Not All Tokens Should Be Treated Equally: Context Credits Reassignment

Shujian Gao[1,2,*] **Yuan Wang**[4,*] **Jiamei Yan**[3] **Penghao Zhou**[3,‡]
Qinglei Wang[3,‡] **Jiangtao Yan**[5] **Zuxuan Wu**[1,2,†] **Yu-Gang Jiang**[1,†]

[1]Institute of Trustworthy Embodied AI, Fudan University
[2]Shanghai Innovation Institute [3]ByteDance
[4]Zhejiang University [5]Wuhan University

Project page: `https://daryl-gsj.github.io/HiCCR/`

Abstract

Multimodal Reinforcement learning with verifiable rewards (RLVR) algorithms assign the same advantage to every token in a generated sequence, providing no differential signal for tokens with varying degrees of dependence on visual evidence. A simple experiment confirms this issue: **scaling token advantages by uninformative random noise already outperforms uniform credit**, suggesting that uniform credit is empirically suboptimal. Beyond that, we observe RLVR tuning under uniform credit weakens visual grounding along three complementary axes: attention, gradient attribution, and functional dependence on image embeddings, **collectively indicating insufficient utilization of visual evidence**. To understand this degradation, we provide a theoretical analysis showing that, under language-prior dominance and regularity conditions, uniform credit can amplify a pretrained text-favored gradient imbalance, contributing to weaker visual attention. In response, we propose Hierarchical Context Credit Reassignment (**HiCCR**), which reassigns credit at two levels: a token-level weight amplifies the advantage for visually grounded tokens, and a trajectory-level weight upweights rollouts with stronger visual engagement. The entire mechanism **adds less than 3% wall-clock overhead** and applies to mainstream RLVR algorithms without auxiliary models or additional data. On ten benchmarks spanning mathematical and general multimodal reasoning, HiCCR consistently improves upon its corresponding base algorithm and achieves state-of-the-art results among open-source models of comparable scale, while alleviating the weakening of visual grounding observed under RLVR training.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) is widely recognized as an effective paradigm for improving reasoning in Multimodal Large Language Models (MLLMs) [6, 24, 23, 18]. Existing methods rely on a uniform token credit assumption where a sequence reward is distributed equally across all generated tokens [49, 69]. However, multimodal reasoning inherently builds on the heterogeneous integration of visual evidence and textual priors [39, 28]. Forcing all tokens to share a single scalar advantage $\hat{A}$ directly contradicts the intuition that token contributions differ significantly in multimodal tasks. This simplification fails to align with the mechanism of multimodal reasoning,

[*]Equal contribution. [‡]Project lead. [†]Corresponding authors.

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

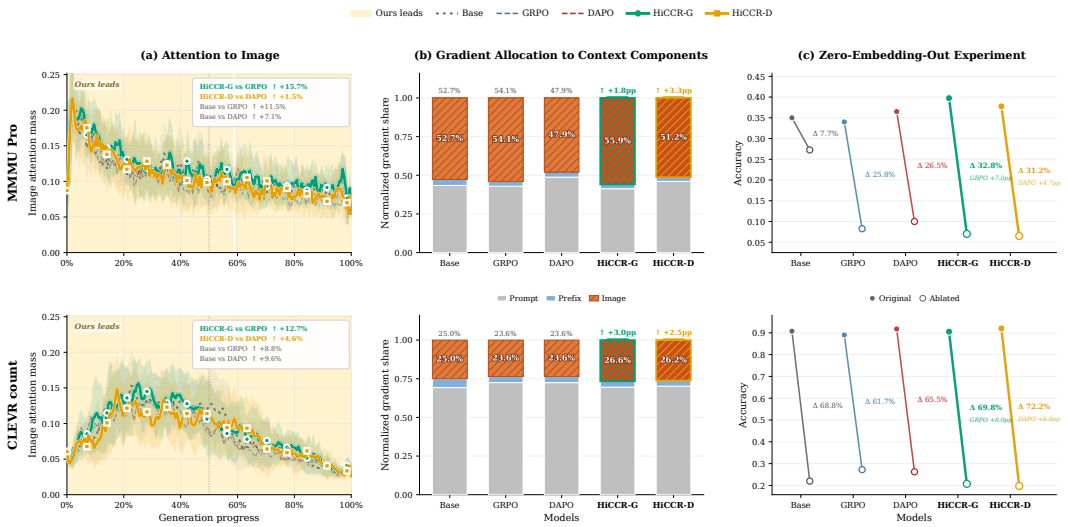


Figure 1: **RLVR fine-tuning weakens visual grounding in MLLMs, and HiCCR mitigates it.**

as many tokens must extract information from the visual context instead of relying on language priors [36, 21]. *Thus, it naturally raises the question: should all tokens in a trajectory really receive the same credit in multimodal RLVR?*

A simple experiment illustrates that uniform credit is overly coarse and empirically suboptimal. When we scale token advantages by **uninformative random noise** that carries no information about token quality, the resulting model already outperforms the base RLVR algorithm (Table 2). That is, *even a marginal perturbation of uniformity yields gains over the default assignment.*

We begin by answering this question empirically. Figure 1 tracks three independent signatures of visual grounding on the same MLLM backbone [2] before and after RLVR fine-tuning. Panel (a) tracks attention to image tokens across generation: the base model sustains higher image attention than GRPO and DAPO throughout, and the gap widens as generation proceeds. Panel (b) quantifies the gradient-level consequence via Input×Gradient attribution: RLVR reduces rather than increases the gradient share routed to image tokens. Panel (c) probes functional dependence via a zero-embedding-out ablation: zeroing out image embeddings hurts RLVR baselines less than the base model, suggesting weaker dependence on visual input. All three axes converge on the same finding: *RLVR can improve answer accuracy while weakening the visual grounding that should underpin multimodal reasoning (full results in Appendix D).*

Why does uniform credit produce this effect? We provide a theoretical analysis offering a structural explanation. Under a mild language-prior dominance condition (Appendix B), we establish two results that answer this question. **Theorem B.3** shows that uniform positive-advantage assignment preserves and aggregates an existing text-favored local gradient imbalance, rather than correcting it: the expected policy-gradient magnitude through textual context exceeds that through visual context. **Theorem B.4** shows that this per-step bias accumulates into a progressive drift of attention away from image tokens over generation, matching the trend in Figure 1(a). Because language-prior dominance is a property of the pretrained model [21], the bias is already present before fine-tuning, as reflected by the gradient gap in Figure 1(b); *RLVR amplifies rather than creates it by reinforcing tokens without distinguishing their degree of visual grounding.*

Motivated by above analysis, we propose Hierarchical Context Credit Reassignment (**HiCCR**), a simple method that reallocates credit at two complementary levels. At the **token level**, HiCCR increases the effective learning signal for tokens that appear more visually engaged; at the **trajectory level**, it upweights rollouts whose overall generation exhibits stronger visual engagement. To instantiate this idea efficiently, we use the concentration of a token's visual attention as a lightweight forward-pass proxy, avoiding auxiliary models, additional supervision, or extra backward passes. HiCCR is designed as a drop-in modification to existing RLVR algorithms, and adds less than 3% wall-clock overhead in our implementation. Compared with task-agnostic heuristics [58], applying our formula to the text context, and raw attention energy (Table 2), concentration-based proxy

improves both backbones and base algorithms, identifying the operative signal as concentration within the visual subset, not non-uniformity, not arbitrary context, not attention magnitude.

Together, our strategy closes the loop opened by our diagnosis: as shown in Figure 1, HiCCR mitigates attention drift, increases the gradient share routed to visual tokens, and yields stronger counterfactual evidence of visual dependence (Appendix. D). Our findings suggest effective multimodal RLVR requires revisiting token-level credit assignment, and HiCCR provides a direct instantiation by reallocating outcome rewards toward visually engaged tokens and trajectories. Our contributions are:

- **Finding.** We identify a visual grounding degradation in RLVR fine-tuned MLLMs on three independent axes (attention tracking, gradient attribution, and counterfactual visual ablation), already present in the pretrained base model and further amplified by RLVR (Appendix D).
- **Justification.** We attribute this finding to a structural interaction between uniform credit assignment and pretrained language-prior dominance: uniform advantage induces a systematic gradient bias favoring text over vision (Theorem B.3), which compounds into visual-attention drift (Theorem B.4), explaining both the base-model gap and its amplification after RLVR training.
- **HiCCR.** We propose a hierarchical credit-reassignment method that mitigates the imbalance at both token and trajectory levels, using visual attention concentration as a lightweight proxy for visual dependence.
- **Strong results.** Applied to GRPO and DAPO on Qwen3-VL across ten multimodal reasoning benchmarks, HiCCR consistently improves its base algorithm and matches or exceeds open-source models several times its scale. Additional experiments on InternVL3 [77] and the smaller Geo3K [38] split further verify robustness across a different architecture and a lower-data regime.

2 Method

We first formalize the token-level credit assignment problem in RLVR and then present **HiCCR**, a hierarchical framework that reassigns credit at both the token and trajectory levels to better align policy updates with visual dependence in multimodal reasoning.

2.1 Preliminaries: Token-Level Credit Assignment in RLVR

Consider an MLLM parameterized by θ that generates a response $\mathbf{y} = (y_1, \ldots, y_T)$ given a multimodal prompt $\mathbf{x}$ consisting of textual tokens $\mathbf{x}^{\text{text}}$ and visual tokens $\mathbf{x}^{\text{vis}}$. In RLVR, each complete trajectory $(\mathbf{x}, \mathbf{y})$ receives a scalar reward $R(\mathbf{x}, \mathbf{y}) \in \{0, 1\}$ from a deterministic verifier.

Methods such as GRPO [49] and DAPO [69] optimize the policy using a trajectory-level advantage computed over a group of sampled responses $\{\mathbf{y}^{(i)}\}_{i=1}^{G}$:

$$\hat{A}^{(i)} = \frac{R(\mathbf{x}, \mathbf{y}^{(i)}) - \mu_R}{\sigma_R + \epsilon}, \tag{1}$$

where μ_R and σ_R denote the mean and standard deviation of group rewards. The clipped objective is

$$\mathcal{L}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \min\Big(r_t(\theta)\, \hat{A},\ \ \text{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)\, \hat{A} \Big), \tag{2}$$

where $r_t(\theta) = \pi_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) \,/\, \pi_{\theta_{\text{old}}}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})$ is the per-token importance ratio.

The uniform credit assignment problem. Equations (1)–(2) imply that every token y_t in a trajectory receives the same advantage signal $\hat{A}^{(i)}$. This uniform assignment treats all tokens as equally responsible for the trajectory outcome, even though, in multimodal reasoning, tokens may differ substantially in their reliance on visual evidence.

To make this explicit, the policy gradient for a single token takes the form

$$\nabla_\theta \mathcal{L}_t \propto \hat{A}\, \nabla_\theta \log \pi_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t}). \tag{3}$$

Under standard RLVR, the same scalar advantage modulates every token update regardless of the degree of visual dependence. This mismatch motivates the need for a credit assignment mechanism that better reflects visual dependence in multimodal reasoning.

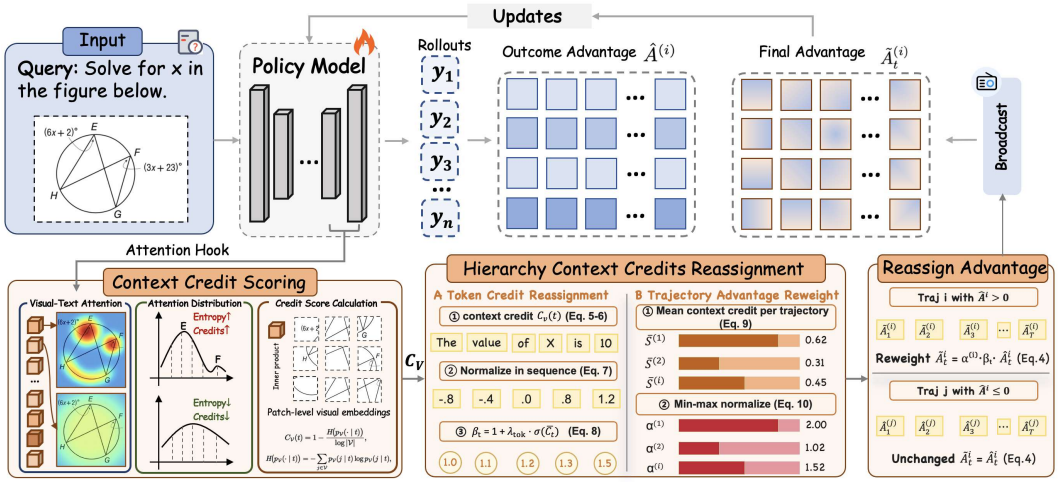


Figure 2: **Overview of HiCCR.** An attention hook on the policy model produces a per-token context credit $C_V(t)$ from intra-visual attention entropy. Token- and trajectory-level weights $\beta_t, \alpha^{(i)}$ are derived from C_V and reassign the outcome advantage as $\tilde{A}_t^{(i)} = \alpha^{(i)} \beta_t \hat{A}^{(i)}$. The reassignment is applied only to positive-advantage trajectories.

2.2 HiCCR: Hierarchical Context Credit Reassignment

To address this token-level credit mismatch, we propose HiCCR, which augments the standard RLVR pipeline with two complementary mechanisms operating at different granularities:

1. **Token Credit Reassignment** (§2.2.2), a per-token weight β_t that amplifies the policy gradient for tokens with stronger estimated dependence on visual context.
2. **Trajectory Advantage Reweighting** (§2.2.3), a per-trajectory scaling factor α that preferentially reinforces trajectories exhibiting stronger overall visual engagement.

Both mechanisms derive their signals from a lightweight *context credit score* (§2.2.1) computed from the internal attention patterns of the model, without requiring an additional reward model, value network, or external supervision. The reassigned advantage is

$$\tilde{A}_t^{(i)} = \begin{cases} \alpha^{(i)} \cdot \beta_t^{(i)} \cdot \hat{A}^{(i)}, & \text{if } \hat{A}^{(i)} > 0, \\ \hat{A}^{(i)}, & \text{otherwise,} \end{cases} \tag{4}$$

where reassignment is applied only to positive-advantage trajectories. For correct trajectories, the goal is to differentially reinforce tokens and trajectories that exhibit stronger reliance on visual evidence. For incorrect trajectories, the negative advantage already provides a uniform discouragement signal, and further modulation is unnecessary. A discussion of asymmetric shaping is provided in Table 12.

2.2.1 Context Credit Score

The foundation of both mechanisms is a per-token *context credit score* that serves as a lightweight proxy for the degree to which the generation of token y_t depends on visual context. We define this score as visual information concentration, namely the complement of the normalized entropy of the attention distribution restricted to visual positions.

Setup. For a target layer $\ell \in \mathcal{L}_{\text{target}}$, let $\mathbf{q}_t^{(h)}, \mathbf{k}_j^{(h)} \in \mathbb{R}^d$ denote the query and key vectors at positions t and j for head h, and let $\mathcal{V} \subset \{1, \ldots, |\mathbf{x}|\}$ be the set of visual token positions. We compute the intra-visual attention distribution by restricting the softmax to visual positions only:

$$p_{\mathcal{V}}(j \mid t, h) = \frac{\exp\left(e_{t,j}^{(h)}/\tau\right)}{\sum_{j' \in \mathcal{V}} \exp\left(e_{t,j'}^{(h)}/\tau\right)}, \qquad e_{t,j}^{(h)} = \frac{\mathbf{q}_t^{(h)\top} \mathbf{k}_j^{(h)}}{\sqrt{d}}, \tag{5}$$

where $\tau > 0$ is a temperature parameter. The resulting context credit score is

$$C_{\mathcal{V}}(t) = 1 - \frac{H\big(p_{\mathcal{V}}(\cdot \mid t)\big)}{\log |\mathcal{V}|}, \qquad H\big(p_{\mathcal{V}}(\cdot \mid t)\big) = -\sum_{j \in \mathcal{V}} p_{\mathcal{V}}(j \mid t) \log p_{\mathcal{V}}(j \mid t), \tag{6}$$

averaged over heads h and layers ℓ, with indices omitted for clarity. Dividing by $\log |\mathcal{V}|$, the maximum entropy of a uniform distribution over $|\mathcal{V}|$ positions, normalizes $C_{\mathcal{V}}(t)$ to the range $[0, 1]$ regardless of the number of visual tokens.

This definition admits a clear interpretation. A large value of $C_{\mathcal{V}}(t)$ indicates that token t concentrates its visual attention on a small subset of image positions, whereas a small value indicates a more diffuse attention pattern [73]. Because the softmax in Eq. (5) is computed only over visual positions, $C_{\mathcal{V}}(t)$ is invariant to global additive shifts in the energy scale across layers and training stages [33].

2.2.2 Token Credit Reassignment (TCR)

Given the context credit scores $\{C_{\mathcal{V}}(t)\}_{t=1}^{T}$ for a trajectory, we normalize them to zero mean and unit variance within each sequence:

$$\bar{c}_t = \frac{C_{\mathcal{V}}(t) - \mu_c}{\sigma_c + \epsilon}, \tag{7}$$

where μ_c and σ_c are the per-sequence mean and standard deviation. The resulting token-level weight is

$$\beta_t = 1 + \lambda_{\text{tok}} \cdot \sigma(\bar{c}_t), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\lambda_{\text{tok}} > 0$ controls the maximum amplification. Since $\sigma(\bar{c}_t) \in (0, 1)$, the weight satisfies $\beta_t \in (1,\ 1{+}\lambda_{\text{tok}})$. Tokens with above-average visual concentration therefore receive stronger reinforcement, while those with below-average concentration remain close to unity. No token is suppressed. The mechanism does not reverse the sign of any per-token update; it only modulates the relative magnitude of positive reinforcement.

2.2.3 Trajectory Advantage Reweighting (TAR)

Token credit reassignment operates within a single trajectory, whereas trajectory reweighting provides a complementary signal across trajectories. Among correct trajectories, those with stronger visual engagement are more likely to reflect solutions that make substantive use of visual evidence and should therefore receive proportionally larger reinforcement.

For each trajectory i, we compute the mean context credit score:

$$\bar{S}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} C_{\mathcal{V}}^{(i)}(t). \tag{9}$$

The trajectory-level weight is obtained by min-max normalization across the current batch:

$$\alpha^{(i)} = 1 + \frac{\bar{S}^{(i)} - \min_k \bar{S}^{(k)}}{\max_k \bar{S}^{(k)} - \min_k \bar{S}^{(k)} + \epsilon}, \tag{10}$$

Under this normalization, the reweighting is strictly non-diminishing, which helps preserve the stability of the underlying RL objective. The procedure and design principles are summarized in F.3.

3 Experiments

3.1 Experimental Setup and Details

Data, baselines, benchmarks, and evaluation. We train on **ViRL39K** [56], a set of 39K multimodal questions with verifiable answers spanning geometry, algebra, and word problems. We compare against **nine RLVR algorithms reproduced under identical conditions** (GRPO [49], DAPO [69], SAPO [14], GSPO [76], CISPO [42], Reinforce++ [22], PAPO [63], DVRP [17], VPPO [23]), a broad set of proprietary and open-weight multimodal models, and alternative token-level credit assignment strategies. Evaluation spans ten benchmarks across three categories: Mathematical Reasoning, including Geometry3K [38], MathVerse [74], MathVista [37], WeMath [48], and MathVerse-Vision [74]; General and Logical Reasoning, including MMK12 [40], CLEVR-Count [27], LogicVista [65], and MMMU [71]; and Hallucination Robustness, including HallusionBench [19]. We report accuracy (%) on each benchmark, averaged over eight independent forward passes (AVG@8). The final training checkpoint is selected for evaluation to ensure fairness. For further details, please refer to Appendix F.

Table 1: Performance on Qwen3-VL-4B and 8B across multimodal reasoning benchmarks. Within each backbone we place the **Ours vs. matched base RL** pairs at the bottom. Best in **bold**.

Method	Mathematical Reasoning						General & Logical Reasoning					AVG@8
	Geo3K	MVerse	MVista	WeMath	MV-V	*Avg.*	MMK12	Clevr	Logic	MMMU	*Avg.*	
***Qwen3-VL-4B** Backbone*												
Base Model (Instruct)	45.42	43.39	65.80	53.39	40.90	*49.78*	39.44	90.00	42.73	20.52	*48.17*	49.07
Other RLVR baselines (context)												
SAPO	40.47	54.17	48.70	62.99	52.80	*51.83*	54.02	63.50	50.43	42.89	*52.71*	52.22
GSPO	59.07	70.69	**74.90**	74.89	67.13	*69.34*	56.86	91.00	53.47	41.04	*60.59*	65.45
CISPO	54.08	64.72	70.20	68.51	59.10	*63.32*	51.38	88.50	50.11	41.33	*57.83*	60.88
Reinforce++	55.24	71.19	69.00	74.66	66.13	*67.24*	67.37	83.00	54.81	42.54	*61.93*	64.88
PAPO	52.94	69.29	68.60	73.97	64.13	*65.80*	65.28	83.00	53.20	41.87	*60.84*	63.59
DVRP	59.43	73.18	70.80	75.42	69.13	*69.60*	61.35	90.00	54.17	43.26	*61.95*	66.20
VPPO	57.08	70.31	70.20	73.12	66.44	*67.39*	58.22	90.50	53.17	42.06	*60.99*	64.19
GRPO	53.41	67.16	68.70	68.45	63.15	*64.17*	55.97	87.50	52.35	40.87	*59.17*	61.95
HiCCR-G (Ours)	61.73	**77.34**	71.70	78.62	**73.85**	***72.65*** *(+8.48)*	**68.64**	**94.00**	57.05	**46.53**	***66.55*** *(+7.38)*	**69.94** (+7.99)
DAPO	60.57	74.40	72.30	**79.48**	70.26	*71.40*	60.56	92.00	54.59	44.05	*62.80*	67.58
HiCCR-D (Ours)	**63.23**	73.21	73.50	75.69	71.18	*71.36* *(−0.04)*	65.54	91.50	**60.40**	46.30	*65.93* *(+3.13)*	68.95 (+1.37)
***Qwen3-VL-8B** Backbone*												
Base Model (Instruct)	51.58	53.99	65.70	60.29	50.08	*56.33*	45.91	89.50	47.87	36.47	*54.94*	55.71
Other RLVR baselines (context)												
SAPO	53.35	60.21	57.20	63.47	56.62	*58.17*	50.36	72.05	52.11	41.84	*54.09*	56.35
GSPO	65.22	76.10	**76.70**	78.97	72.40	*73.88*	63.61	90.50	57.49	47.86	*64.86*	69.87
CISPO	57.57	66.15	67.70	69.89	62.23	*64.71*	53.76	91.50	52.35	45.26	*60.72*	62.93
Reinforce++	65.89	76.59	72.80	79.20	77.22	*74.94*	70.23	90.50	**63.98**	**50.08**	***69.95***	72.32
PAPO	63.23	76.03	71.40	75.84	75.06	*72.31*	70.75	89.00	60.98	50.01	*67.68*	70.25
DVRP	64.42	76.15	72.30	78.06	74.41	*73.06*	69.53	89.00	59.45	48.54	*66.63*	70.20
VPPO	60.41	69.28	72.00	73.66	67.52	*68.17*	60.73	90.50	55.38	47.21	*63.46*	65.58
GRPO	56.74	65.23	68.40	68.91	61.93	*64.24*	54.81	90.50	52.57	43.35	*60.31*	62.49
HiCCR-G[1] (Ours)	58.40	72.29	73.50	76.49	68.27	*69.79* *(+5.55)*	62.11	90.00	57.49	46.18	*63.94* *(+3.63)*	67.19 (+4.70)
DAPO	64.23	73.39	71.10	77.13	68.88	*70.95*	63.99	89.00	57.49	49.36	*64.96*	68.29
HiCCR-D (Ours)	**68.39**	**78.94**	75.70	**81.03**	**75.99**	***76.01*** *(+5.06)*	**70.58**	**92.00**	60.63	51.56	*68.69* *(+3.73)*	**72.76** (+4.47)

(+Δ) / (−Δ) are gains against the matched base RL row above. HiCCR wins all 4 AVG@8 aggregates. All baseline experiments are produced in identical environment with same conditions for fair comparison.

3.2 Main Experiments

Comparison with RLVR algorithms and Computation Efficiency. HiCCR consistently attains the best or near-best accuracy across both backbones while remaining highly efficient, under identical-condition comparisons with nine RLVR baselines. Table 1 shows HiCCR-G reaches **69.94%** AVG@8 on 4B and HiCCR-D **72.76%** on 8B, surpassing the strongest baselines on both scales and improving their respective base algorithms, with the largest gains on inherently vision-dependent benchmarks such as MathVerse, which is consistent with our hypothesis that uniform credit assignment disproportionately harms visually grounded reasoning. These gains come at negligible cost: HiCCR adds only marginal wall-clock and peak-memory overhead, far below perturbation-based alternatives such as PAPO [63] (Computation Cost Details in Appendix A.1).

Comparison with multimodal models. HiCCR-D matches or exceeds several larger open-weight and proprietary baselines on the evaluated benchmark aggregate, demonstrating strong performance beyond the controlled RLVR setting. As reported in Table 9(Appendix A.5), HiCCR-D achieves **72.76%**, surpassing the 72B Qwen2.5-VL-Instruct [3], underscoring HiCCR's effectiveness in unlocking the multimodal reasoning potential of smaller models.

Analysis of token-level credit computation strategies. The gains of HiCCR come from the concentration structure of visual attention, not from non-uniform weighting in general. Table 2 compares our score against five alternatives under GRPO and DAPO: Uniform, Random ($w_t \sim \mathcal{N}(1, \sigma^2)$, σ=0.2), Time-decay ($w_t = \gamma^{T-t}$), Entropy-mask (top-20% [58]), HiCCR-T (same formula over the text context $\mathcal{X}_{\text{text}} \cup y_{<t}$), and Raw Energy (logsumexp of $e_{t,j}^{(h)}$ over $\mathcal{V}$, no intra-visual normalization). The results form a clear ladder. Random beats Uniform on every configuration, so uniform credit is not even locally optimal. Yet non-uniformity alone is insufficient: Time-decay and Entropy-mask can hurt DAPO, and HiCCR-T often falls below Uniform although it uses the same formula as ours. Concentration helps only when computed over the visual subset. Within that subset, concentration outperforms magnitude: Raw Energy beats every heuristic but trails our full score across all four configurations. These comparisons suggest that, among the tested alternatives, the

Table 2: Token-level credit assignment across various reasoning benchmarks under GRPO and DAPO.

Strategy	Mathematical Reasoning					General & Logical Reasoning				AVG@8	Δ
	Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	Clevr	Logic	MMMU		
***Qwen3-VL-4B** Backbone*											
Base Model (Instruct)	45.42	43.39	65.80	53.39	40.90	39.44	90.00	42.73	20.52	49.07	–
GRPO-based											
Uniform Allocation (Baseline)	53.41	67.16	68.70	68.45	63.15	55.97	87.50	52.35	40.87	61.95	–
Random Allocation	56.08	70.33	66.80	68.80	64.44	58.19	85.00	53.39	40.60	62.63	(+0.68)
Time-decay (γ-decay)	56.40	73.19	70.80	69.43	62.86	55.54	87.00	52.55	39.55	63.04	(+1.09)
Entropy-mask (Top-20%)	55.91	70.62	69.30	66.64	65.22	59.27	89.50	53.64	41.36	63.50	(+1.55)
Text-Context-based (HiCCR-T)	54.07	66.51	69.25	68.12	60.78	54.37	86.50	53.70	40.16	61.50	(−0.45)
Raw Attention Energy *(w/o Eq. 6)*	58.42	73.67	70.20	75.19	69.58	64.83	91.50	53.94	42.17	66.61	(+4.66)
Context-based (HiCCR-G, Ours)	**61.73**	**77.34**	**71.70**	**78.62**	**73.85**	**68.64**	**94.00**	**57.05**	**46.53**	**69.94**	(+7.99)
DAPO-based											
Uniform Allocation (Baseline)	60.57	**74.40**	72.30	**79.48**	70.26	60.56	**92.00**	54.59	44.05	67.58	–
Random Allocation	63.06	74.27	72.60	75.80	70.48	**65.76**	**92.00**	58.84	44.45	68.58	(+1.00)
Time-decay (γ-decay)	57.74	70.09	71.30	74.37	68.35	61.01	78.50	54.36	40.23	64.00	(−3.58)
Entropy-mask (Top-20%)	61.06	72.84	69.00	77.39	68.88	61.67	90.50	58.97	39.71	66.67	(−0.91)
Text-Context-based (HiCCR-T)	61.90	73.86	72.70	77.05	69.36	59.29	91.00	56.27	43.29	67.19	(−0.39)
Raw Attention Energy *(w/o Eq. 6)*	62.97	73.01	73.20	75.34	70.96	65.17	91.00	60.14	46.21	68.67	(+1.09)
Context-based (HiCCR-D, Ours)	**63.23**	73.21	**73.50**	75.69	**71.18**	65.54	91.50	**60.40**	**46.30**	**68.95**	**(+1.37)**
***Qwen3-VL-8B** Backbone*											
Base Model (Instruct)	51.58	53.99	65.70	60.29	50.08	45.91	89.50	47.87	36.47	55.71	–
GRPO-based											
Uniform Allocation (Baseline)	56.74	65.23	68.40	68.91	61.93	54.81	**90.50**	52.57	43.35	62.49	–
Random Allocation	57.32	66.17	69.30	67.75	62.94	55.86	90.00	52.23	42.87	62.72	(+0.23)
Time-decay (γ-decay)	57.06	65.19	71.20	69.31	67.24	57.09	88.50	48.39	43.17	63.02	(+0.53)
Entropy-mask (Top-20%)	55.98	67.46	69.65	71.72	63.29	58.69	90.00	53.61	43.25	63.74	(+1.25)
Text-Context-based (HiCCR-T)	53.24	70.13	70.60	73.85	66.82	58.96	89.00	54.59	44.39	64.62	(+2.13)
Raw Attention Energy *(w/o Eq. 6)*	57.64	70.18	72.00	73.96	65.82	60.43	89.50	55.27	44.71	65.50	(+3.01)
Context-based (HiCCR-G, Ours)	**58.40**	**72.29**	**73.50**	**76.49**	**68.27**	**62.11**	90.00	**57.49**	**46.18**	**67.19**	**(+4.70)**
DAPO-based											
Uniform Allocation (Baseline)	64.23	73.39	71.10	77.13	68.88	63.99	89.00	57.49	49.36	68.29	–
Random Allocation	66.39	76.53	74.60	80.55	73.84	68.20	**93.00**	60.19	49.27	71.40	(+3.11)
Time-decay (γ-decay)	62.90	71.92	72.70	76.97	73.99	62.34	88.50	57.72	49.54	68.51	(+0.22)
Entropy-mask (Top-20%)	65.72	77.03	74.40	79.48	73.35	66.92	91.00	56.60	47.98	70.28	(+1.99)
Text-Context-based (HiCCR-T)	62.93	74.50	73.90	77.28	70.71	66.21	89.50	58.96	48.61	69.18	(+0.89)
Raw Attention Energy *(w/o Eq. 6)*	67.64	78.13	75.20	80.36	75.17	69.87	91.50	59.94	50.89	72.08	(+3.79)
Context-based (HiCCR-D, Ours)	**68.39**	**78.94**	**75.70**	**81.03**	**75.99**	**70.58**	92.00	**60.63**	**51.56**	**72.76**	**(+4.47)**

Best per framework in **bold**; Δ is improvement over Uniform (red = gain, blue = drop). Raw Attention Energy is our ablation without Eq. 6, Context-based is the full HiCCR score. Text-Context-based (HiCCR-T) replaces the visual reference set $\mathcal{V}$ in Eq. 6 with the non-visual context $\mathcal{X}_{\text{text}} \cup y_{<t}$ (prompt text plus previously generated tokens), isolating the effect of restricting the reference set to $\mathcal{V}$.

most effective signal is concentration within the visual subset rather than generic non-uniformity, raw attention magnitude, or applying the same formula to textual context.

Generalization to Perception and Hallucination Benchmarks. HiCCR also improves performance on perception- and hallucination-oriented benchmarks. On MMStar [7] and HalluBench [19], HiCCR-G consistently outperforms all baselines across both backbones (Fig. 5, Appendix A.5), reaching 71.33%/72.20% on MMStar and 59.73%/62.67% on HalluBench. This pattern is consistent with HiCCR's design: by up-weighting visually engaged tokens during training, it encourages more faithful use of image evidence and reduces reliance on spurious textual priors.

Data size verification and different mllm architecture generalization. Restricting training to the small Geo3K split (Appendix A.2) and transplanting the full pipeline to a different MLLM family with InternVL3-8B [8] (Appendix A.3), HiCCR continues to outperform its base RL algorithms in both settings. This indicates that the gains are robust to data scarcity and transfer across distinct vision encoders, connectors, and language priors, supporting HiCCR as a data-efficient and architecture-agnostic credit-reassignment mechanism.

Table 3: Ablation of the two core components (**TCR** and **TAR**) in HiCCR on Qwen3-VL-4B and 8B.(+Δ) / (−Δ) are per-cell differences against the matched base RL row (grey).

Method	Components		Mathematical Reasoning					General & Logical Reasoning				AVG@8
	TCR	TAR	Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	Clevr	Logic	MMMU	
***Qwen3-VL-4B** Backbone*												
Base Model (Instruct)			45.42	43.39	65.80	53.39	40.90	39.44	90.00	42.73	20.52	49.07
DAPO	✗	✗	60.57	74.40	72.30	79.48	70.26	60.56	92.00	54.59	44.05	67.58
+ TCR	✓	✗	62.45 (+1.88)	75.12 (+0.72)	72.80 (+0.50)	79.56 (+0.08)	71.34 (+1.08)	63.89 (+3.33)	91.50 (−0.50)	57.23 (+2.64)	45.67 (+1.62)	68.84 **(+1.26)**
+ TAR	✗	✓	61.78 (+1.21)	74.65 (+0.25)	72.55 (+0.25)	78.23 (−1.25)	70.89 (+0.63)	62.34 (+1.78)	91.85 (−0.15)	56.12 (+1.53)	44.92 (+0.87)	68.15 **(+0.57)**
HiCCR-D (Ours)	✓	✓	63.23 (+2.66)	73.21 (−1.19)	**73.50** (+1.20)	75.69 (−3.79)	71.18 (+0.92)	65.54 (+4.98)	91.50 (−0.50)	**60.40** (+5.81)	46.30 (+2.25)	**68.95 (+1.37)**
***Qwen3-VL-8B** Backbone*												
Base Model (Instruct)			51.58	53.99	65.70	60.29	50.08	45.91	89.50	47.87	36.47	55.71
GRPO	✗	✗	56.74	65.23	68.40	68.91	61.93	54.81	90.50	52.57	43.35	62.49
+ TCR	✓	✗	58.76 (+2.02)	71.34 (+6.11)	71.50 (+3.10)	75.23 (+6.32)	67.45 (+5.52)	61.12 (+6.31)	90.50 (+0.00)	56.34 (+3.77)	46.78 (+3.43)	66.56 **(+4.07)**
+ TAR	✗	✓	56.23 (−0.51)	69.85 (+4.62)	70.50 (+2.10)	72.89 (+3.98)	65.12 (+3.19)	58.76 (+3.95)	89.75 (−0.75)	54.91 (+2.34)	44.56 (+1.21)	64.73 **(+2.24)**
HiCCR-G (Ours)	✓	✓	58.40 (+1.66)	72.29 (+7.06)	73.50 (+5.10)	76.49 (+7.58)	68.27 (+6.34)	62.11 (+7.30)	90.00 (−0.50)	57.49 (+4.92)	46.18 (+2.83)	**67.19 (+4.70)**

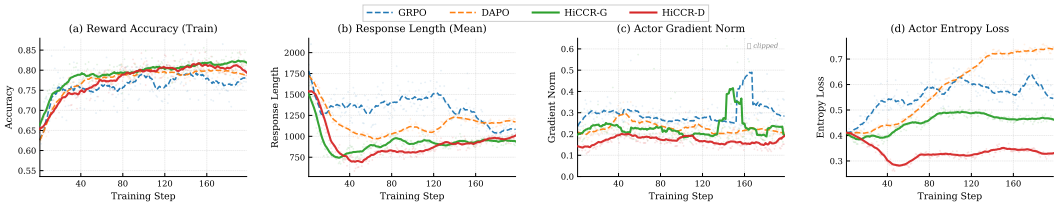


Figure 3: Training dynamics on Qwen3-VL-4B. **(a)** Reward accuracy. **(b)** Mean response length. **(c)** Actor gradient norm. **(d)** Policy entropy. Faint dots show per-step values; solid/dashed lines are smoothed averages. HiCCR converges to higher accuracy with shorter responses, stable gradients, and lower entropy, indicating efficient and confident visual reasoning.

3.3 Ablation Study and Extented Discussion

Component Contribution and Training Log. We incrementally add TCR and TAR to DAPO (4B) and GRPO (8B) to isolate their contributions. Table 3 shows the two modules are complementary: on 8B, TCR or TAR alone lifts GRPO's AVG@8 from 62.49% to 66.56% and 64.73%, while combining both in HiCCR-G yields the best overall score, confirming that token- and trajectory-level signals capture distinct aspects of reasoning. Consistently, Figure 3 shows HiCCR reaches higher reward accuracy with shorter responses, smaller gradient norms, and lower entropy than GRPO/DAPO, indicating more stable and confident optimization (Appendix A.4).

Visualization: How HiCCR redistributes credit. A fine-grained credit analysis (Fig. 4) shows that HiCCR reallocates optimization pressure in a distinctly vision-aware manner rather than uniformly across the response. Its token-level weights are smooth rather than binary, its trajectory-level weights favor visually engaged rollouts, and its highest-credit tokens are dominated by geometry and diagram semantics. Moreover, credit is concentrated earlier in generation and decays over token position, suggesting that visual evidence is used most heavily when establishing the reasoning scaffold. Consistently, semantically visual tokens receive significantly higher credit than non-visual ones, indicating that HiCCR assigns higher weights to tokens that are semantically associated with visual reasoning, without token-level supervision. Full analysis in Appendix A.6.

Ablation: Credit Distribution Strategy. Motivated by prior work that updates only the top-$k\%$ of token gradients [23, 58], we compare **hard masking** against our **soft weighting** under identical credit scores (Table 10, Appendix A.5). Hard masking consistently underperforms: Top-20% degrades both backbones, and Top-40% barely matches Uniform ($+0.58$), as discarding sub-threshold tokens forfeits gradient mass from partially informative positions. The continuous weighting of HiCCR preserves the full token spectrum while amplifying visually grounded steps, and further eliminates the need to tune the sensitive masking ratio k, making it both more effective and more practical to deploy.

Ablation: Target Layers for Credit Estimation. Later transformer layers yield progressively more informative credit signals. We ablate which layers supply the visual-attention proxy for HiCCR-G on Qwen3-VL-4B (Table 11, Appendix A.5). Accuracy improves monotonically from

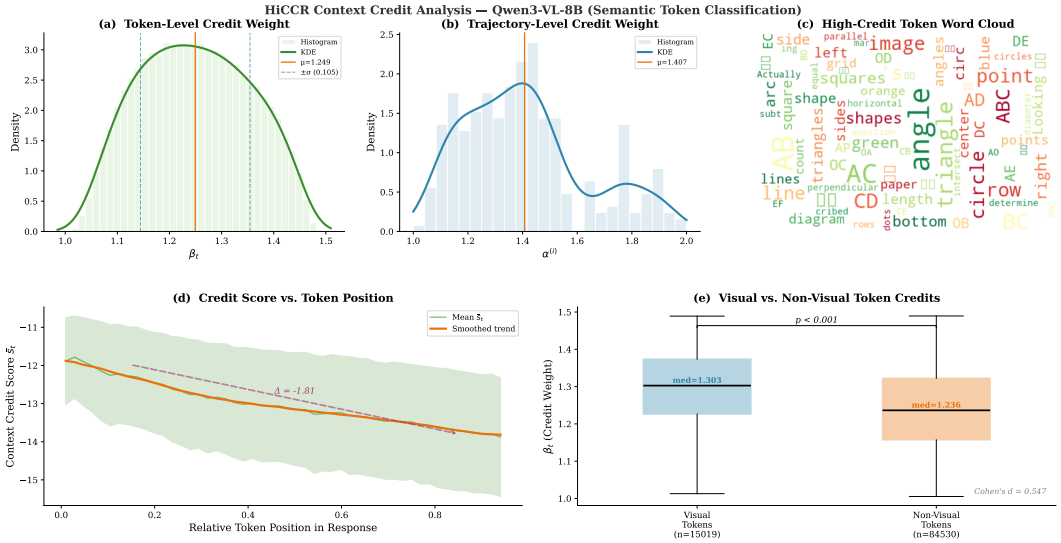


Figure 4: **HiCCR context credit analysis on Qwen3-VL-8B.** (a) Token-level credit weight β_t distribution. (b) Trajectory-level credit weight $\alpha^{(i)}$ distribution. (c) Word cloud of high-credit tokens, sized by average $\beta_t \times$ frequency. (d) Context credit score $\bar{s}_t$ vs. relative token position, showing monotonic decay ($\Delta{=}-1.81$). (e) Credit weight comparison between semantically visual and non-visual tokens ($p < 0.001$, Cohen's $d{=}0.547$).

the first to the middle to the last two layers, with the largest margins on visually demanding tasks (MMK12 +6.86). This reflects a representational hierarchy: early layers capture local visual-text alignment but lack task specificity, middle layers begin cross-modal fusion, and the last layers, where the model commits to its reasoning trajectory, best indicate whether attention to the image is causally consequential [70, 43]. Credit assignment should therefore be grounded in where the model acts on visual information, not merely where it attends to it.

Ablation: Asymmetric Advantage Shaping **Credit reassignment should be restricted to positive-advantage trajectories.** HiCCR modulates only correct rollouts ($\hat{A} > 0$), leaving negative-advantage trajectories unchanged (Eq. 4). We compare this **asymmetric** design against a **symmetric** variant that reshapes both signs under HiCCR-G at 4B and 8B scales (Table 12, Appendix A.5). Asymmetric shaping yields substantially larger gains, with the widest margins on geometry- and diagram-intensive tasks (Geo3K, WeMath, MV-V). The mechanism is intuitive: symmetric reshaping amplifies negative gradients for incorrect rollouts that happen to attend strongly to the image, thereby penalising visual attention indiscriminately and conflating visual engagement with correctness. Restricting reassignment to positive-advantage trajectories avoids this failure mode, ensuring that only correct reasoning paths receive visually guided credit amplification.

Sensitivity to Token-Level Scaling Magnitude λ_{tok} **Moderate scaling yields the best trade-off; both under- and over-amplification hurt.** The hyper-parameter λ_{tok} in Eq. 8 controls the maximum amplitude of the per-token credit multiplier $\beta_t = 1 + \lambda_{\text{tok}} \cdot \sigma(\bar{s}_t)$. We sweep $\lambda_{\text{tok}} \in \{0.1, 0.2, 0.5, 0.8, 1.5\}$ under HiCCR-D on Qwen3-VL-4B (Table 13, Appendix A.5). Small values (0.1–0.2) under-differentiate tokens and approach the uniform baseline (+0.25/+0.57); $\lambda_{\text{tok}} = 0.8$ achieves the best AVG (69.31, +1.73); $\lambda_{\text{tok}} = 1.5$ over-amplifies the credit gap and destabilises training (-1.34), excessively down-weighting tokens that are useful but not directly attending to the image. Our default $\lambda_{\text{tok}} = 0.5$ (+1.39) trades a marginal 0.34 in AVG for greater stability, and performance remains strong across the $[0.5, 0.8]$ range, indicating low sensitivity to exact tuning.

Ablation: Matched Random Controls **Visual credit alignment matters beyond generic non-uniform weighting.** To disentangle these effects, we introduce two controls. *Matched Shuffle* permutes HiCCR-derived weights, preserving their empirical distribution and scale while disrupting their correspondence with visual engagement. *Mean-Preserving HiCCR* retains visual ranking but normalizes the mean positive-advantage multiplier to one. Table 4 shows that HiCCR outperforms

Table 4: **Matched random controls.** AVG@8 (%) averaged over five training seeds (different from main experiments). 4B-D and 8B-D denote the Qwen3-VL-4B and 8B settings with DAPO, respectively.

Credit Rule	4B-D	8B-D
Uniform	67.51	68.21
Gaussian Random	68.53	71.31
Matched Shuffle	68.47	71.48
Mean-Preserving HiCCR	68.84	72.24
HiCCR (Full)	**69.51**	**72.65**

Table 5: **Grounding–accuracy correlations.** Spearman ρ between matched accuracy gains and grounding changes. Each setting contains ten benchmark-level pairs; Pooled combines all 40 pairs. Image-attention mass is summarized by its area under the generation curve (AUC). Zero-out drop measures the accuracy decrease after zeroing image embeddings.

Grounding Diagnostic	4B-G	4B-D	8B-G	8B-D	Pooled
Visual-gradient share	0.64	0.36	0.15	0.31	**0.42**
Image-attention mass (AUC)	0.58	0.39	0.20	0.44	**0.45**
Zero-out accuracy drop	0.71	0.47	0.18	0.51	**0.50**

Matched Shuffle, while Mean-Preserving HiCCR also improves over Uniform and Random. These results support the value of visually aligned credit: neither the weight distribution nor average positive-update amplification alone explains the gains.

Grounding–Accuracy Correlation **Grounding recovery is positively associated with accuracy gains.** We compute Spearman correlations between accuracy improvements and changes in three grounding diagnostics, each measured as HiCCR minus its matched GRPO/DAPO baseline. Across ten benchmarks and four settings, Table 5 shows positive pooled correlations for visual-gradient share (0.42), image-attention mass (0.45), and zero-out accuracy drop (0.50). The associations vary across settings and are weaker for 8B-G, indicating that grounding recovery does not map one-to-one to accuracy gains. These results are consistent with grounding contributing to the improvements, but do not establish causal mediation or exclude other mechanisms.

4 Conclusion

We studied the token-level credit mismatch in multimodal RLVR and showed that uniform credit assignment can bias learning toward textual rather than visual context. To address this issue, we proposed HiCCR, which reassigns credit at both the token and trajectory levels using a lightweight context attention concentration score. HiCCR requires no auxiliary models or additional data, adds less than 3% wall-clock overhead, and integrates easily into group-relative RLVR algorithms. Experiments on Qwen3-VL-4B and 8B across ten benchmarks show consistent improvements over base RL baselines, suggesting that effective multimodal RL benefits from credit signals that better reflect visual dependence.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 62521004) and the Science and Technology Commission of Shanghai Municipality (No. 24511103100). The authors declare no competing interests.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang,

Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025.

[2] 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.

[3] 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.

[4] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025.

[5] Hongzhan Chen, Tao Yang, Shiping Gao, Ruijun Chen, Xiaojun Quan, Hongtao Tian, and Ting Yao. Discriminative policy optimization for token-level reward models, 2025.

[6] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V`, 2025. Accessed: 2025-02-02.

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.

[9] Jie Cheng, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Gang Xiong, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. *arXiv preprint arXiv:2504.15275*, 2025.

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

[11] Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking, 2025.

[12] Yuhao Dong, Zuyan Liu, Shulin Tian, Yongming Rao, and Ziwei Liu. Insight-v++: Towards advanced long-chain visual reasoning with multimodal large language models, 2026.

[13] Henghui Du, Chang Zhou, Xi Chen, and Di Hu. Appo: Attention-guided perception policy optimization for video reasoning, 2026.

[14] Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025.

[15] Shu-Jian Gao, Jia-Mei Yan, Yu-Chen Yang, Peng-Hao Zhou, Qing-Lei Wang, Tie-Han Fan, Yuan Wang, Zu-Xuan Wu, and Yu-Gang Jiang. Live assistant: Learning whether, when, and whom to assist in real-world live social streams. 2026.

[16] Shujian Gao, Yuan Wang, Chenglong Ma, Xin Gao, Jiangtao Yan, Junzhi Ning, Cheng Tang, Changkai Ji, Huihui Xu, Wei Li, Ziyan Huang, Jiashi Lin, Ming Hu, Jiyao Liu, Wenhao Tang, Ye Du, Tianbin Li, Jin Ye, and Junjun He. Escaping low-rank traps: Interpretable visual concept learning via implicit vector quantization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 95771–95805, 2026.

[17] Shujian Gao, Yuan Wang, Jiangtao Yan, Zuxuan Wu, and Yu-Gang Jiang. Thinking with deltas: Incentivizing reinforcement learning via differential visual reasoning policy, 2026.

[18] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, and Yu Zhang. Reasoning-aligned perception decoupling for scalable multi-modal reasoning, 2026.

[19] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024.

[20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025.

[21] Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from language pre-training, 2025.

[22] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025.

[23] Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025.

[24] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2026.

[25] Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019.

[26] Zhengbo Jiao, Shaobo Wang, Zifan Zhang, Wei Wang, Bing Zhao, Hu Wei, and Linfeng Zhang. Credit where it is due: Cross-modality connectivity drives precise reinforcement learning for mllm reasoning. *arXiv preprint arXiv:2602.11455*, 2026.

[27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016.

[28] Zheng-Lun Kong, Yi-Ze Li, Fan-Hu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models - from vision, language to multimodality. *ArXiv*, abs/2505.18227, 2025.

[29] Nathan Lambert. Reinforcement learning from human feedback, 2026.

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.

[31] 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.

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[33] Xu Liu, Guikun Chen, and Wenguan Wang. Sinktrack: Attention sink based context anchoring for large language models. In *ICLR*, 2026.

[34] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning, 2025.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[36] Jinda Lu, Junkang Wu, Jinghan Li, Kexin Huang, Shuo Yang, Guoyin Wang, Jiancan Wu, Xiang Wang, and Xiangnan He. Bridging perception and reasoning: Token reweighting for rlvr in multimodal llms, 2026.

[37] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[39] Ruilin Luo, Chufan Shi, Yizhen Zhang, Cheng Yang, Songtao Jiang, Tongkun Guan, Ruizhe Chen, Ruihang Chu, Peng Wang, Mingkun Yang, et al. From narrow to panoramic vision: Attention-guided cold-start reshapes multimodal reasoning. *arXiv preprint arXiv:2603.03825*, 2026.

[40] 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.

[41] Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. Sparse but critical: A token-level analysis of distributional shifts in rlvr fine-tuning of llms, 2026.

[42] MiniMax, :, Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention, 2025.

[43] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2025.

[44] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card, 2024.

[45] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[46] Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. Grpo-λ: Credit assignment improves llm reasoning, 2025.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

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

[49] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[50] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM, March 2025.

[51] Marco Simoni, Aleksandar Fontana, Giulio Rossolini, Andrea Saracino, and Paolo Mori. Gtpo: Stabilizing group relative policy optimization via gradient and entropy control, 2025.

[52] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.

[53] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[54] Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Peter Tu, and Jing Zhang. All roads lead to rome: Incentivizing divergent thinking in vision-language models, 2026.

[55] Hieu Tran, Zonghai Yao, and Hong Yu. Exploiting tree structure for credit assignment in rl training of llms, 2025.

[56] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[57] 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.

[58] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

[59] Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. Llava-critic-r1: Your critic model is secretly a strong policy model, 2025.

[60] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.

[61] Yuan Wang, Shu-Jian Gao, Jiaxiang Liu, Songtao Jiang, Haoxiang Xia, Xiaotian Zhang, Zhao-Lu Kang, Ye-Min Wang, and Zuozhu Liu. Beyond n-grams: A hierarchical reward learning framework for clinically-aware medical report generation. In *AAAI Conference on Artificial Intelligence*, 2025.

[62] Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-guided policy optimization for multimodal reasoning, 2026.

[63] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025.

[64] Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-ktr: Reinforcing video reasoning via key token attribution, 2026.

[65] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024.

[66] Guofu Xie, Yunsheng Shi, Hongtao Tian, Ting Yao, and Xiao Zhang. Capo: Towards enhancing llm reasoning through generative credit assignment, 2025.

[67] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025.

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

[69] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan 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, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025.

[70] Zhuoran Yu and Yong Jae Lee. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding, 2025.

[71] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

[72] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025.

[73] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023.

[74] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

[75] Zheng Zhang, Ao Lu, Yuanhao Zeng, Ziwei Shan, Jinjin Guo, Lufei Li, Yexin Li, and Kan Ren. Grad2reward: From sparse judgment to dense rewards for improving open-ended llm reasoning, 2026.

[76] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025.

[77] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

Appendix Roadmap

This appendix provides supplementary experiments, theory, proxy validation, mechanistic diagnostics, reproducibility details, qualitative examples, limitations, and broader impacts.

Appendix A **Additional Experimental Evidence** EXPERIMENTS

Additional results on efficiency, scaling, architecture transfer, training dynamics, benchmarks, and context-credit quality.

- **A.1 Computational Overhead**: wall-clock and memory cost.
- **A.2 Data Scaling Verification**: Geo3K-only low-data training.
- **A.3 Model Architecture Generalization**: transfer to InternVL3-8B.
- **A.4 Training Dynamics**: accuracy, length, gradient norm, and entropy.
- **A.5 Remaining Results**: additional benchmarks and ablations.
- **A.6 Context Credit Analysis**: credit distributions and semantic alignment.

Appendix B **Theoretical Analysis: From Uniform Credit to Attention Drift** THEORY

Formalizes text-favored gradient bias under uniform credit and explains how HiCCR mitigates the resulting visual attention drift.

Appendix C **Context Credit as a Lightweight Visual-Dependence Proxy** PROXY VALIDATION

Justifies visual-attention concentration as an efficient proxy for visual dependence in RLVR training.

Appendix D **Mechanistic Interpretation for Visual Grounding** DIAGNOSTICS

Reports gradient attribution, counterfactual visual ablation, and attention dynamics across training stages.

Appendix E **Related Work** POSITIONING

Discusses multimodal reasoning, RLVR, attention-based grounding, and token-level credit assignment.

Appendix F **Training and Evaluation Details** REPRODUCIBILITY

Provides implementation details, hyperparameters, evaluation protocol, and the HiCCR algorithm pipeline.

Appendix G **Qualitative Case Studies** CASE ANALYSIS

Shows representative examples of visually grounded credit reassignment.

Appendix H **Limitations** SCOPE

Discusses assumptions, scope, and possible extensions.

Appendix I **Broader Impacts** IMPACT

Discusses potential benefits, risks, and responsible-use considerations.

A More Experiments Results

A.1 Computational Overhead

A practical credit-reassignment method must not introduce prohibitive overhead into the already expensive RLVR pipeline. Table 6 compares the computational cost of HiCCR against uniform baselines (GRPO, DAPO) and the

Table 6: **Computational overhead comparison (Qwen3-VL-8B).** All methods are trained on 8× High-Performance GPUs with identical batch size, rollout configuration, and total training steps. *Rel. Overhead* = wall-clock increase vs. the matched uniform baseline (GRPO for HiCCR-G, DAPO for HiCCR-D); (+Δ) on overhead columns marks cost increase (lower is better), (+Δ) on AVG@8 marks accuracy gain. HiCCR adds only 2.7% wall-clock overhead while delivering +4.70 / +4.47 AVG@8; PAPO requires >20% overhead for a smaller gain and an extra backward pass.

Method	Step Time (sec)	Rel. Overhead (%)	Peak Mem (GB)	Mem Overhead (%)	Extra Backward	AVG@8
GRPO	619	–	115.8	–	✗	62.49
HiCCR-G (Ours)	636 (+2.7)	+2.7	121.0 (+4.5)	+4.5	✗	67.19 **(+4.70)**
DAPO	634	–	118.1	–	✗	68.29
HiCCR-D (Ours)	651 (+2.7)	+2.7	123.4 (+4.5)	+4.5	✗	**72.76 (+4.47)**
PAPO *(perturbation-based, context)*	782 (+26.3)	+26.3	129.7 (+12.0)	+12.0	✓	70.25 (+1.96)

Table 7: Data scaling verification. Models are fine-tuned exclusively on the Geo3K training split to evaluate data efficiency. (+Δ) / (−Δ) are per-cell differences against the matched base RL row (grey). HiCCR variants consistently surpass base RL algorithms across both model scales under this limited data regime; best AVG@8 per backbone in **bold**.

Method	Mathematical Reasoning					General & Logical Reasoning				AVG@8
	Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	Clevr	Logic	MMMU	
***Qwen3-VL-4B** Backbone (Trained on Geo3K Train)*										
Base Model (Instruct)	45.42	43.39	65.80	53.39	40.90	39.44	90.00	42.73	20.52	49.07
GRPO	52.91	53.85	65.40	62.76	47.71	45.35	87.50	44.74	25.84	54.01
HiCCR-G (Ours)	54.81 (+1.90)	56.93 (+3.08)	68.50 (+3.10)	67.18 (+4.42)	52.06 (+4.35)	49.28 (+3.93)	88.50 (+1.00)	45.64 (+0.90)	27.34 (+1.50)	**56.69 (+2.68)**
DAPO	48.42	49.63	64.60	59.22	45.18	39.66	88.00	42.73	25.84	51.48
HiCCR-D (Ours)	51.75 (+3.33)	53.35 (+3.72)	67.20 (+2.60)	64.25 (+5.03)	49.62 (+4.44)	45.35 (+5.69)	89.00 (+1.00)	43.62 (+0.89)	27.75 (+1.91)	54.65 **(+3.17)**
***Qwen3-VL-8B** Backbone (Trained on Geo3K Train)*										
Base Model (Instruct)	51.58	53.99	65.70	60.29	50.08	45.91	89.50	47.87	36.47	55.71
GRPO	54.58	59.86	69.20	67.53	55.05	49.00	86.00	50.56	38.21	58.89
HiCCR-G (Ours)	59.90 (+5.32)	65.05 (+5.19)	69.30 (+0.10)	70.06 (+2.53)	58.64 (+3.59)	53.10 (+4.10)	87.00 (+1.00)	53.91 (+3.35)	41.68 (+3.47)	**62.07 (+3.18)**
DAPO	55.41	58.72	65.80	65.06	52.98	49.00	89.50	50.78	39.94	58.58
HiCCR-D (Ours)	60.12 (+4.71)	62.97 (+4.25)	67.40 (+1.60)	68.74 (+3.68)	55.12 (+2.14)	52.36 (+3.36)	89.50 (+0.00)	53.66 (+2.88)	42.78 (+2.84)	61.41 **(+2.83)**

perturbation-based method PAPO under identical hardware and training configurations (8× High-Performance GPUs, same batch size, rollout length, and total training steps).

HiCCR computes the context credit score via forward-pass attention hooks on the last two decoder layers, requiring no additional backward pass and no attention-weight materialization. As shown in Table 6, the per-step wall-clock overhead is only 2.7% relative to the uniform baselines, well within the $< 3\%$ budget claimed in §1. Peak GPU memory increases by 4.5%, as the only additional storage is the pre-softmax attention logits at visual key positions for the target layers. In contrast, PAPO requires counterfactual perturbation passes, incurring 26.3% wall-clock overhead and 12.0% additional memory for comparable or lower accuracy.

A.2 Data Scaling Verification

To validate data efficiency, we restrict training to the Geo3K training split alone, a minimal corpus covering only geometry problems. As shown in Table 7, HiCCR consistently outperforms its base RL algorithm across both backbones: HiCCR-G exceeds GRPO by +2.69 (4B) and +3.18 (8B), while HiCCR-D surpasses DAPO by +3.18 (4B) and +2.83 (8B). These gains confirm that the effectiveness of HiCCR is algorithm-driven and robust to data scarcity, rather than an artifact of large-scale data memorization.

A.3 Model Architecture Generalization

To verify that HiCCR's benefits are not tied to a single MLLM family, we transplant the full pipeline to InternVL3-8B [77], whose vision encoder, connector, and language backbone differ substantially from Qwen3-VL, while keeping the ViRL39K training set, rollout configuration ($G = 5$, temperature 1.0), and HiCCR hyperparameters unchanged. As shown in Table 8, HiCCR-G improves GRPO by $+3.21$ AVG@8 points, and HiCCR-D improves DAPO by $+3.23$ points, with consistent gains across mathematical reasoning, visual counting, and logical reasoning benchmarks. These results indicate that HiCCR provides additional benefits on top of both GRPO and

Table 8: Model architecture generalization. We replace the Qwen3-VL backbone with InternVL3-8B and re-run the full RLVR pipeline on ViRL39K under identical hyperparameters. (+Δ) / (−Δ) are per-cell differences against the matched base RL row (grey). HiCCR variants consistently surpass base RL algorithms across both frameworks, confirming that the gains transfer across VLM families with distinct vision encoders, connectors, and language priors; best AVG@8 in **bold**.

Method	Mathematical Reasoning					General & Logical Reasoning				AVG@8
	Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	Clevr	Logic	MMMU	
***InternVL3-8B** Backbone (Trained on ViRL39K)*										
Base Model (Instruct)	36.77	48.17	54.70	54.77	42.35	48.29	71.50	40.94	29.77	47.47
GRPO	42.28	58.62	63.70	63.68	52.28	45.96	82.00	51.39	39.31	55.47
HiCCR-G (Ours)	45.87 (+3.59)	63.46 (+4.84)	66.15 (+2.45)	66.21 (+2.53)	54.69 (+2.41)	48.54 (+2.58)	85.50 (+3.50)	55.92 (+4.53)	41.78 (+2.47)	58.68 **(+3.21)**
DAPO	44.16	60.55	65.50	65.91	54.12	48.03	84.00	53.64	41.42	57.48
HiCCR-D (Ours)	48.74 (+4.58)	63.21 (+2.66)	69.05 (+3.55)	69.47 (+3.56)	56.55 (+2.43)	51.71 (+3.68)	86.50 (+2.50)	56.12 (+2.48)	45.07 (+3.65)	**60.71 (+3.23)**

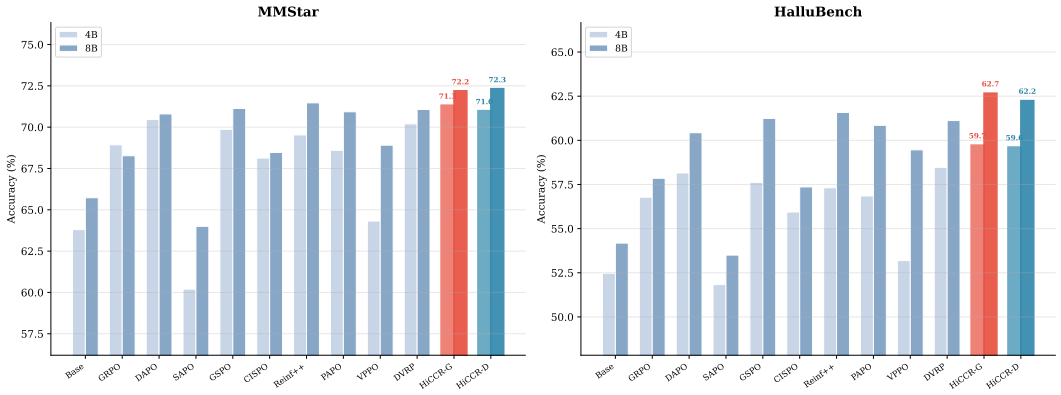


Figure 5: **Performance on MMStar and HalluBench.** Grouped bars compare all methods on Qwen3-VL-4B and 8B backbones. HiCCR variants (red/blue) consistently achieve the highest accuracy on both benchmarks, indicating improved visual perception and reduced hallucination.

DAPO, and that its concentration-based credit signal transfers across distinct VLM architectures rather than relying on architecture-specific tuning.

A.4 Training Dynamics

Figure 3 traces four diagnostic metrics across training for all methods on the Qwen3-VL-4B backbone. **(a) Reward accuracy:** HiCCR variants converge faster and stabilize at a higher plateau (∼0.80–0.82) with markedly lower variance, whereas GRPO oscillates persistently in the 0.72–0.80 range. **(b) Response length:** Both baselines exhibit monotonic length inflation—a hallmark of reward hacking—while HiCCR *shortens* responses over training (HiCCR-D reaches ∼800 tokens), achieving higher accuracy with more concise reasoning chains. **(c) Gradient norm:** GRPO gradients remain near-zero throughout, signaling weak optimization signal; DAPO norms escalate sharply in later stages, risking instability. HiCCR maintains gradient norms in a healthy, bounded regime (0.15–0.25), consistent with the bounded-modulation design ($\alpha \geq 1,\ \beta \geq 1$) that amplifies without destabilizing. **(d) Entropy:** Baseline entropies rise continuously, reflecting growing policy uncertainty. HiCCR-D entropy *decreases* to ∼0.30, indicating that the model learns a confident, visually grounded policy rather than hedging between textual and visual cues. Taken together, these dynamics show that HiCCR's credit reassignment not only improves final accuracy but also yields a qualitatively healthier optimization trajectory: shorter outputs, stable gradients, and decisive policies.

A.5 Remaining Results

We present remaining experiment results according to the main content order.

Table 9: Comprehensive comparison with multimodal models. Benchmarks are grouped into **Mathematical Reasoning** (with sub-average) and **General & Logical Reasoning** (with sub-average). Missing scores are denoted by "–"; sub-averages are computed over available entries only. Subscripts $_D$ and $_G$ denote HiCCR built on DAPO and GRPO. (+Δ) marks absolute gain over the matched Base Model (grey). Best within each backbone in **bold**.

Method	Mathematical Reasoning						General & Logical Reasoning					AVG@8
	Geo3K	MVista	WeMath	MVerse	MV-V	*Avg.*	MMK12	Clevr	Logic	MMMU	*Avg.*	
Proprietary Models												
GPT-4o	42.31	64.70	62.84	50.15	53.77	*54.75*	55.80	95.00	64.83	59.12	*68.69*	60.95
GPT-4o-mini	34.58	59.90	56.27	42.36	45.14	*47.65*	51.90	88.00	54.37	46.19	*60.12*	53.19
Gemini-2.0-flash	48.86	70.40	47.45	47.79	48.62	*52.62*	65.20	96.00	67.94	60.31	*72.36*	61.40
General-Purpose Multimodal VLMs												
Qwen2.5-VL-72B	45.41	74.20	49.08	47.34	48.55	*52.92*	70.50	92.50	65.82	61.47	*72.57*	61.65
InternVL2.5-8B	33.19	64.90	44.86	37.04	40.15	*44.03*	46.80	85.00	51.76	42.09	*56.41*	49.53
InternVL2.5-78B	44.57	64.90	44.88	36.95	40.24	*46.31*	59.80	90.00	61.23	54.81	*66.46*	55.26
LLaVA-OV-7B	32.11	58.50	44.17	–	–	*44.93*	43.89	82.00	46.21	37.94	*52.51*	49.26
LLaVA-OV-72B	41.32	67.10	31.97	27.24	30.08	*39.54*	54.27	88.50	54.83	51.62	*62.31*	49.66
LLaVA-OV-1.5-8B	36.85	69.60	61.47	–	–	*55.97*	48.55	86.00	53.79	43.15	*57.87*	57.06
LLaVA-Critic-R1-7B	35.43	68.70	62.58	58.85	53.14	*55.74*	57.40	87.50	55.83	42.67	*60.85*	58.01
R1-OneVision-7B	30.56	64.90	55.23	61.72	44.27	*51.34*	43.30	84.00	52.19	39.71	*54.80*	52.88
Domain-Specific Multimodal VLMs												
MM-Eureka-7B	36.38	59.10	45.34	57.55	56.46	*50.97*	60.60	80.00	49.12	36.24	*56.49*	53.42
MM-Eureka-7B-CPGD	37.64	64.20	64.28	63.66	59.25	*57.81*	64.70	81.50	48.73	37.51	*58.11*	57.94
ADORA-7B	41.17	61.10	53.06	45.18	41.83	*48.47*	49.80	78.00	44.62	33.95	*51.59*	49.86
R1-VL-7B	31.86	63.50	56.14	42.05	43.18	*47.35*	55.30	82.50	48.16	39.27	*56.31*	51.33
VLAA-Thinker-7B	24.23	67.40	65.87	47.92	51.96	*51.48*	63.20	83.00	49.88	40.54	*59.16*	54.89
VL-Rethinker-7B	33.64	61.30	66.46	63.95	60.83	*57.24*	59.80	84.50	54.12	41.73	*60.04*	58.48
RACRO-7B-CRO-GRPO	41.42	61.70	68.87	65.73	61.68	*59.88*	70.50	85.00	53.81	44.15	*63.37*	61.43
ThinkLite-7B-VL	34.36	68.90	63.52	49.48	46.05	*52.46*	56.20	81.00	50.36	38.82	*56.60*	54.30
AVAR-Thinker-7B	32.95	58.80	49.66	53.42	52.16	*49.40*	46.92	80.00	41.48	38.34	*51.69*	50.41
Ours												
Base Model *(Qwen3-VL-4B)*	45.42	65.80	53.39	43.39	40.90	*49.78*	39.44	90.00	42.73	20.52	*48.17*	49.07
+ **HiCCR**$_D$	**63.23** (+17.81)	73.50 (+7.70)	75.69 (+22.30)	73.21 (+29.82)	71.18 (+30.28)	*71.36* ***(+21.58)***	65.54 (+26.10)	91.50 (+1.50)	**60.40** (+17.67)	46.30 (+25.78)	*65.93* ***(+17.76)***	68.95 **(+19.88)**
+ **HiCCR**$_G$	61.73 (+16.31)	71.70 (+5.90)	78.62 (+25.23)	**77.34** (+33.95)	**73.85** (+32.95)	***72.65 (+22.87)***	**68.64** (+29.20)	**94.00** (+4.00)	57.05 (+14.32)	**46.53** (+26.01)	***66.55 (+18.38)***	**69.94 (+20.87)**
Base Model *(Qwen3-VL-8B)*	51.58	65.70	60.29	53.99	50.08	*56.33*	45.91	89.50	47.87	36.47	*54.94*	55.71
+ **HiCCR**$_D$	**68.39** (+16.81)	**75.70** (+10.00)	**81.03** (+20.74)	**78.94** (+24.95)	**75.99** (+25.91)	***76.01 (+19.68)***	**70.58** (+24.67)	**92.00** (+2.50)	**60.63** (+12.76)	**51.56** (+15.09)	***68.69 (+13.75)***	**72.76 (+17.05)**
+ **HiCCR**$_G$	58.40 (+6.82)	73.50 (+7.80)	76.49 (+16.20)	72.29 (+18.30)	68.27 (+18.19)	*69.79* ***(+13.46)***	62.11 (+16.20)	90.00 (+0.50)	57.49 (+9.62)	46.18 (+9.71)	*63.94* ***(+9.00)***	67.19 **(+11.48)**

Table 10: Ablation of credit *distribution* strategies under **DAPO**. All non-uniform rows use the same credit scores from our visual-attention proxy; only the score-to-weight mapping differs. *Hard*: tokens below threshold receive zero weight. *Soft*: all tokens are rescaled continuously. Δ: change vs. Uniform. Best per backbone in **bold**.

Category	Strategy	Mathematical Reasoning					General & Logical Reasoning				AVG	Δ
		Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	CLEVR	Logic	MMMU		
Qwen3-VL-4B												
–	Uniform (Baseline)	60.57	74.40	72.30	**79.48**	70.26	60.56	**92.00**	54.59	44.05	67.58	–
Hard	Top-20% Mask	59.07	70.92	73.20	74.20	67.43	59.46	85.50	55.70	41.45	65.21	-2.37
	Top-40% Mask	63.19	73.50	73.30	76.18	70.26	61.28	90.00	59.73	41.39	67.65	+0.07
Soft	**HiCCR (Ours)**	**63.23**	**73.21**	**73.50**	75.69	**71.18**	**65.54**	91.50	**60.40**	**46.30**	**68.95**	**+1.37**
Qwen3-VL-8B												
–	Uniform (Baseline)	64.23	73.39	71.10	77.13	68.88	63.99	89.00	57.49	49.36	68.29	–
Hard	Top-20% Mask	63.06	72.48	74.60	76.78	68.43	61.28	89.00	58.17	47.57	67.93	-0.36
	Top-40% Mask	62.40	74.40	73.70	77.18	70.03	64.66	**92.50**	58.39	46.59	68.87	+0.58
Soft	**HiCCR (Ours)**	**68.39**	**78.94**	**75.70**	**81.03**	**75.99**	**70.58**	92.00	**60.63**	**51.56**	**72.76**	**+4.47**

A.6 Context Credit Quantization and Quality Analysis

To understand how HiCCR redistributes optimization pressure across tokens and trajectories, we conduct a fine-grained credit analysis on 500 samples from the ViRL training set using a Qwen3-VL-8B checkpoint trained with HiCCR-G. Figure 4 visualizes the resulting credit distributions at both the token and trajectory levels.

Token-level credit weight β_t. As shown in Figure 4(a), the token-level credit weight $\beta_t = 1 + \lambda_{\text{tok}} \cdot \sigma(\bar{s}_t)$ follows an approximately Gaussian distribution centered at μ=1.249 with σ=0.105, spanning the effective range $[1.0, 1.5]$. This confirms that the context credit mechanism produces a smooth, well-spread weighting rather than a degenerate or binary assignment, enabling fine-grained differentiation among tokens based on their visual grounding.

Trajectory-level credit weight $\alpha^{(i)}$. Figure 4(b) reveals a right-skewed distribution of trajectory-level weights, with the majority of trajectories receiving moderate weights ($\alpha \in [1.1, 1.5]$) and a long tail extending toward α=2.0. This indicates that HiCCR selectively amplifies gradients from trajectories whose reasoning

Table 11: Ablation of **target layers** used for context credit estimation under **GRPO** on the **Qwen3-VL-4B** backbone. We report the original GRPO baseline and HiCCR-G with different target-layer choices. All HiCCR variants share the same training recipe; the only difference is which two transformer layers are used to extract the visual-attention proxy.

Method / Target Layers	Mathematical Reasoning					General & Logical Reasoning				AVG@8	Δ
	Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	Clevr	Logic	MMMU		
Original GRPO (Baseline)	53.41	67.16	68.70	68.45	63.15	55.97	87.50	52.35	40.87	61.95	–
HiCCR-G w/ First Two Layers	61.56	72.94	70.10	76.32	70.34	61.78	91.00	55.96	44.74	67.19	+5.24
HiCCR-G w/ Middle Two Layers	60.43	76.56	71.20	76.66	72.34	65.92	91.50	55.92	45.86	68.49	+6.54
HiCCR-G w/ Last Two Layers (Ours)	**61.73**	**77.34**	**71.70**	**78.62**	**73.85**	**68.64**	**94.00**	**57.05**	**46.53**	**69.94**	**+7.99**

Table 12: Ablation of **asymmetric** *vs.* **symmetric** advantage shaping under **GRPO** on two model scales. **Asymmetric** applies HiCCR credit reassignment only to positive-advantage trajectories; **Symmetric** applies it to both positive and negative advantages.

Method	Mathematical Reasoning					General & Logical Reasoning				AVG@8	Δ
	Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	Clevr	Logic	MMMU		
***Qwen3-VL-4B** Backbone*											
Original GRPO (Baseline)	53.41	67.16	68.70	68.45	63.15	55.97	87.50	52.35	40.87	61.95	–
HiCCR-G w/ Symmetric	55.37	71.03	69.50	70.39	66.94	62.46	89.50	55.64	42.38	64.80	+2.85
HiCCR-G w/ Asymmetric (Ours)	**61.73**	**77.34**	**71.70**	**78.62**	**73.85**	**68.64**	**94.00**	**57.05**	**46.53**	**69.94**	**+7.99**
***Qwen3-VL-8B** Backbone*											
Original GRPO (Baseline)	56.74	65.23	68.40	68.91	61.93	54.81	90.50	52.57	43.35	62.49	–
HiCCR-G w/ Symmetric	55.31	67.93	69.10	66.78	64.87	57.67	89.00	53.96	44.11	63.19	+0.70
HiCCR-G w/ Asymmetric (Ours)	**58.40**	**72.29**	**73.50**	**76.49**	**68.27**	**62.11**	**90.00**	**57.49**	**46.18**	**67.19**	**+4.70**

chains exhibit strong visual engagement, while down-weighting trajectories that rely less on visual context—a desirable property for multimodal RLVR.

High-credit token semantics. The word cloud in Figure 4(c) displays tokens with the highest average credit scores. Dominant terms include geometric primitives (*triangle*, *circle*, *angle*, *square*), spatial relations (*parallel*, *perpendicular*, *row*, *side*), diagram references (*image*, *diagram*, *point*), and geometric vertex labels (*AB*, *BC*, *AC*, *AD*). This qualitatively demonstrates that HiCCR assigns higher credit to tokens that are semantically grounded in visual content.

Credit decay along generation. Figure 4(d) plots the mean context credit score $\bar{s}_t$ against relative token position in the response. A monotonic decay of $\Delta{=}-1.81$ is observed from the beginning to the end of generation. This pattern is consistent with the intuition that early response tokens—which typically establish the visual interpretation and set up the reasoning scaffold—attend more strongly to the image, whereas later tokens progressively shift toward symbolic computation and answer formatting. And this finding aligns with our attention pattern experiments D.3.

Visual vs. non-visual token credits. Finally, we classify each response token as *visual* or *non-visual* based on an independent semantic lexicon of $\sim$150 geometry and spatial terms. As shown in Figure 4(e), visual tokens receive a significantly higher median credit weight ($\tilde{\beta}{=}1.303$) than non-visual tokens ($\tilde{\beta}{=}1.236$), with $p < 0.001$ under a Mann–Whitney U test and a medium effect size of Cohen's $d{=}0.547$. This provides direct quantitative evidence that HiCCR's credit mechanism successfully up-weights visually grounded tokens without any explicit supervision on token-level labels.

B Theoretical Analysis: From Uniform Credit to Attention Drift

We analyze how uniform credit assignment can bias multimodal RLVR updates toward non-visual context. The goal of this section is not to prove a global training guarantee, but to isolate a local first-order mechanism: under mild structural conditions, the same trajectory-level advantage can allocate a larger gradient budget to non-visual tokens and induce a local drift of attention away from visual tokens. We then show how HiCCR changes the corresponding diagonal update through token- and trajectory-level reweighting.

Throughout this section, we analyze a single attention head in a single layer ℓ and suppress layer/head superscripts unless needed. The extension to the full transformer is discussed at the end as a local decomposition argument rather than a separate global theorem.

Table 13: Sensitivity of λ_{tok} (token-level scaling magnitude) under **DAPO** on the **Qwen3-VL-4B** backbone. λ_{tok} controls the maximum amplitude of the per-token credit multiplier β_t. Δ: improvement over the DAPO baseline. Best in **bold**. Default is the setting used in other experiments.

Method / λ_{tok}	**Mathematical Reasoning**					**General & Logical Reasoning**				**AVG@8**	Δ
	Geo3K	MVerse	MVista	WeMath	MV-V	MMK12	Clevr	Logic	MMMU		
***Qwen3-VL-4B** Backbone*											
Original DAPO (Baseline)	60.59	**74.42**	72.30	**79.45**	70.29	60.58	**92.00**	54.55	44.02	67.58	–
HiCCR-D w/ $\lambda_{\text{tok}} = 0.1$	61.18	74.28	72.25	78.92	70.48	61.29	91.95	55.68	44.48	67.83	+0.25
HiCCR-D w/ $\lambda_{\text{tok}} = 0.2$	61.89	73.91	72.55	78.07	70.72	62.71	91.75	56.82	44.91	68.15	+0.57
HiCCR-D w/ $\lambda_{\text{tok}} = 0.5$ (Default)	63.27	73.26	73.50	75.62	71.15	**65.58**	91.50	**60.47**	**46.37**	68.97	+1.39
HiCCR-D w/ $\lambda_{\text{tok}} = 0.8$	**63.48**	74.37	**73.85**	77.28	**71.59**	65.32	91.80	59.81	46.28	**69.31**	+1.73
HiCCR-D w/ $\lambda_{\text{tok}} = 1.5$	59.82	73.08	71.75	75.19	69.41	59.87	90.75	52.83	43.49	66.24	-1.34

B.1 Setup and Notation

Let $x = (x_{\text{vis}}, x_{\text{text}})$ denote the multimodal prompt. At generation step t, let $\mathcal{V} \subset \{1, \dots, |x|\}$ index visual tokens, and let $\mathcal{T}_t \subset \{1, \dots, |x| + t - 1\}$ index all non-visual positions, including textual prompt tokens and the generated prefix $y_{<t}$. Thus, throughout this section, "textual" should be read as "non-visual" unless otherwise specified. We write $\mathcal{T}$ for $\mathcal{T}_t$ when t is clear.

Let

$$z_j^{(\ell)} = \mathrm{LN}\Big(h_j^{(\ell)}\Big)$$

be the normalized hidden state used by a pre-norm attention block. The attention branch computes

$$a_t^{(\ell)} = \sum_{j=1}^{|x|+t-1} \alpha_{tj} W_V z_j^{(\ell)}, \qquad \alpha_{tj} = \mathrm{softmax}_j(e_{tj}), \qquad e_{tj} = \frac{q_t^\top k_j}{\sqrt{d_h}}, \tag{11}$$

where

$$q_t = W_Q z_t^{(\ell)}, \qquad k_j = W_K z_j^{(\ell)}.$$

The residual branch is omitted from the derivation because it does not depend on W_Q and therefore does not affect the W_Q-gradient considered below. For notational economy, we write $h_j^{(\ell)}$ for the normalized state $z_j^{(\ell)}$ in the rest of the analysis.

The visual and non-visual attention masses are

$$\alpha_t^{\mathcal{V}} = \sum_{j \in \mathcal{V}} \alpha_{tj}, \qquad \alpha_t^{\mathcal{T}} = \sum_{j \in \mathcal{T}} \alpha_{tj}. \tag{12}$$

Since $\mathcal{V} \cup \mathcal{T}$ covers the attended positions, $\alpha_t^{\mathcal{V}} + \alpha_t^{\mathcal{T}} = 1$.

We focus on the local unclipped regime of group-relative RLVR, where the importance ratio satisfies $r_t(\theta) \approx 1$ and the clipped objective reduces locally to the standard policy-gradient term. For one trajectory with scalar advantage $\hat{A}$, the local objective is

$$\mathcal{L}(\theta) = -\hat{A} \sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{13}$$

This isolates the credit-assignment structure; clipping gates the magnitude of updates but does not change the fact that standard RLVR assigns the same $\hat{A}$ to every token in the trajectory. Unless otherwise stated, the drift results below are conditional on $\hat{A} > 0$. For group-relative objectives, the full batch-level update additionally depends on the correlation between $\hat{A}^{(i)}$ and the local drift coefficients across sampled trajectories.

B.2 Gradient Decomposition for Attention Parameters

Let

$$d_t \triangleq \frac{\partial \log \pi_\theta(y_t \mid x, y_{<t})}{\partial h_t^{(\ell+1)}} \in \mathbb{R}^d, \qquad v_j = W_V h_j^{(\ell)}.$$

By the softmax Jacobian,

$$\frac{\partial \alpha_{tj}}{\partial e_{tk}} = \alpha_{tj}(\mathbf{1}\{j = k\} - \alpha_{tk}), \tag{14}$$

the gradient of the token log-probability with respect to W_Q can be written as

$$\frac{\partial \log \pi_\theta(y_t \mid x, y_{<t})}{\partial W_Q} = \sum_j \alpha_{tj} \left(d_t^\top v_j - \bar{s}_t\right) \frac{\partial e_{tj}}{\partial W_Q}, \tag{15}$$

where

$$\bar{s}_t = \sum_k \alpha_{tk} d_t^\top v_k = \alpha_t^{\mathcal{V}} \mu_{\mathcal{V},t} + \alpha_t^{\mathcal{T}} \mu_{\mathcal{T},t}, \tag{16}$$

with

$$\mu_{\mathcal{S},t} = \frac{1}{\alpha_t^{\mathcal{S}}} \sum_{j \in \mathcal{S}} \alpha_{tj} d_t^\top v_j, \qquad \mathcal{S} \in \{\mathcal{V}, \mathcal{T}\}. \tag{17}$$

Definition B.1 (Per-modality gradient contribution)**.** *For $\mathcal{S} \in \{\mathcal{V}, \mathcal{T}\}$, define*

$$G_t^{\mathcal{S}} = \sum_{j \in \mathcal{S}} \alpha_{tj} \left(d_t^\top v_j - \bar{s}_t \right) \frac{\partial e_{tj}}{\partial W_Q}. \tag{18}$$

Then

$$\frac{\partial \log \pi_\theta(y_t \mid x, y_{<t})}{\partial W_Q} = G_t^{\mathcal{V}} + G_t^{\mathcal{T}}.$$

B.3 Structural Conditions

We use two types of conditions. The first captures the language-prior asymmetry of pretrained VLMs. The remaining conditions are local regularity assumptions needed to convert activation asymmetry into gradient and attention-update statements.

Assumption B.1 (Language-prior dominance)**.**

(a) ***Absolute-scale dominance.*** *For each generation step t, there exists $\kappa_t > 1$ such that*

$$\mathbb{E}_{j \sim \mathrm{Unif}(\mathcal{T})} \left[|d_t^\top v_j| \right] \geq \kappa_t \mathbb{E}_{j \sim \mathrm{Unif}(\mathcal{V})} \left[|d_t^\top v_j| \right]. \tag{19}$$

(b) ***Signed alignment gap.*** *For a subset $\mathcal{H}$ of generation steps, there exists $\delta_t > 0$ such that*

$$\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t} \geq \delta_t, \qquad t \in \mathcal{H}. \tag{20}$$

We call such steps text-dominant *steps.*

The two parts are logically distinct. Assumption B.1(a) states that non-visual value activations have larger absolute scale on average. Assumption B.1(b) states that, on text-dominant steps, the attention-weighted signed activation is larger on the non-visual side. The empirical analysis below supports the absolute-scale component; the signed component is used as a local condition for the drift result.

Assumption B.2 (Low cancellation)**.** *For $\mathcal{S} \in \{\mathcal{V}, \mathcal{T}\}$, there exists $c_{\mathcal{S}} \in (0, 1]$ such that*

$$\left\| \sum_{j \in \mathcal{S}} \alpha_{tj} \left(d_t^\top v_j - \bar{s}_t \right) \frac{\partial e_{tj}}{\partial W_Q} \right\|_F \geq c_{\mathcal{S}} \sum_{j \in \mathcal{S}} \alpha_{tj} \left| d_t^\top v_j - \bar{s}_t \right| \left\| \frac{\partial e_{tj}}{\partial W_Q} \right\|_F. \tag{21}$$

This excludes adversarial cancellation among same-modality gradient directions.

Assumption B.3 (Jacobian-scale regularity)**.** *There exist constants $0 < c_J \leq C_J < \infty$ and a step-dependent scale $J_t > 0$ such that for all attended positions j,*

$$c_J J_t \leq \left\| \frac{\partial e_{tj}}{\partial W_Q} \right\|_F \leq C_J J_t. \tag{22}$$

Since

$$\frac{\partial e_{tj}}{\partial W_Q} = \frac{k_j h_t^{(\ell)\top}}{\sqrt{d_h}},$$

this condition corresponds to comparable key-vector norms at the analyzed layer.

Assumption B.4 (Local key-metric alignment)**.** *Let*

$$G_t = G_t^{\mathcal{V}} + G_t^{\mathcal{T}}.$$

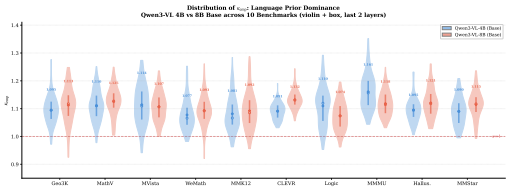
For text-dominant steps $t \in \mathcal{H}$, the key-induced metric does not reverse the logit-space direction of the signed gap:

$$-\left\langle \frac{\partial \alpha_t^{\mathcal{V}}}{\partial W_Q}, G_t \right\rangle_F \geq \zeta \, C_t \, \alpha_t^{\mathcal{V}} \alpha_t^{\mathcal{T}} \left(\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t} \right), \qquad \zeta > 0, \tag{23}$$

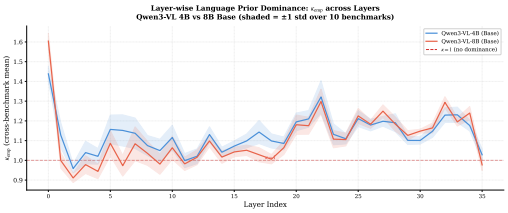
where

$$C_t = \frac{\|h_t^{(\ell)}\|^2}{d_h} \sum_j \alpha_{tj} \|k_j\|^2 > 0. \tag{24}$$

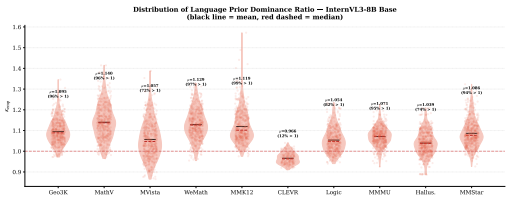
This assumption is used only to sign the same-step diagonal attention update. It is weaker and more explicit than requiring a global attention-dynamics guarantee.



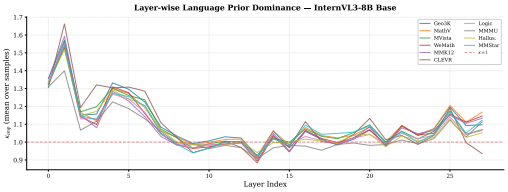
(a) Qwen3-VL per-benchmark κ_{emp} (last two layers).



(b) Qwen3-VL layer-wise κ_{emp}.



(c) InternVL3-8B per-benchmark κ_{emp} (last two layers).



(d) InternVL3-8B layer-wise κ_{emp}.

Figure 6: Empirical support for the absolute-scale component of language-prior dominance. Across the evaluated benchmarks and model families, κ_{emp} is consistently above one in the measured layers, suggesting that non-visual context often carries larger downstream-projected activation magnitude than visual context.

Empirical support for absolute-scale LPD. We measure

$$\kappa_{\text{emp}} = \frac{\mathbb{E}_{j\in\mathcal{T}}\left[|d_t^\top v_j|\right]}{\mathbb{E}_{j\in\mathcal{V}}\left[|d_t^\top v_j|\right]}$$

on Qwen3-VL and InternVL3 using full forward–backward passes. This experiment supports Assumption B.1(a), but does not claim to verify all regularity assumptions.

B.4 Main Results

We first relate the signed gap to the centered quantities that appear in the attention-gradient decomposition.

Lemma B.2 (Centered activation transfer)**.** *Fix a text-dominant step $t \in \mathcal{H}$. Define*

$$\sigma_{\mathcal{V},t} = \frac{1}{\alpha_t^{\mathcal{V}}} \sum_{j\in\mathcal{V}} \alpha_{tj} \left| d_t^\top v_j - \mu_{\mathcal{V},t} \right|.$$

Then

$$\rho_t \triangleq \frac{\sum_{j\in\mathcal{T}} \alpha_{tj} |d_t^\top v_j - \bar{s}_t|}{\sum_{j\in\mathcal{V}} \alpha_{tj} |d_t^\top v_j - \bar{s}_t|} \geq \gamma_t, \tag{25}$$

where

$$\gamma_t = \frac{\alpha_t^{\mathcal{T}} (\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t})}{\sigma_{\mathcal{V},t} + \alpha_t^{\mathcal{T}} (\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t})}. \tag{26}$$

Proof. Since $t \in \mathcal{H}$,

$$\mu_{\mathcal{T},t} \geq \mu_{\mathcal{V},t} + \delta_t.$$

Using

$$\bar{s}_t = \alpha_t^{\mathcal{V}} \mu_{\mathcal{V},t} + \alpha_t^{\mathcal{T}} \mu_{\mathcal{T},t},$$

we have

$$\mu_{\mathcal{T},t} - \bar{s}_t = \alpha_t^{\mathcal{V}} (\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t}), \qquad \bar{s}_t - \mu_{\mathcal{V},t} = \alpha_t^{\mathcal{T}} (\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t}).$$

By Jensen's inequality,

$$\begin{aligned} \sum_{j\in\mathcal{T}} \alpha_{tj} |d_t^\top v_j - \bar{s}_t| &\geq \left| \sum_{j\in\mathcal{T}} \alpha_{tj} (d_t^\top v_j - \bar{s}_t) \right| \\ &= \alpha_t^{\mathcal{T}} \alpha_t^{\mathcal{V}} (\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t}). \end{aligned} \tag{27}$$

For the visual side, the triangle inequality gives

$$\sum_{j\in\mathcal{V}} \alpha_{tj}|d_t^\top v_j - \bar{s}_t| \le \sum_{j\in\mathcal{V}} \alpha_{tj}|d_t^\top v_j - \mu_{\mathcal{V},t}| + \sum_{j\in\mathcal{V}} \alpha_{tj}|\mu_{\mathcal{V},t} - \bar{s}_t|$$
$$= \alpha_t^{\mathcal{V}}\sigma_{\mathcal{V},t} + \alpha_t^{\mathcal{V}}\alpha_t^{\mathcal{T}}(\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t}). \tag{28}$$

Dividing (27) by (28) yields (25). □

Lemma B.2 shows that a signed non-visual gap yields a nonzero lower bound on the centered text-to-vision activation ratio. The bound is intentionally conservative; the next theorem uses the actual centered ratio ρ_t as the sufficient condition.

Theorem B.3 (Gradient-budget bias under uniform credit)**.** *Consider a positive-advantage local update with uniform credit $\hat{A} > 0$. Under Assumptions B.2 and B.3, if a step t satisfies*

$$\rho_t > \frac{C_J}{c_{\mathcal{T}} c_J}, \tag{29}$$

then the non-visual gradient contribution dominates the visual contribution:

$$\frac{\|G_t^{\mathcal{T}}\|_F}{\|G_t^{\mathcal{V}}\|_F} \ge \frac{c_{\mathcal{T}} c_J}{C_J}\rho_t > 1. \tag{30}$$

Consequently, for any subset $\mathcal{H}_\rho$ of steps satisfying (29)*,*

$$\sum_{t\in\mathcal{H}_\rho} \|\hat{A}G_t^{\mathcal{T}}\|_F \ge \underline{\kappa} \sum_{t\in\mathcal{H}_\rho} \|\hat{A}G_t^{\mathcal{V}}\|_F, \qquad \underline{\kappa} = \min_{t\in\mathcal{H}_\rho} \frac{c_{\mathcal{T}} c_J}{C_J}\rho_t > 1. \tag{31}$$

Proof. By Assumption B.2,

$$\|G_t^{\mathcal{T}}\|_F \ge c_{\mathcal{T}} \sum_{j\in\mathcal{T}} \alpha_{tj}|d_t^\top v_j - \bar{s}_t| \left\| \frac{\partial e_{tj}}{\partial W_Q} \right\|_F. \tag{32}$$

By the triangle inequality,

$$\|G_t^{\mathcal{V}}\|_F \le \sum_{j\in\mathcal{V}} \alpha_{tj}|d_t^\top v_j - \bar{s}_t| \left\| \frac{\partial e_{tj}}{\partial W_Q} \right\|_F. \tag{33}$$

Using Assumption B.3, the numerator in (32) is lower-bounded by $c_{\mathcal{T}} c_J J_t \sum_{j\in\mathcal{T}} \alpha_{tj}|d_t^\top v_j - \bar{s}_t|$, while the denominator in (33) is upper-bounded by $C_J J_t \sum_{j\in\mathcal{V}} \alpha_{tj}|d_t^\top v_j - \bar{s}_t|$. Therefore

$$\frac{\|G_t^{\mathcal{T}}\|_F}{\|G_t^{\mathcal{V}}\|_F} \ge \frac{c_{\mathcal{T}} c_J}{C_J}\rho_t.$$

Condition (29) gives the strict dominance. Multiplying by the same positive advantage $\hat{A}$ and summing over $\mathcal{H}_\rho$ gives (31). □

Remark 1 (Interpretation)**.** *Theorem B.3 is a sufficient-condition result. It does not require the conservative lower bound γ_t itself to exceed the threshold. Instead, γ_t explains how a signed non-visual gap contributes to the centered ratio ρ_t, while the actual gradient-budget dominance is controlled by ρ_t together with cancellation and Jacobian-scale regularity.*

Theorem B.4 (Local visual-attention drift)**.** *Consider a gradient-descent update on* (13) *with $\hat{A} > 0$:*

$$\Delta W_Q = -\eta \nabla_{W_Q} \mathcal{L} = \eta \hat{A} \sum_{t'=1}^{T} G_{t'}. \tag{34}$$

Define the same-step diagonal drift coefficient

$$D_t \triangleq -\left\langle \frac{\partial \alpha_t^{\mathcal{V}}}{\partial W_Q}, G_t \right\rangle_F, \tag{35}$$

and the cross-step remainder

$$R_t \triangleq \sum_{t'\neq t} \left\langle \frac{\partial \alpha_t^{\mathcal{V}}}{\partial W_Q}, G_{t'} \right\rangle_F. \tag{36}$$

Then the first-order change in visual attention mass is

$$\Delta\alpha_t^{\mathcal{V}} = -\eta\hat{A}D_t + \eta\hat{A}R_t + O(\|\Delta W_Q\|_F^2). \tag{37}$$

If $t \in \mathcal{H}$ and Assumption B.4 holds, then

$$D_t \ge \zeta C_t \alpha_t^{\mathcal{V}} \alpha_t^{\mathcal{T}} (\mu_{\mathcal{T},t} - \mu_{\mathcal{V},t}) > 0. \tag{38}$$

Therefore, for positive-advantage trajectories, visual attention mass locally decreases whenever

$$|R_t| < D_t. \tag{39}$$

Proof. A first-order Taylor expansion gives

$$\Delta\alpha_t^{\mathcal{V}} = \left\langle \frac{\partial \alpha_t^{\mathcal{V}}}{\partial W_Q}, \Delta W_Q \right\rangle_F + O(\|\Delta W_Q\|_F^2).$$

Substituting (34) and separating the $t' = t$ term from the $t' \neq t$ terms yields

$$\Delta\alpha_t^{\mathcal{V}} = \eta\hat{A}\left\langle \frac{\partial \alpha_t^{\mathcal{V}}}{\partial W_Q}, G_t \right\rangle_F + \eta\hat{A}\sum_{t'\neq t}\left\langle \frac{\partial \alpha_t^{\mathcal{V}}}{\partial W_Q}, G_{t'} \right\rangle_F + O(\|\Delta W_Q\|_F^2).$$

Using the definitions of D_t and R_t gives (37). For $t \in \mathcal{H}$, Assumption B.4 gives (38). If $|R_t| < D_t$, the first-order term in (37) is strictly negative, so $\alpha_t^{\mathcal{V}}$ decreases for sufficiently small effective update size. □

Remark 2 (Scope of the drift result)**.** *Theorem B.4 is a local first-order statement. It does not require the cross-step term R_t to vanish; it only states a sufficient regime in which the same-step diagonal drift dominates cross-step interference. For $\hat{A} < 0$, the first-order sign reverses. In group-relative RLVR, the batch-level drift depends on the correlation between trajectory advantages and the local coefficients D_t across sampled rollouts.*

Remark 3 (Second-order term)**.** *The remainder in* (37) *is controlled by the effective parameter displacement $\|\Delta W_Q\|_F$. Thus the result should be interpreted in the standard local-update regime of policy-gradient analysis. We do not claim that the first-order expression characterizes large-step or long-horizon training dynamics globally.*

B.5 Effect of HiCCR Reweighting

HiCCR replaces the uniform positive-trajectory advantage by

$$\widetilde{A}_t^{(i)} = \alpha^{(i)}\beta_t^{(i)}\hat{A}^{(i)}, \qquad \hat{A}^{(i)} > 0,$$

while leaving non-positive-advantage trajectories unmodified. For a positive-advantage trajectory, the same first-order expansion gives

$$\Delta\alpha_t^{\mathcal{V}}\big|_{\mathrm{HiCCR}} = -\eta\alpha^{(i)}\beta_t\hat{A}D_t + \eta\alpha^{(i)}\hat{A}R_t^{(\beta)} + O(\|\Delta W_Q\|_F^2), \tag{40}$$

where

$$R_t^{(\beta)} = \sum_{t'\neq t}\beta_{t'}\left\langle \frac{\partial \alpha_t^{\mathcal{V}}}{\partial W_Q}, G_{t'} \right\rangle_F. \tag{41}$$

Thus, at the diagonal level, HiCCR changes the aggregate within-trajectory coefficient from

$$S_1 = \sum_t D_t \qquad \text{to} \qquad S_\beta = \sum_t \beta_t D_t. \tag{42}$$

Uniform credit corresponds to $\beta_t = 1$.

Define

$$\mathcal{N} = \{t : D_t > 0\}, \qquad \mathcal{P} = \{t : D_t < 0\}.$$

Steps in $\mathcal{N}$ contribute to decreasing visual attention under positive advantage, whereas steps in $\mathcal{P}$ contribute in the opposite direction. Under the diagonal approximation, HiCCR reverses the aggregate visual-attention drift if

$$\sum_{t\in\mathcal{P}}\beta_t|D_t| \geq \sum_{t\in\mathcal{N}}\beta_t D_t. \tag{43}$$

It mitigates the magnitude of a negative uniform drift whenever

$$0 \leq S_\beta < S_1. \tag{44}$$

HiCCR is designed to make such mitigation more likely by assigning larger β_t to tokens with stronger visual attention concentration. This proxy does not assert that attention concentration is a causal saliency measure for every token; rather, it provides an efficient operational signal for redistributing positive credit toward visually engaged generation steps. The trajectory-level factor $\alpha^{(i)}$ is constant within a trajectory and therefore does not change the sign of S_β for that trajectory. Its role is batch-level: among positive-advantage rollouts, it increases the contribution of trajectories with stronger overall visual engagement.

Remark 4 (Empirical justification of the β_t proxy)**.** *The analysis in* (42)–(44) *shows that any reweighting that assigns larger β_t to steps with larger positive D_t (or smaller negative D_t) will reduce S_β relative to the uniform-credit baseline S_1. It does not uniquely determine the functional form of β_t. HiCCR instantiates β_t via visual attention concentration $C_{\mathcal{V}}(t)$ as a cheap, differentiable proxy available at every rollout step without additional forward passes. This design choice is not a consequence of the theorem but an empirical one: Section 3 (Table 2) compares the attention-based β_t against random, uniform, and gradient-magnitude alternatives under identical $\alpha^{(i)}$, and shows that the attention-based proxy yields the strongest downstream gains. Thus the role of the theoretical analysis is to identify* the direction of a useful reweighting*, while the particular proxy is selected by ablation.*

Remark 5 (Positive-advantage asymmetry)**.** *The above correction applies only to $\hat{A} > 0$, matching the design of HiCCR. Applying the same amplification to negative-advantage trajectories would also amplify penalties on visually engaged but incorrect rollouts, which can discourage visual use. This is why HiCCR reassigns credit asymmetrically: it amplifies visually engaged successful trajectories while leaving unsuccessful trajectories under the base RLVR signal. The related experiments have confirm this conclusion 12.*

B.6 Multi-Head and Multilayer Scope

The preceding analysis is a single-head, single-layer local component analysis. In a full transformer, attention-gradient contributions add across heads and layers but are coupled through output projections, residual streams, and subsequent nonlinear blocks. Therefore, we do not claim a headwise theorem automatically becomes a full-network theorem without additional assumptions.

The mechanism nevertheless extends qualitatively under limited cross-head and cross-layer cancellation. If a non-vanishing subset of heads/layers satisfies the centered-dominance and local key-metric conditions, their contributions create a text-favored component in the aggregate update. Conversely, heads/layers that do not satisfy these conditions can weaken or cancel the effect. This is why the empirical analysis of κ_{emp} across layers is used as support for where the mechanism is likely active, rather than as a proof that every layer satisfies the same condition.

In practice, HiCCR computes its context-credit score from late decoder layers, where the measured non-visual activation dominance remains consistently above one and where token predictions are closer to the final policy decision. This motivates the choice of the target layer set $\mathcal{L}^\star$ for computing $C_{\mathcal{V}}(t)$ in Section 2.

C On the Validity of the Context Credit Score

This section clarifies why $C_{\mathcal{V}}(t)$ is used as the token-level credit signal in HiCCR. We do not claim that it is a gradient-based saliency score or a token-level causal explanation. Instead, we use it as an operational measure of visual-attention concentration and evaluate whether it provides a useful signal for visually grounded credit reassignment.

C.1 Definition

Recall that

$$C_{\mathcal{V}}(t) = 1 - \frac{H\big(p_{\mathcal{V}}(\cdot \mid t)\big)}{\log |\mathcal{V}|}, \qquad p_{\mathcal{V}}(j \mid t) \propto \exp(e_{tj}/\tau), \;\; j \in \mathcal{V}. \tag{45}$$

Here $p_{\mathcal{V}}(\cdot \mid t)$ is the attention distribution from generation token t to visual positions. Thus, $C_{\mathcal{V}}(t)$ is high when a token selectively attends to a small subset of visual tokens, and low when its visual attention is diffuse. We interpret this score as a lightweight proxy for selective visual querying, rather than as a direct measure of causal saliency.

C.2 Ablation Evidence

Table 2 tests whether the proposed score is more useful than alternative credit signals under the same HiCCR pipeline. We compare Uniform Allocation, Random Allocation, Time-decay, Entropy-mask, Text-Context-based HiCCR-T, Raw Attention Energy, and the proposed Context-based HiCCR.

The results support three observations. First, Random Allocation improves over Uniform Allocation, suggesting that uniform token credit is overly coarse. Second, Time-decay, Entropy-mask, and HiCCR-T are weaker or unstable, indicating that generic non-uniformity, hard token selection, and textual-context concentration are insufficient. Third, Raw Attention Energy improves over several heuristics but remains below $C_{\mathcal{V}}(t)$, showing that raw visual-attention magnitude is less effective than normalized concentration within the visual token subset.

These comparisons support a limited conclusion: among the tested alternatives, visual-attention concentration provides the most effective credit signal. They do not imply that $C_{\mathcal{V}}(t)$ perfectly measures token-level visual dependence.

C.3 Semantic and Counterfactual Evidence

Beyond credit-score ablations, we further examine whether the learned credit weights align with visually related tokens. As shown in Figure 4(e), semantically visual tokens receive higher credit weights than non-visual tokens, with a median weight of 1.303 versus 1.236 ($p < 0.001$, Cohen's $d = 0.547$). **This result suggests that the proposed credit score is not merely assigning arbitrary non-uniform weights, but tends to amplify tokens that are semantically associated with visual reasoning**. We emphasize that this is a semantic alignment test rather than a token-level causal proof.

We also conduct a model-level visual ablation in Appendix D.2. During inference, we zero out the visual embeddings and measure the resulting accuracy drop across benchmarks. As shown in Figure 7, HiCCR variants show larger degradation than their corresponding uniform-credit RLVR baselines across both 4B and 8B scales. This suggests that HiCCR's gains are accompanied by stronger functional reliance on visual input at the model level.

Together, the credit-score ablations, semantic token analysis, and visual counterfactual results support $C_{\mathcal{V}}(t)$ as a practical proxy for visually grounded credit reassignment. They do not require the stronger claim that attention concentration is equivalent to gradient-based saliency or token-level causal dependence.

D Mechanistic Interpretability: From Attention to Functional Dependence

A fundamental critique of attention-guided methods is the well-known "attention is not explanation" debate [25]. To rigorously verify whether HiCCR genuinely alters the model's functional dependence on visual inputs—rather than merely manipulating superficial attention patterns—we conduct a three-stage mechanistic and causal analysis with increasing evidential strength: (i) **gradient-based attribution** quantifies how much each modality contributes to the policy gradient (§D.1); (ii) **counterfactual ablation** removes visual input entirely and measures performance collapse, establishing a causal link (§D.2); and (iii) **attention dynamics** traces the temporal evolution of visual attention across generation steps, revealing how different training objectives reshape the model's reliance on image tokens over the course of a rollout (§D.3).

D.1 Gradient-Based Attribution Analysis

We ask: *Does the preservation of visual attention in HiCCR translate to actual gradient contribution?* Using the Input×Gradient method [53], we compute, for each generated token, the fraction of the gradient's L2 norm derived from visual embeddings relative to the entire context, then average over all tokens and prompts per benchmark.

As shown in Table 14, uniform RLVR baselines consistently *diminish* visual gradient attribution relative to the Base model (e.g., DAPO reduces the 4B average from 25.0% to 22.7%, and the 8B average from 22.0% to 20.8%), confirming the gradient-budget bias predicted by Theorem B.3. HiCCR reverses this degradation: HiCCR-G restores the average to 25.8% (4B) and 26.1% (8B), with the largest gains on geometry- and diagram-intensive benchmarks where continuous visual extraction is indispensable—Geo3K improves by +5.3 pp (4B) and +6.1 pp (8B) over DAPO, and WeMath by +4.0 pp and +6.6 pp respectively. Notably, the attribution gains are consistent across both HiCCR variants and both model scales, confirming that the credit reassignment mechanism reliably redirects optimization pressure toward the visual modality.

Table 14: **Gradient Attribution Analysis.** Image Attr. (%): average fraction of total Input×Gradient attribution originating from visual tokens. Best results compared to respective uniform baselines are in **bold**.

Method	Geo3K	MathV	MVista	WeMath	MMK12	CLEVR	Logic	MMMU	Hallus.	MMStar	Avg
Qwen3-VL-4B Backbone											
Base	18.7	21.6	21.2	16.1	13.5	**28.5**	26.0	53.0	**32.3**	18.8	25.0
GRPO	17.0	20.3	20.9	15.0	12.8	25.8	25.8	52.5	31.8	17.4	23.9
DAPO	16.7	19.6	20.6	14.2	13.5	27.5	24.1	46.6	28.5	15.8	22.7
HiCCR-G	22.0	**23.4**	**22.0**	**18.2**	**15.3**	27.9	26.9	**56.1**	27.1	**19.5**	**25.8**
HiCCR-D	**22.6**	23.3	21.8	17.5	14.5	27.5	**27.6**	**56.2**	31.6	18.4	**26.1**
Qwen3-VL-8B Backbone											
Base	17.5	20.8	18.9	13.6	11.8	21.4	23.8	52.4	25.3	14.9	22.0
GRPO	20.4	22.2	20.3	17.0	13.3	21.4	28.0	55.6	29.4	14.6	24.2
DAPO	15.8	18.6	17.8	13.7	11.2	19.8	23.4	49.2	24.9	14.1	20.8
HiCCR-G	**21.9**	**22.8**	**22.3**	**20.3**	**14.7**	**25.3**	**30.5**	**55.7**	**31.0**	**16.7**	**26.1**
HiCCR-D	17.2	17.9	18.9	15.4	13.3	24.8	26.7	46.2	29.7	14.2	**22.4**

D.2 Counterfactual Visual Ablation

Gradient attribution reveals how the model *uses* visual information, but does not establish whether visual input is *necessary*. To establish stronger causal evidence between our credit reassignment and functional visual dependence, we perform a counterfactual visual ablation study. During inference, we corrupt the visual embeddings by applying *Zero-Ablation* (zeroing out visual tokens), forcing the model to rely entirely on its textual priors and previously generated context. The rationale is straightforward: if a model genuinely grounds its reasoning in the visual input, destroying that input should result in a catastrophic performance drop. Conversely, if a model relies on textual shortcuts or hallucination, the performance drop will be minimal (or even negative).

As illustrated in Figure 7, while most models suffer to some extent when visual evidence is destroyed, **HiCCR variants (HiCCR-G and HiCCR-D) exhibit a steeper accuracy drop (Δ) compared to both the Base models and uniform baselines across both 4B and 8B scales**.

This disparity is most striking in highly complex benchmarks. For instance, on MMMU_Pro, the Base-4B model paradoxically shows a negative drop (i.e., performance slightly improves when the image is removed), exposing a severe reliance on textual hallucination rather than visual comprehension. In contrast, HiCCR models consistently suffer the largest accuracy penalties when blinded, strongly suggesting that their superior performance stems from functional visual grounding. This profound vulnerability to visual corruption provides stronger causal evidence: HiCCR's optimization explicitly forces the policy to ground its reasoning in the actual image.

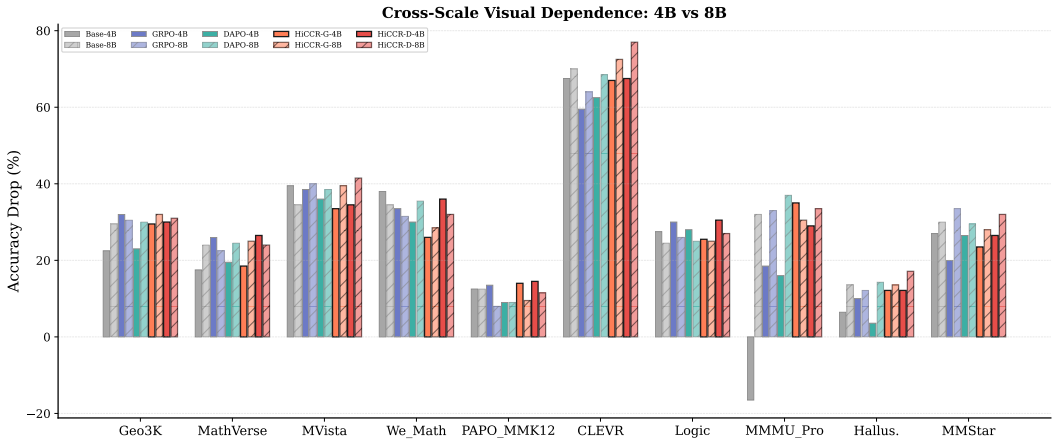


Figure 7: **Cross-Scale Counterfactual Visual Ablation.** We report the accuracy drop (%) when visual embeddings are zeroed out during inference. A larger drop indicates stronger functional dependence on visual inputs. Across both 4B (solid) and 8B (hatched) scales, HiCCR variants consistently exhibit the most severe performance degradation upon visual corruption, confirming that our method substantially enforces stronger visual functional dependence.

Together with the gradient evidence in §D.1, this ablation establishes both a correlational and a causal link between HiCCR's credit-reassignment mechanism and enhanced visual functional dependence.

D.3 Attention Dynamics During Generation

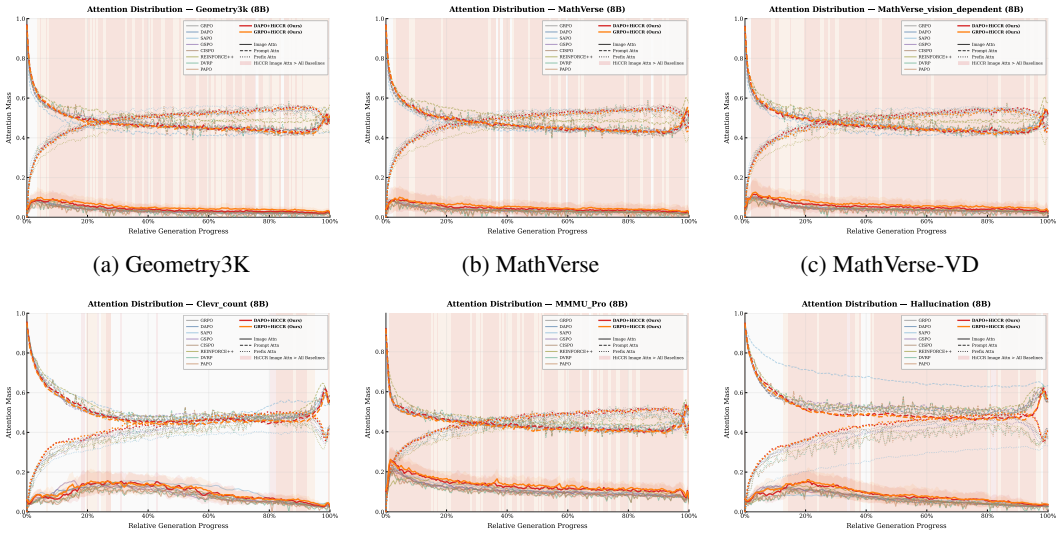
We examine how visual attention evolves during generation. As predicted by Theorem B.4, uniform credit assignment can progressively shift policy updates away from image tokens, leading to attention drift. This subsection provides temporal evidence for this drift and shows that HiCCR counteracts it by reallocating credit toward visually engaged tokens and trajectories.

Protocol. For each model and method, we sample 500 instances and track the normalized attention mass at every decoding step over three context components: image tokens, prompt tokens, and prefix tokens. All trajectories are interpolated onto a uniform 200-point grid and visualized with mean and standard deviation bands. We evaluate six representative benchmarks covering mathematical reasoning, vision-dependent reasoning, and general multimodal understanding: Geometry3K, MathVerse, MathVerse-VD, CLEVR-Count, MMMU-Pro, and HallusionBench. Results are reported under four settings: RLVR baseline comparisons at 8B and 4B, and base-model comparisons at 8B and 4B. In all figures, solid lines denote image attention, dashed lines denote prompt attention, and dotted lines denote prefix attention; shaded regions mark intervals where HiCCR image attention exceeds all compared baselines.

Main findings. Three findings are consistent across benchmarks and model scales. First, the pretrained model already shows decreasing image attention as generation proceeds, indicating that visual under-utilization is an inherited tendency of the backbone. Second, standard RLVR methods such as GRPO and DAPO amplify this tendency: they improve answer accuracy but further reduce sustained attention to image tokens and increase reliance on generated prefix tokens. Third, HiCCR reverses this pattern. Across all settings, HiCCR maintains higher image attention than GRPO, DAPO, and other RLVR baselines, often forming a stable visual attention plateau throughout generation. The gain mainly comes from reduced prefix reliance while prompt attention remains comparable, suggesting that HiCCR encourages the model to repeatedly consult visual evidence rather than reinforce its own textual continuation.

D.3.1 RLVR Baselines at 8B

Figure 8 compares HiCCR with nine RLVR baselines on Qwen3-VL-8B. Across all six benchmarks, GRPO, DAPO, and related RLVR methods show a clear decline in image attention from early to late generation. In contrast, HiCCR variants sustain higher image attention and exceed all baselines over long decoding intervals. The advantage is strongest on vision-dependent tasks such as MathVerse-VD and CLEVR-Count, where correct reasoning requires repeated extraction of visual evidence. Prompt attention remains comparable across methods, while prefix attention is lower under HiCCR. This indicates that HiCCR improves visual grounding not by weakening prompt processing, but by reducing self-reinforcing dependence on generated text.



(a) Geometry3K (b) MathVerse (c) MathVerse-VD

(d) CLEVR-Count (e) MMMU-Pro (f) HallusionBench

Figure 8: **8B RLVR attention dynamics.** HiCCR sustains higher image attention across benchmarks, while GRPO, DAPO, and related RLVR baselines exhibit progressive visual-attention drift.

D.3.2 RLVR Baselines at 4B

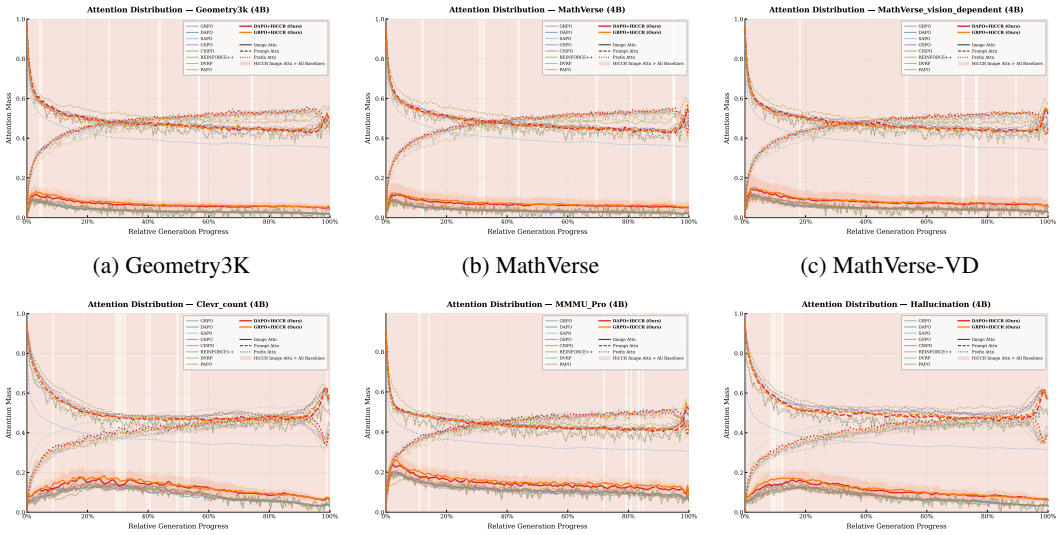
We repeat the comparison on Qwen3-VL-4B to test whether the same dynamics hold under tighter capacity constraints. Figure 9 shows that attention drift becomes more severe at the smaller scale. The 4B RLVR baselines lose image attention more rapidly than their 8B counterparts, suggesting that limited capacity makes the model more vulnerable to language-prior amplification during RLVR. Despite this stronger drift, HiCCR again maintains a persistent visual attention plateau and surpasses the RLVR baselines across benchmarks. This consistency indicates that sustained visual grounding is not a byproduct of model capacity, but a direct effect of credit reassignment.

D.3.3 Base Model Comparisons at 8B

To isolate the effect of HiCCR from the pretrained model's intrinsic attention pattern, we compare HiCCR with the original Qwen3-VL-8B-Instruct model, GRPO, and DAPO. Figure 10 shows that the base model already reduces image attention as generation progresses. GRPO and DAPO further intensify this tendency, confirming that standard RLVR amplifies the attention drift inherited from pretraining. HiCCR substantially reshapes the trajectory by raising image attention above both the base model and RLVR baselines. Since prompt attention remains comparable, the improvement reflects targeted visual grounding rather than a general redistribution away from text.

D.3.4 Base Model Comparisons at 4B

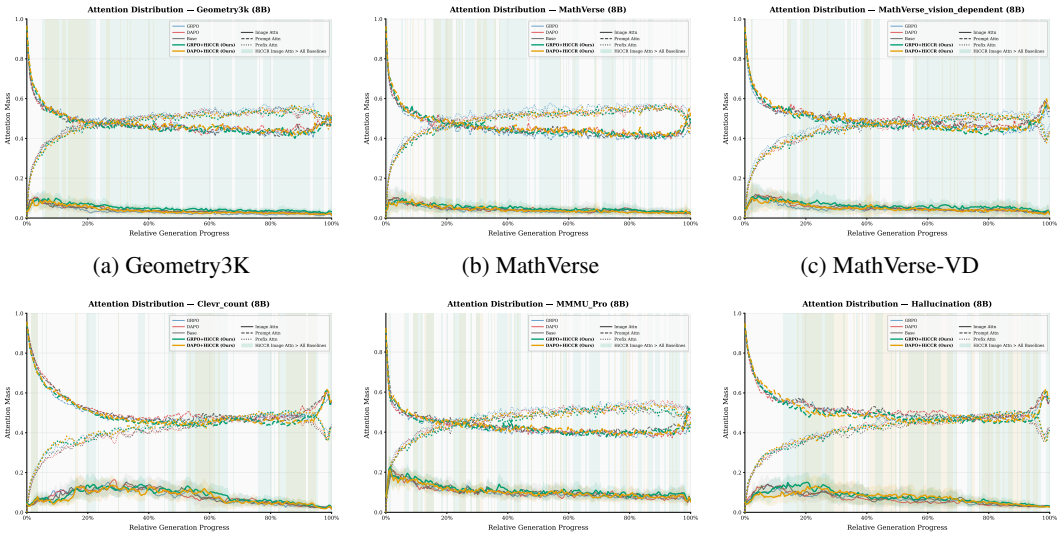
We further compare against the Qwen3-VL-4B-Instruct base model. As shown in Figure 11, the smaller base model exhibits a sharper decline in image attention than the 8B base model, often approaching very low visual attention in later decoding steps. GRPO and DAPO do not correct this collapse and instead reinforce the same drift pattern. HiCCR effectively rescues visual grounding at this scale, keeping image attention consistently higher throughout generation while maintaining balanced prompt and prefix attention. This result shows that visual credit reassignment is valuable when standard RLVR is more likely to reinforce language-prior shortcuts.



(a) Geometry3K (b) MathVerse (c) MathVerse-VD

(d) CLEVR-Count (e) MMMU-Pro (f) HallusionBench

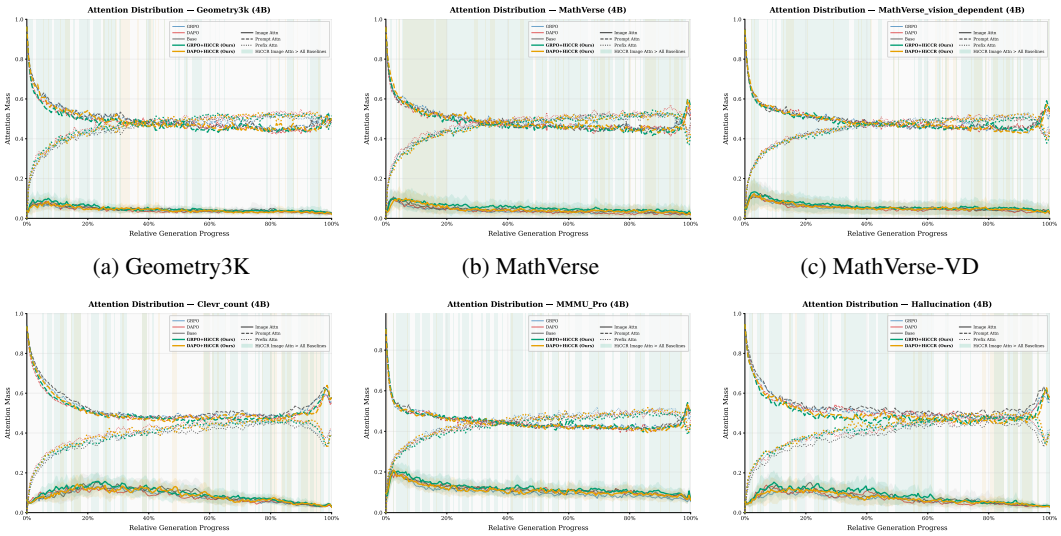
Figure 9: **4B RLVR attention dynamics.** Attention drift becomes stronger at the smaller scale, but HiCCR still preserves a stable visual attention plateau beyond all RLVR baselines.



(a) Geometry3K (b) MathVerse (c) MathVerse-VD

(d) CLEVR-Count (e) MMMU-Pro (f) HallusionBench

Figure 10: **8B base-model attention comparison.** The pretrained base model already shows visual-attention decay, GRPO and DAPO amplify it, and HiCCR reverses the trend.



(a) Geometry3K (b) MathVerse (c) MathVerse-VD

(d) CLEVR-Count (e) MMMU-Pro (f) HallusionBench

Figure 11: **4B base-model attention comparison.** The 4B base model suffers stronger visual-attention collapse, which GRPO and DAPO further reinforce while HiCCR effectively restores sustained image attention.

D.4 Summary of Findings

Taken together, these results provide converging evidence, from correlational to causal, that HiCCR's hierarchical credit reassignment induces functional visual grounding rather than superficial attention patterns. The gradient attribution confirms where the learning signal flows (§D.1); the counterfactual ablation confirms that visual input is functionally necessary for the learned policy (§D.2); and the attention dynamics reveal the temporal mechanism through which HiCCR counteracts the drift predicted by Theorem B.4 (§D.3).

E Related Work

E.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) extend the reasoning and instruction-following capabilities of LLMs to visual inputs, typically by aligning a vision encoder with a language backbone through projection layers or cross-attention modules [32, 57, 8, 2, 16]. Recent work has scaled MLLMs along both the model and data axes. LLaVA [32] demonstrated that a simple linear projection suffices for effective visual instruction tuning, while subsequent models such as InternVL-2.5 [8] and Qwen2-VL [57] advanced state-of-the-art performance through improved training recipes, dynamic resolution strategies, and curated multimodal corpora. On the reasoning front, efforts like Insight-V [12] and Vision-R1 [24] explore scalable approaches for generating long-chain reasoning data and applying reinforcement learning to multimodal settings, respectively.

E.2 Reinforcement Learning with Verifiable Rewards

Reinforcement learning from human feedback (RLHF) [45, 52] has become the dominant paradigm for aligning LLMs with human preferences, yet it relies on a learned reward model that is expensive to train and prone to reward hacking. Reinforcement learning with verifiable rewards (RLVR) offers a compelling alternative for tasks where correctness can be automatically verified—such as mathematics, coding, and visual question answering—by replacing the reward model with a deterministic verifier that provides binary (correct/incorrect) feedback [20, 29]. DeepSeek-R1 [20] demonstrated that RLVR alone can elicit sophisticated chain-of-thought reasoning in large-scale models. On the algorithmic side, GRPO [49] eliminates the critic network by estimating baselines from group-sampled trajectories, substantially reducing memory overhead. DAPO [69] further improves training stability through clip-higher, dynamic sampling, and token-level loss normalization. Recent studies have begun probing the boundaries and failure modes of RLVR: [72] show that RLVR gains largely stem from amplifying capabilities already present after supervised fine-tuning, and that performance saturates without sufficiently diverse training distributions [54]. In the multimodal domain, R1-V [6] and Vision-R1 [24] apply RLVR to MLLMs but retain the uniform per-token advantage formulation, leaving the credit assignment

problem unaddressed. Despite these advances, how to effectively reinforce *visual grounding* during post-training remains understudied [34]. Most RLVR pipelines treat visual and textual tokens uniformly [39, 61], overlooking the unique role of visual context in grounding the reasoning process—a gap that motivates our work.

E.3 Token-Level Credit Assignment in RLVR

The credit assignment problem—determining which tokens in a generated sequence are responsible for the final reward—is central to efficient policy optimization. Standard RLVR algorithms such as GRPO [49] dispense with the critic model and broadcast the trajectory-level advantage uniformly to every token, implicitly assuming equal contribution. This assumption is particularly problematic for vision-centric understanding and long-form reasoning, where pivotal tokens are grounded on visual content or only a sparse subset actually drives the outcome [41, 15]. A growing body of work seeks finer-grained credit signals; we organize these efforts by modality setting.

Credit assignment in text-only RLVR. One line of work introduces critic-free temporal-difference corrections. GRPO-λ [46] propagates credit backward via an eligibility trace reparameterised from the GAE λ-return, while TEMPO [55] derives nonparametric token-level TD corrections from the implicit prefix tree formed by grouped rollouts. A second line leverages entropy for reward shaping: GTPO [51] redistributes trajectory reward proportionally to per-token policy entropy, treating it as a proxy for cognitive effort, and Entropy-RL [10] provides a theoretical account linking entropy dynamics to the probability-advantage covariance. A third category employs external models for dense credit—Grad2Reward [75] extracts token-level rewards from a judge model's gradient attribution, and CAPO [66] uses an LLM-as-GenPRM to generate step-wise critiques with asymmetric outcome–process reward shaping. A fourth line learns explicit value functions: Q-RM [5] optimizes a discriminative Q-function over preference data for token-level value estimates, and PURE [9] replaces the canonical summation-form credit with a min-form value function $V(s_t) = \min_{k \geq t} r_k$ to alleviate PRM-induced reward hacking. All the above are designed for text-only LLMs and do not account for the interplay between generated tokens and *visual context*.

Credit assignment in multimodal RLVR. Several recent efforts extend token-level credit to the multimodal setting. PEPO [31], AT-RL [26], and VGPO [62] derive visual grounding signals from hidden-state similarity between response and image tokens. These approaches operate in the *representation* space and rely on the assumption that high cosine similarity implies visual dependence—an assumption that is not formally justified. In contrast, HiCCR measures the *concentration* of the attention distribution over visual positions, which directly reflects how selectively the model extracts visual information during generation. A second family identifies perception-critical tokens via counterfactual perturbation: VPPO [23], ToR-GRPO [36], and PGPO [68] compare next-token distributions with and without visual input to score per-token visual dependency, while Video-KTR [64] extends this to the video domain with masked and frame-shuffled branches. All such counterfactual methods require one or more additional forward passes, incurring non-trivial overhead. TW-GRPO [11] avoids perturbation by measuring intra-group distributional divergence at each token position as a proxy for information density. APPO [13] is also based on intra-group similarities, by calculating attention-to-frame scores to target perception tokens and reinforce them with more credits.

Despite recent progress, existing multimodal credit-assignment methods largely leave the training-time evolution of visual grounding under RLVR underexplored. In particular, they do not diagnose the grounding drift we identify: standard uniform-credit RLVR can reduce image attention and the gradient share routed to visual tokens even while improving task accuracy, making it unclear whether gains come from stronger visual use or greater exploitation of textual/reasoning priors. HiCCR is designed around this diagnosis: it estimates per-token visual engagement through the information concentration of visual attention, derives both token- and trajectory-level credit weights from this lightweight forward-pass score, and pairs the mechanism with attention, gradient, and visual counterfactual diagnostics to verify that accuracy gains are accompanied by mitigated grounding drift.

F Additions about Training and Evaluation

F.1 Experiments Details

Implementation framework. All experiments are built on veRL [50] with PyTorch 2.8.0 [47] and CUDA 12.8, and are conducted on $8\times$ High-Performance GPUs. We fine-tune the Qwen3-VL-4B and 8B backbones [2] following the RLVR paradigm of DeepSeek-R1 [20].

Training details & computational efficiency. We integrate HiCCR with two representative RLVR algorithms—GRPO [49] and DAPO [69]—yielding **HiCCR-G** and **HiCCR-D**, respectively. Key training hyperparameters are shared across all runs: G=5 rollouts per prompt, temperature 1.0, maximum generation length 4096, learning rate 2×10^{-6} with AdamW [35] and cosine decay, and PPO clip ratio ϵ=0.2. For HiCCR-specific settings, we fix λ_{tok}=0.5, τ=0.5, and select the last two transformer layers as $\mathcal{L}_{\text{target}}$ throughout. Ablation on the target layer has been presented in Table 11.

Implementation. Eq. (6) does *not* require materializing the full $T \times T$ attention matrix. We register lightweight forward hooks on the query and key projection modules of the target layers, cache their detached outputs, and compute the score via a two-pass chunked algorithm: (i) accumulate the global log-sum-exp over visual positions using `logaddexp`, and (ii) compute $\sum_j p \log p$ chunk-by-chunk with the cached normalizer. For Grouped Query Attention, keys are expanded to match the query head count. The resulting $\mathcal{O}(S)$ token scores are immediately offloaded to CPU, so GPU memory overhead is near zero. The entire procedure runs in a single `no_grad` forward pass before the policy update, retaining only the $[B, T]$-shaped score tensor, and adds less than 3% wall-clock time per iteration (see Table 6 for detailed profiling). Since the computation operates solely on intercepted Q/K projections, it is fully compatible with FlashAttention and other fused attention backends. Full hyperparameter configurations are provided in Appendix F.2.

Compared Multimodal Models. To contextualize HiCCR beyond controlled RLVR baselines, we compare against a broad set of proprietary, general-purpose, and domain-specific multimodal models. Proprietary baselines include GPT-4o, GPT-4o-mini, and Gemini-2.0-Flash [44]. General-purpose open models include Qwen2.5-VL-72B [3], InternVL2.5-8B/78B [8], LLaVA-OV-7B/72B and LLaVA-OV-1.5-8B [30, 1], LLaVA-Critic-R1-7B [59], and R1-OneVision-7B [67]. We further include domain-specific multimodal reasoning models, including MM-Eureka-7B and MM-Eureka-7B-CPGD [40], R1-VL-7B [6], VLAA-Thinker-7B [4], VL-Rethinker-7B [56], RACRO-7B-CRO-GRPO [18], ThinkLite-7B-VL [60], and AVAR-Thinker-7B [39]. This comparison evaluates whether HiCCR-trained Qwen3-VL models [2] remain competitive beyond matched RLVR baselines under the same benchmark suite.

Prompt Template. Evaluation shares the chat template below, applied identically to every baseline and HiCCR variant, so that differences are attributable to credit reassignment rather than prompting.

Rollout Prompt

System. You first think about the reasoning process internally, then provide the final answer. Enclose them in `<think>...</think>` and `<answer>...</answer>` respectively.
User. `<image>` {question}
Assistant. `<think>`

F.2 Hyperparams

Table 15 lists all hyperparameters used in our experiments. All runs share the same training, rollout and optimizer settings unless otherwise stated; only HiCCR-specific entries are introduced in the dedicated block. Values confirmed in the main text are kept identical here, and any unspecified defaults follow the corresponding open-source implementation of veRL [50], GRPO [49] and DAPO [69]. Values marked with † follow the default setting of the corresponding base algorithm and are kept untouched for fair comparison.

Notes on robustness to hyperparameter choices. We verified that HiCCR is insensitive to moderate perturbations of its two new hyperparameters: Table 13 shows stable performance in the range $\lambda_{\text{tok}} \in [0.5, 0.8]$, and Table 11 reports monotonic improvement as $\mathcal{L}_{\text{target}}$ moves toward the last layers. All remaining entries follow the defaults of the corresponding base algorithm and were *not* re-tuned for HiCCR, which isolates the gains to credit reassignment rather than to hyperparameter search.

F.3 Algorithm Pipeline

HiCCR integrates into mainstream group-based RLVR algorithm as a pre-processing step that transforms advantages before the standard policy update. The procedure is summarized in Alg. 1.

Design principles. HiCCR is built on four properties that jointly underpin its robustness. (a) *Information-theoretic credit signal*: the context credit score is derived from the normalized entropy of visual attention, offering a principled, shift-invariant measure of grounding that is stable across layers and training stages. (b) *Minimal overhead*: the score is obtained from the policy's own query-key representations in a single forward pass, with no trainable reward or value network and only λ_{tok} to specify (Appendix A.1). (c) *Bounded modulation*: $\alpha \in [1, 2]$ and $\beta \in (1, 1+\lambda_{\text{tok}})$ by construction, so HiCCR only *amplifies* gradients for visually engaged tokens without reversing or suppressing any signal, preserving the convergence guarantees of the base algorithm (§B.5, §3.3). (d) *Asymmetric application*: reassignment is applied to positive-advantage trajectories only, since negative advantages already serve as corrective signals and modulating them would risk penalizing exploratory rollouts that attend to the image yet reach an incorrect conclusion through otherwise valid reasoning (§3.3).

Table 15: Hyperparameter configuration for HiCCR and the baseline RLVR algorithms. Shared values apply to both HiCCR-G and HiCCR-D.

Hyperparameter	Value	Hyperparameter	Value
Model & Data		*RL Objective (shared)*	
Backbone	Qwen3-VL, InternVL3	PPO clip ratio ε	0.2
Training dataset	ViRL39K [56]	KL coefficient β_{KL}	$0.0^{\dagger}$
Max prompt length	8,192 tokens	Advantage norm.	group-relative (Eq. 1)
Max response length	4,096 tokens	Reward type	$\{0, 1\}$ verifiable
Image resolution	up to 1024^2		
Optimization		*DAPO-specific (HiCCR-D)*	
Optimizer	AdamW (β_1=.9, β_2=.999)	Clip-higher ε_{hi}	$0.28^{\dagger}$
Learning rate	2×10^{-6}	Clip-lower ε_{lo}	$0.20^{\dagger}$
LR schedule	cosine, 10 warm-up	Dynamic sampling	enabled†
Weight decay	1×10^{-2}	Overlong soft penalty	enabled†
Gradient clipping	1.0		
Training steps	198	*HiCCR-specific*	
Global batch (prompts)	256	Token weight λ_{tok}	0.5
Mini-batch (per update)	128	Softmax temperature	0.5
PPO epochs (μ)	2	Target layers $\mathcal{L}_{\text{target}}$	last 2 decoder layers
Precision	bf16	Attention heads	all (mean)
		Stability ϵ	1×10^{-8}
		Reassignment cond.	$\hat{A}^{(i)}>0$ only
Rollout & Sampling		*Hardware & Runtime*	
Rollouts per prompt (G)	5	Framework	veRL [50]
Sampling temperature	1.0	Backend	PyTorch 2.8, CUDA 12.8
Top-p / Top-k	1.0 / -1	GPUs	8× High-Performance GPUs
Rollout engine	vLLM 0.11	Attention kernel	FlashAttention-2
		Wall-clock overhead	$< 3\%$
Evaluation			
Decode temperature	1.0		
Samples per query	8 (AVG@8)		
Max gen. length	4,096		
Checkpoint choice	last training ckpt		

G Case Study

To complement the quantitative results in Section 3 and the aggregate credit analysis in Appendix A.6, we visualize four rollouts sampled from the training set (ViRL39K [56]) produced by HiCCR-G on Qwen3-VL-8B, to illustrate how the context credit weight β_t is allocated at the token level. Each figure pairs the input image with the generated rationale, where every token is color-coded by its final credit weight β_t (blue: low visual dependency; red: high visual dependency). All five rollouts yield correct predictions.

Three consistent patterns emerge across cases. **(i)** High-credit tokens concentrate on *visually grounded semantic units*—clock hands and numerical readings in Figure 12, color-shape descriptors in Figure 13, object counts and grid structure in Figures 14–15—exactly the tokens whose generation requires reading the image. **(ii)** Low-credit tokens cluster on *language-driven connectives and meta-reasoning* ("*Therefore*", "*That means*", "⟨*/thinking*⟩", arithmetic closers such as "$= 10$"), where the next token is determined by textual priors rather than visual evidence. **(iii)** Credit decays along the response: early perceptual tokens receive the strongest amplification, while later symbolic-manipulation tokens receive weights close to unity, matching the quantitative trend in Figure 4. Together, these patterns confirm that HiCCR reshapes gradient flow in a vision-aware, rather than uniform, manner—without any explicit token-level supervision.

H Limitations

HiCCR is designed for RLVR settings where verifiable outcome rewards are available and the model exposes attention statistics that can be efficiently reused during training. While our theory identifies gradient-based saliency as a principled signal for visual dependence, computing it requires an additional backward pass, so HiCCR adopts visual-attention concentration as a forward-only proxy; Table 2 and Appendix C show that this proxy outperforms raw attention energy and task-agnostic heuristics, but richer causal or semantic estimators may further improve credit reassignment. Extending the idea to open-ended rewards, video and audio inputs, or architectures with substantially different visual-token representations may therefore require adapting the credit estimator. Our experiments focus on image-based multimodal reasoning and hallucination-oriented benchmarks, with additional validation on a smaller-data setting and another MLLM family; broader evaluation on more diverse real-world multimodal tasks would further clarify the generality of the approach. Finally, our theoretical

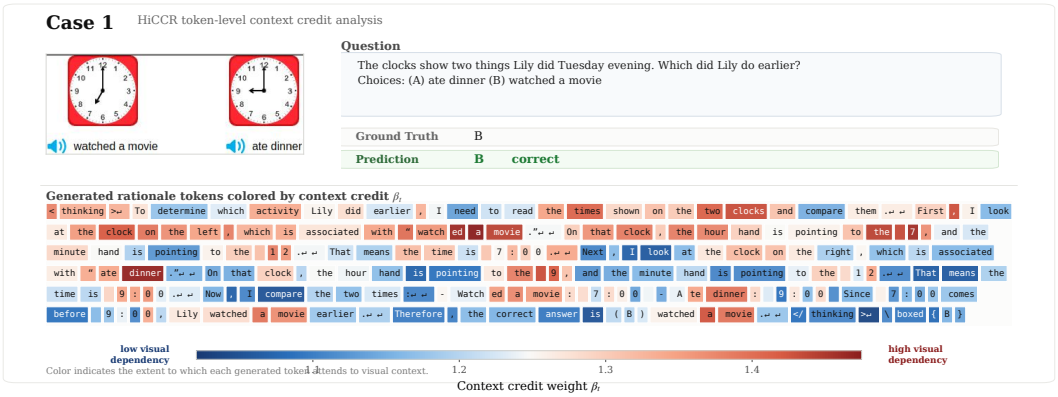


Figure 12: **Case 1: clock reading (ViRL39K training rollout).** Tokens referring to clock positions ("*7*", "*9*", "*pointing*", "*clock*", "*movie/dinner*") receive the highest credit, while temporal comparison tokens ("*before*", "*Since*", "*Therefore*") drop to near-baseline weight, showing that HiCCR amplifies visually grounded time reading rather than purely textual reasoning.

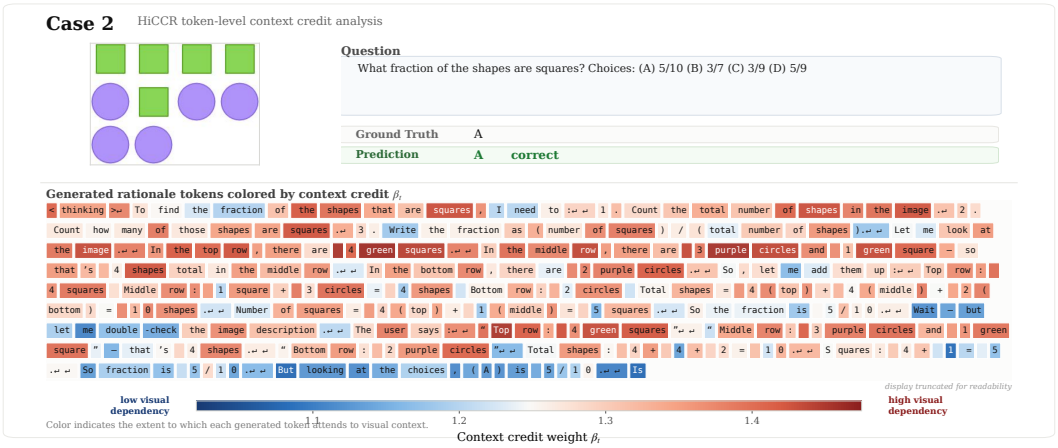


Figure 13: **Case 2: fraction of shapes (ViRL39K training rollout).** Shape- and color-descriptive tokens ("*green squares*", "*purple circles*", row-wise counts) carry the largest weights, while arithmetic tokens ("$5/10$", "*Is*") fall into the low-dependency band, indicating that credit follows perceptual evidence rather than downstream calculation.

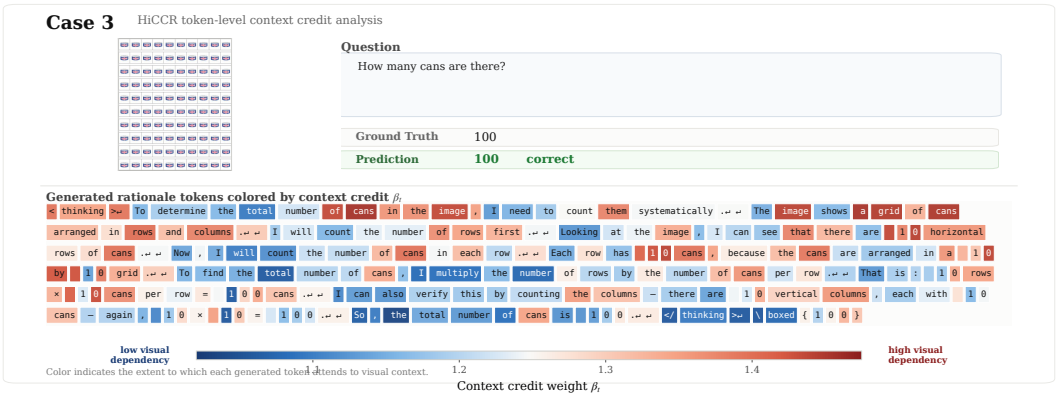


Figure 14: **Case 3: counting 10×10 cans (ViRL39K training rollout).** HiCCR assigns high credit to structural-perception tokens ("*grid*", "*rows*", "*columns*", "*10 horizontal rows*"), while the multiplicative closure ("$10 \times 10 = 100$") and the answer tag receive near-unity weight, a clean separation between "looking" and "computing".

Algorithm 1 HiCCR: Hierarchical Context Credit Reassignment

Require: Policy π_θ, dataset $\mathcal{D}$, target layers $\mathcal{L}_{\text{target}}$, hyperparameter λ_{tok}, temperature τ

for each training iteration **do**
  Sample prompts $\{\mathbf{x}^{(i)}\}$ from $\mathcal{D}$; generate G trajectories per prompt
  Compute verifiable rewards $R^{(i)}$ and group advantages $\hat{A}^{(i)}$ via Eq. (1)
  `/* Stage 1: Context Credit Scoring */`
  Register forward hooks on $\{\texttt{q_proj}, \texttt{k_proj}\}$ in layers $\mathcal{L}_{\text{target}}$
  for each micro-batch **do**
    Forward pass through π_θ (`eval`, `no_grad`)
    Compute $p_{\mathcal{V}}(\cdot \mid t)$ via Eq. (5); compute $C_{\mathcal{V}}(t)$ via Eq. (6)
  end for
  Remove all hooks
  `/* Stage 2: Hierarchical Advantage Reassignment */`
  Compute token weights $\beta_t^{(i)}$ via Eqs. (7)–(8)
  Compute trajectory weights $\alpha^{(i)}$ via Eqs. (9)–(10)
  $\tilde{A}_t^{(i)} \leftarrow \alpha^{(i)} \cdot \beta_t^{(i)} \cdot \hat{A}^{(i)}$ if $\hat{A}^{(i)} > 0$; $\tilde{A}_t^{(i)} \leftarrow \hat{A}^{(i)}$ otherwise
  `/* Stage 3: Standard Policy Update (unchanged) */`
  for each PPO epoch **do**
    Update θ using Eq. (2) with $\tilde{A}_t^{(i)}$ replacing $\hat{A}^{(i)}$
  end for
end for

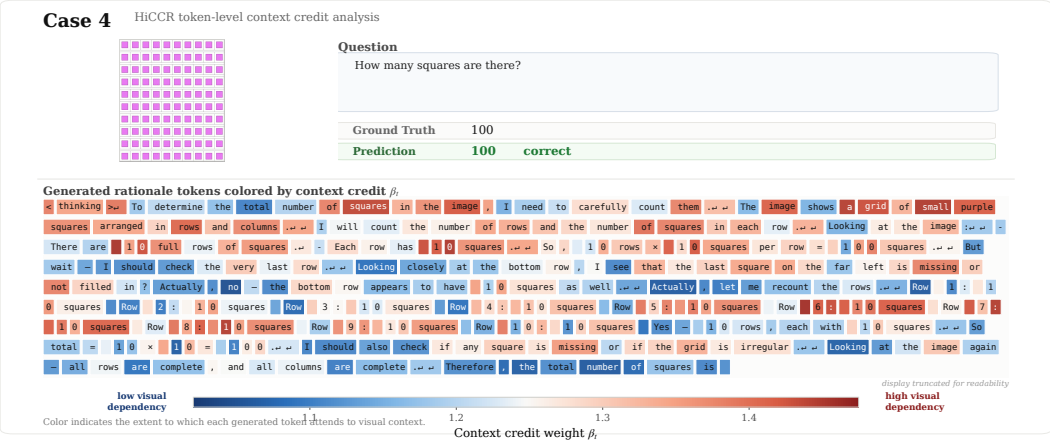


Figure 15: **Case 4: counting** 10×10 **squares with self-verification (ViRL39K training rollout).** Even in a rollout containing a self-check sub-trajectory ("*But wait — I should check the very last row*"), HiCCR consistently upweights the row-wise enumeration tokens (Row 1–Row 10, "*squares*"), showing that the credit signal is robust to reasoning detours.

analysis focuses on the credit-assignment mechanism itself and abstracts from some engineering details of large-scale RL training; nevertheless, the resulting predictions are consistent with the attention, gradient, and counterfactual diagnostics observed empirically.

I Broader Impacts

This work aims to improve the visual grounding of RLVR-trained MLLMs by reducing excessive reliance on language priors when visual evidence is required. It may benefit educational, scientific, and assistive applications where multimodal answers should be supported by image content. However, HiCCR does not guarantee faithful visual understanding, and stronger multimodal reasoning can introduce risks in high-stakes domains such as medical, legal, surveillance, or safety-critical settings. Since RL fine-tuning large models also incurs computational costs, HiCCR should be deployed with task-specific evaluation, privacy safeguards, and appropriate human oversight.

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: The Abstract and Introduction state the main diagnosis, theoretical explanation, HiCCR design, and empirical claims, which are supported by Figure 1, Theorems B.3–B.4, and Table 1. The claims are scoped to RLVR-trained MLLMs with verifiable rewards, rather than arbitrary multimodal tasks.

   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 discusses limitations in §H.

   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: The theoretical results are stated under explicit assumptions in Appendix B, including language-prior dominance and key-correlation conditions. Theorems B.3 and B.4 are accompanied by proof sketches, complete proofs, and empirical verification of central assumptions.

   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: The paper describes the algorithmic objective, credit score, token- and trajectory-level reweighting, implementation hooks, datasets, baselines, benchmarks, hyperparameters, and compute setup in §2, §3.1, Appendix F, and Appendix A.1. These details provide a reasonable path to reproduce the main experimental conclusions.

   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: Code is zipped in supplement files.

   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: The paper specifies the training data, benchmarks, evaluation protocol, baselines, optimizer, learning rate, rollout configuration, decoding setup, checkpoint selection, and HiCCR-specific hyperparameters in §3.1 and Appendix F. Additional implementation details for attention hooks and memory-efficient credit computation are also provided.

   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: [No]

   Justification: Similar to prior large-scale RL studies [63, 17, 58, 20, 39], we do not report error bars due to the high computational cost of repeatedly training all baselines. Instead, we report AVG@8, where each benchmark accuracy is averaged over eight independent forward passes to reduce stochastic decoding variance, and further support the conclusions with ablation and diagnostic analyses as described in §3.1. Fair evaluation protocol and metrics are ensured in our experiments.

   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: The paper reports the hardware and software setup configurations in Appendix F. Appendix A.1 and Table 6 further provide step time, peak memory, wall-clock overhead.

   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 uses publicly datasets and complies with the NeurIPS Code of Ethics.

   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: The paper discusses potential impacts in educational, scientific, and assistive applications in §I.

    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: The current submission does not release a high-risk pretrained model, image generator, or newly scraped dataset requiring special release safeguards.

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: All datasets and pretrained models are cited with their original sources throughout the paper

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 introduces a new algorithmic method, but the current submission does not release a new dataset, benchmark, model checkpoint, or software package as a documented asset.

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 work does not involve crowdsourcing or 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 work does not involve human subjects research.

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: [N/A]

Justification: LLMs are only used for editing.

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.