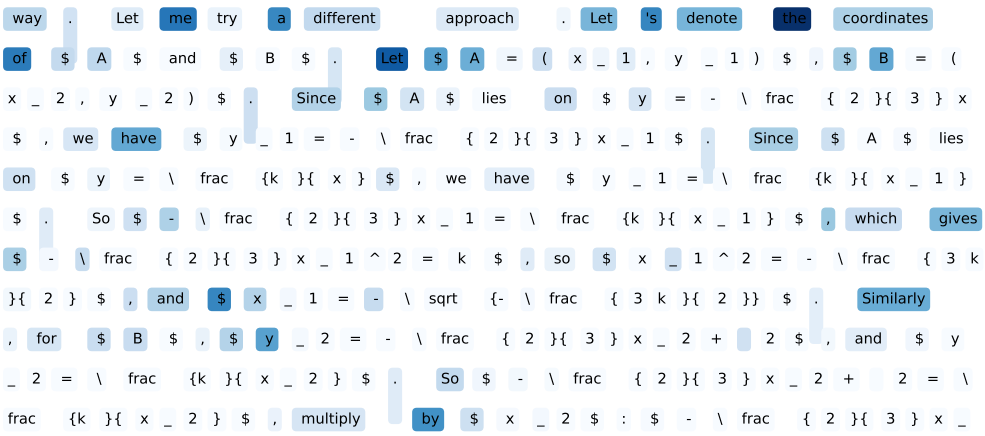
K Case Study

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.



(a) **Case 1.**



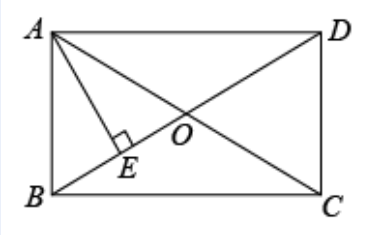
(b) **Case 2.**

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

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{}`.



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



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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide our theory and algorithm design in Section 3 and Section 4 and provide our experimental results in Section 5 to justify the effectiveness of our algorithm and the correctness of our theoretical insights

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section at Appendix J.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide our theory in Section 3 and Section 4 and the proof in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental settings in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets and benchmarks used in this work are open-source and publicly available, with links provided in footnotes and in Table 5. The submitted supplementary material also includes the core RAPO implementation code, including training code and evaluation code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide the training and evaluation details at Section 5, Appendix C and Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report evaluation-time sampling variability over 8 sampled responses per input in Appendix C and Table 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the computing time in detail in Appendix F and provide the hardware requirements at Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts of this paper in Section 6 and Appendix J.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release a new dataset, pre-trained model, image generator, or fine-tuned model checkpoint with high misuse risk. The released materials consist of the core RAPO implementation, including training and evaluation code, while all datasets, benchmarks, and base models used in the experiments are existing public assets governed by their original licenses and usage terms. Therefore, no additional safeguards for releasing high-risk data or models are applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the original creators of all code, datasets, and models used in this work, and explicitly list their source, version, URL, and license information in Appendix I.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve any crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: This paper does not involve any crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: This work explicitly uses existing LVLMs as the policy models being optimized and evaluated. Their usage, training procedure, and evaluation settings are described in Sections 4 and 5 and Appendix C. No additional unreported LLM usage is part of the core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.