

VIGOR: Outcome-Contrastive Visual Grounding for Multimodal Policy Optimization

Haohao Xu^{1,2}, Yuyao Ge⁵, Qiuyue Zhang^{1,2}, Zhaoli Deng³, HongLin Liu³,
Sheng Ang⁴, Likang Wu^{1,2,3*}, Shenghua Liu⁵, Hongke Zhao^{1,2,3}

¹College of Management and Economics, Tianjin University, ²CACMS, Tianjin University,

³CICO Research Global Co., Ltd., ⁴Hefei University of Technology,

⁵Institute of Computing Technology, Chinese Academy of Sciences

3023205095@tju.edu.cn wul@tju.edu.cn

Abstract

While reinforcement learning with verifiable rewards (RLVR) has advanced multimodal reasoning, coarse outcome rewards offer limited guidance for assigning credit to visual evidence. Analyzing same-prompt rollout groups, we observe that different outcomes exhibit distinct temporal patterns of measured visual dependency across reasoning segments. We therefore propose **VIGOR** (Visual Grounding Outcome-contrastive Reinforcement), a policy optimization method that converts these group-wise contrasts into fine-grained supervision. VIGOR constructs outcome-conditioned prototypes from temporal visual fingerprints to reweight trajectories, uses segment-level contrasts to prioritize token gradients, and allocates the visual budget according to group difficulty. It requires no additional annotations or reward models and introduces no inference-time overhead. Across eight multimodal reasoning benchmarks, VIGOR achieves 57.5% average accuracy with Qwen2.5-VL-7B, improving over VPPO by 1.6 points. Controlled experiments further validate VIGOR on Qwen2.5-VL-3B and InternVL2.5-4B; on the latter, it outperforms VPPO on all eight tasks by 0.96 points on average. Our code is publicly available at <https://github.com/beck83879-bit/VIGOR>.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has driven substantial progress in multimodal reasoning (Guo et al., 2025; Shao et al., 2024; Yu et al., 2025; Wang et al., 2026d). In multimodal settings, an outcome reward effectively distinguishes successful from unsuccessful responses, but provides limited guidance on how visual evidence should influence different parts of a reasoning trajectory. This creates a visual credit-assignment problem: given multiple sampled responses to the same prompt, policy optimization

must determine how strongly each trajectory should contribute, where token-level updates should concentrate, and how much visual-update capacity should be allocated to the group.

This uniform treatment introduces limitations at three levels. First, at the *token level*, the algorithm applies gradient updates equally to perception-critical tokens and filler tokens, diluting the learning signal for visual grounding. Second, at the *trajectory level*, two correct responses receive identical reward despite potentially different measured visual-dependency profiles, preventing the optimizer from exploiting outcome-associated variation already present within the rollout group. Third, at the *group level*, the method allocates the same visual budget to all prompts, ignoring that some problems demand more perceptual capacity than others. Prior work improves token- and trajectory-level visual credit through perturbation-based filtering and advantage shaping (Huang et al., 2025a; Wang et al., 2026d; Li et al., 2026b; Ye et al., 2026), while structural and difficulty-aware methods refine other aspects of multimodal policy optimization (Lu et al., 2026; Li et al., 2026a; Yan et al., 2025). However, these methods largely assess rollouts independently, leaving outcome-conditioned cross-trajectory structure and group-adaptive visual budgeting underexplored.

Within mixed-accuracy prompt groups—where the same question yields both correct and incorrect rollouts—token-level visual dependency reveals a consistent structural difference (Figure 1). Correct rollouts exhibit higher and more temporally distributed visual dependency than incorrect rollouts within the same group, with the difference concentrated in early and mid reasoning segments rather than uniformly distributed. Drawing inspiration from contrastive learning principles (Khosla et al., 2020; Mroueh, 2025; Cui et al., 2026), we use this structural contrast (Gao et al., 2026) as an outcome-associated relative signal

*Corresponding author.

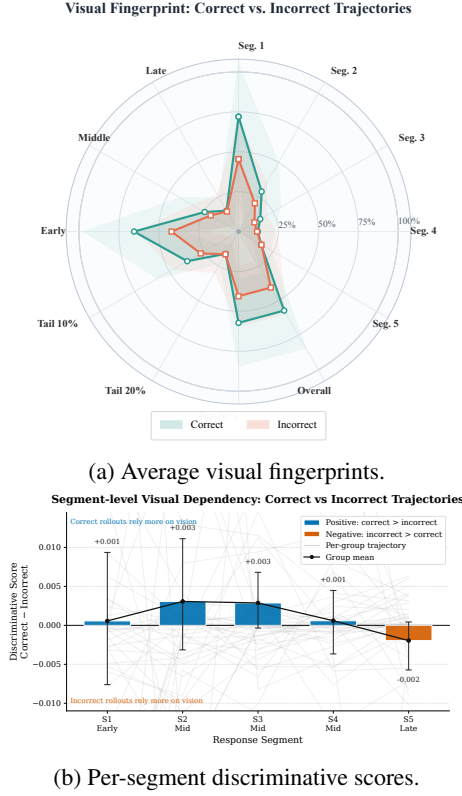


Figure 1: Outcome-conditioned patterns of measured visual dependency within the same prompt group. (a) Average visual fingerprints show that correct and incorrect rollouts differ in the magnitude and temporal distribution of the perturbation-based dependency proxy. (b) Correct-minus-incorrect segment means are positive mainly in early and middle reasoning segments. Statistics are computed from Qwen2.5-VL-7B-Instruct on 64 MMK12 prompts with eight rollouts per prompt. These differences are descriptive associations under the chosen proxy, rather than direct causal grounding attribution.

for ranking visual-dependency patterns within the same prompt group. The key insight is that this outcome-contrastive signal is freely available in any group-based RL framework—it requires no additional annotations or models, only a comparison across rollouts already generated by the base algorithm. Building on this observation, we propose **VIGOR** (Visual Grounding Outcome-contrastive Reinforcement), which transforms coarse outcome rewards into fine-grained, perception-aware supervision. VIGOR introduces three outcome-contrastive mechanisms operating at complementary granularities—trajectory, token, and group levels:

- **Outcome-Contrastive Trajectory Weighting** (§3.2): constructs multi-scale visual fingerprints and scores each correct rollout against puri-

fied correct/incorrect prototypes, assigning larger weights to outcome-consistent dependency profiles.

- **Discriminative Token Gradient Filtering** (§3.3): reranks token selection using segment-level correct–incorrect dependency gaps, concentrating updates where measured dependency is most associated with successful outcomes.
- **Difficulty-Adaptive Visual Budget** (§3.4): adjusts the per-group token filtering ratio based on group accuracy, allocating more update capacity to medium-difficulty groups with informative outcome contrasts.

Additionally, trajectory-level advantage shaping (TAS; §3.5) scales policy advantages using aggregate measured visual dependency, providing a complementary non-contrastive signal. We derive all signals from rollouts already generated by the base algorithm (Chan et al., 2024), requiring no forward passes beyond those already used by VPPO for visual dependency estimation.

We evaluate VIGOR on eight multimodal reasoning benchmarks. Starting from Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and training on ViRL39K (Wang et al., 2026b), VIGOR achieves 57.5% average accuracy, outperforming the matched VPPO baseline (Huang et al., 2025a) by 1.6 points. Controlled experiments on Qwen2.5-VL-3B and InternVL2.5-4B further evaluate transfer across model scale and visual architecture. Our contributions are:

- (1) We identify a freely available outcome-contrastive signal in group-based RLVR: under a common perturbation-based dependency estimator, correct and incorrect rollouts for the same prompt exhibit systematically different temporal patterns.
- (2) We propose VIGOR, which converts this cross-trajectory contrast into perception-aware policy optimization through three complementary mechanisms: trajectory weighting, discriminative token filtering, and difficulty-adaptive visual budgeting.
- (3) We validate VIGOR on eight multimodal reasoning benchmarks, where it improves over strong matched baselines on Qwen2.5-VL-7B. Ablations support the complementarity of the three OC mechanisms, while controlled Qwen2.5-VL-3B and InternVL2.5-4B experiments show transfer across the evaluated model scales and visual architectures.

2 Related Work

Multimodal Reasoning. Vision-language models have progressed through visual instruction tuning (Li et al., 2024; Bai et al., 2025) and chain-of-thought prompting (Wei et al., 2023; Wang et al., 2025c). RLVR further enables reasoning from verifiable outcomes, as prominently shown in the text domain by Guo et al. (2025). In multimodal settings, recent work improves data quality (Meng et al., 2025; Wang et al., 2026b), rollout robustness (Liu et al., 2025b), and perception-aware optimization via visual dependency estimation or separate credit assignment (Huang et al., 2025a; Wang et al., 2026d). However, existing visually-aware RLVR methods mainly shape rewards within individual rollouts, leaving the contrast between correct and incorrect responses to the same prompt underexplored. Since recent analyses show that perception and vision-language grounding remain major bottlenecks in multimodal LLMs (Chung et al., 2025; Yue et al., 2025a), we propose to exploit this cross-trajectory contrast as training-time supervision for stronger perceptual grounding.

Perception-Aware Multimodal RLVR. Visual grounding links language outputs to localized image evidence through region- or mask-level supervision (Ma et al., 2024; Wu et al., 2025). Recent multimodal RLVR methods introduce visual awareness through implicit perception regularization (Wang et al., 2026d), perturbation-based dependency estimation (Huang et al., 2025a; Gao et al., 2026), hidden-state or role-aware token weighting (Lu et al., 2026; Li et al., 2026b; Jiang and Zou, 2026), and token- or trajectory-level visual credit assignment (Ye et al., 2026; Wang et al., 2026c; Liu et al., 2025a). Unlike explicit localization, VIGOR treats measured perturbation-based dependency as a relative signal and organizes same-prompt rollouts into outcome-conditioned temporal prototypes, whose contrasts jointly guide trajectory weighting, token reranking, and group-adaptive visual budgeting.

Credit Assignment in Policy Optimization. Process reward models offer step-level supervision (Wang et al., 2025a; Luo et al., 2025; Ong et al., 2025; Min et al., 2026) but require auxiliary models. Model-free alternatives redistribute outcome rewards via prefix-value estimation (Tran et al., 2025; Wu et al., 2026a), entropy-weighted advantages (Tan et al., 2025), learned discriminators (Xu et al., 2026), or attribution-based

credit (Parthasarathi et al., 2025; Tran et al., 2025). Contrastive learning (Chen et al., 2020; Khosla et al., 2020) has recently entered policy optimization: Wu et al. (2026b) show GRPO is equivalent to a KL-regularized contrastive loss, and CLIPO (Cui et al., 2026) applies contrastive objectives over rollouts. We operationalize this principle for visual dependency structures, deriving credit from the structural contrast between correct and incorrect rollouts within the same prompt group—a signal freely available in group-based training.

3 Methodology

Current multimodal RLVR methods (e.g., GRPO; Shao et al., 2024) broadcast a scalar outcome reward across response tokens and do not use the variation in measured visual dependency across rollouts of the same prompt. We propose **VIGOR** (Visual Grounding Outcome-contrastive Reinforcement), which converts this within-group variation into outcome-aware supervision at trajectory, token, and group levels. VIGOR comprises three OC mechanisms—trajectory weighting (§3.2), discriminative token filtering (§3.3), and adaptive visual budgeting (§3.4)—complemented by trajectory-level advantage shaping (§3.5). We adopt the perturbation-based dependency metric and masking backbone from Huang et al. (2025a); the three OC mechanisms are introduced in this work. Figure 2 illustrates the full pipeline.

3.1 Preliminary: GRPO

Given a multimodal prompt (I, q) , the policy generates G responses $\{o_i\}_{i=1}^G$ with binary rewards R_i . GRPO (Shao et al., 2024) computes group-normalized advantages $\hat{A}_i$ and optimizes:

$$\mathcal{L}^{\text{GRPO}} = \mathbb{E} \left[\frac{1}{G} \sum_i \frac{1}{|o_i|} \sum_t \min \left(r_t \hat{A}_i, \text{clip}(r_t, 1-\epsilon, 1+\epsilon) \hat{A}_i \right) \right], \quad (1)$$

where $r_t = \pi_\theta(o_{i,t} \mid \cdot) / \pi_{\theta_{\text{old}}}(o_{i,t} \mid \cdot)$. A key limitation is that GRPO applies $\hat{A}_i$ uniformly to every token, with no mechanism to distinguish visually-grounded reasoning from text generation. To measure per-token visual reliance, we adopt the perturbation-based metric from Huang et al. (2025a): a second forward pass with a perturbed image I' (random 14×14 patch blackening, $p=0.5$)

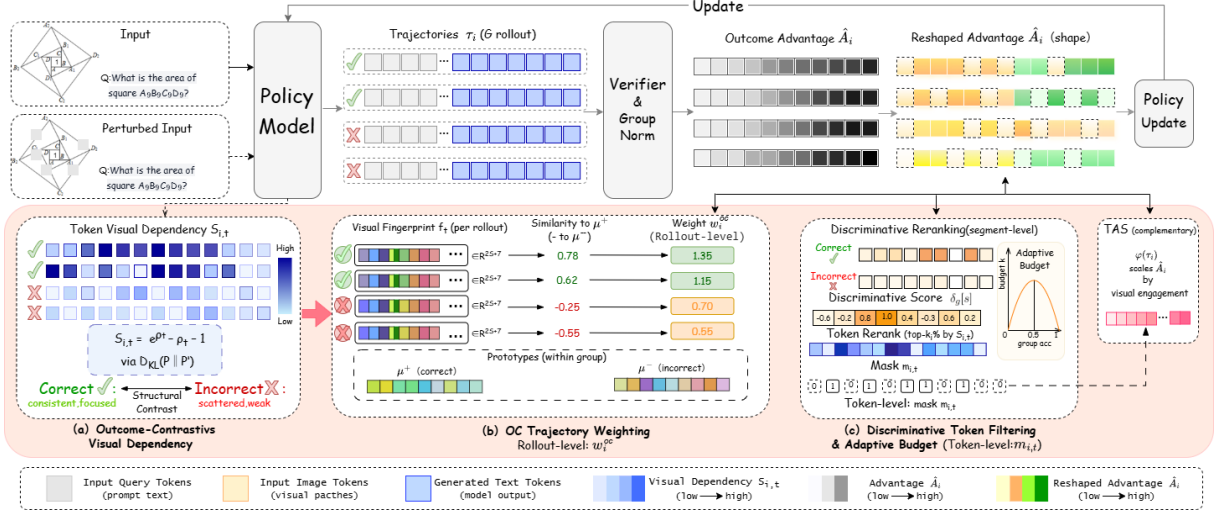


Figure 2: Overview of VIGOR. Given a prompt group with G rollouts, we compute token-level visual dependency via image perturbation, then apply three outcome-contrastive mechanisms: (1) OC Trajectory Weighting constructs visual fingerprints and derives contrastive weights from purified correct/incorrect prototypes; (2) Discriminative TGF reranks token selection using segment-level discriminative profiles; (3) Adaptive Visual Budget adjusts the filtering ratio based on group accuracy. These are complemented by Trajectory-level Advantage Shaping (TAS) that scales advantages by overall visual engagement.

yields a visual dependency score:

$$\begin{aligned} S_{i,t} &= e^{\rho_t} - \rho_t - 1, \\ \rho_t &= \log \pi_{\theta}(o_{i,t} | s_t, I') - \log \pi_{\theta}(o_{i,t} | s_t, I), \end{aligned} \quad (2)$$

The score $S_{i,t}$ measures the change in token probability under the chosen image perturbation and serves as the foundation for all subsequent mechanisms. Throughout the paper, we refer to it as *measured visual dependency*. It is a relative perturbation-based proxy rather than a direct causal measure of visual grounding.

3.2 Outcome-Contrastive Trajectory Weighting

Visual fingerprints. A scalar dependency score ignores when visual information is used. Since Figure 1b shows that correct and incorrect rollouts differ across reasoning segments, we represent each rollout by a fingerprint $\mathbf{f}_i \in \mathbb{R}^{2S+7}$, consisting of segment-level mean/max dependency over S segments and seven global statistics. Details are in Appendix A.

Prototype purification and contrastive scoring.

For each *mixed-accuracy group* g (containing both correct $\mathcal{C}_g$ and incorrect $\mathcal{W}_g$ rollouts), we build prototypes. The incorrect prototype is $\mathbf{p}_g^- = \text{mean}(\{\mathbf{f}_j\}_{j \in \mathcal{W}_g})$. Correct rollouts can still exhibit

heterogeneous measured visual-dependency profiles, and low-dependency instances may provide a weaker reference for the outcome-conditioned structure. We therefore construct a purified correct prototype by softly down-weighting such rollouts:

$$\tilde{w}_j = \begin{cases} 1 & q_j \geq \text{med}(\{q_l\}_{l \in \mathcal{C}_g}) \\ \eta & \text{otherwise} \end{cases}, \quad (3)$$

where $q_j = \bar{S}(\tau_j)$ is the mean visual dependency and $\eta \in (0, 1)$ is a down-weighting factor that softly suppresses low-dependency correct rollouts without fully excluding them (set in §4.1). The purified correct prototype is $\mathbf{p}_g^+ = \sum_j \tilde{w}_j \mathbf{f}_j$ with $\tilde{w}_j = \tilde{w}_j / \sum_l \tilde{w}_l$. For each correct rollout i , we compute a leave-one-out contrast score:

$$c_i = \cos(\mathbf{f}_i, \mathbf{p}_{g \setminus i}^+) - \cos(\mathbf{f}_i, \mathbf{p}_g^-), \quad (4)$$

where $\mathbf{p}_{g \setminus i}^+$ excludes i 's contribution to prevent self-reinforcement. When $|\mathcal{C}_g| = 1$, we set $\mathbf{p}_{g \setminus i}^+ = \mathbf{f}_i$ as a fallback and later halve β_{eff} to reduce the influence of this noisier estimate. A positive c_i indicates that rollout i 's visual fingerprint is closer to the purified correct-outcome prototype than to the incorrect-outcome prototype. We map the contrast score to a bounded weight:

$$w_i^{\text{oc}} = \text{clip}(1 + \beta_{\text{eff}} \tanh(c_i), w_{\min}, w_{\max}), \quad (5)$$

with linear warmup $\beta_{\text{eff}} = \beta \cdot \min(1, t/T_{\text{warm}})$; for singleton correct groups ($|\mathcal{C}_g|=1$), we set $\beta_{\text{eff}} \leftarrow$

$\beta_{\text{eff}}/2$. Incorrect rollouts and non-mixed groups retain $w_i^{\text{oc}}=1$. In practice, online filtering retains 192 mixed-accuracy prompt groups per step, covering 75% of the 256 prompt candidates (Appendix C.3).

3.3 Discriminative Token Gradient Filtering

Token visual dependency $\mathcal{S}_{i,t}$ indicates whether a token relies on the image, but not whether this reliance distinguishes successful from failed reasoning. We therefore refine the gradient mask of Huang et al. (2025a) with OC information, using segment-level correct–incorrect contrasts to rerank token priorities.

For each mixed-accuracy group, we compute segment-level discriminative profiles:

$$\delta_g[s] = \frac{\bar{\mathcal{S}}_{\mathcal{C}_g}^{(s)} - \bar{\mathcal{S}}_{\mathcal{W}_g}^{(s)}}{\max_{s'} |\bar{\mathcal{S}}_{\mathcal{C}_g}^{(s')} - \bar{\mathcal{S}}_{\mathcal{W}_g}^{(s')}|} \in [-1, 1], \quad (6)$$

where $\bar{\mathcal{S}}_{\mathcal{C}_g}^{(s)}$ and $\bar{\mathcal{S}}_{\mathcal{W}_g}^{(s)}$ are the mean visual dependency of correct and incorrect rollouts in segment s , respectively. For each correct rollout, we rerank the token selection priority:

$$\tilde{\mathcal{S}}_{i,t} = \mathcal{S}_{i,t} \cdot \text{clip}(1 + \alpha \delta_g[s(t)], 0.75, 1.25), \quad (7)$$

where the symmetric $[\pm 25\%]$ clip bounds keep discriminative reranking a bounded adjustment to the original visual-dependency order. For incorrect rollouts and non-mixed groups, we keep the original ranking based on $\mathcal{S}_{i,t}$. We then select the top- $k_i\%$ tokens to form the gradient mask $m_{i,t}$, ranked by $\tilde{\mathcal{S}}_{i,t}$ for OC-reranked correct rollouts and by $\mathcal{S}_{i,t}$ otherwise.

3.4 Difficulty-Adaptive Visual Budget

Given the OC-ranked token priorities above, a fixed filtering ratio may still be suboptimal: easy groups provide little contrastive signal, while medium-difficulty groups contain both successful and failed trajectories for the same prompt. We therefore set a per-group token budget:

$$k_i = k_{\text{base}} + \Delta \cdot 4 \text{acc}_g(1 - \text{acc}_g), \quad (8)$$

where acc_g is the group accuracy. This inverted-U schedule peaks at $\text{acc}_g=0.5$, where the OC signal is richest, and falls back to k_{base} for all-correct or all-incorrect groups.

3.5 Trajectory-level Advantage Shaping

Complementary to the OC modules, TAS scales each rollout by its overall visual engagement. We

normalize $\bar{\mathcal{S}}(\tau_i)$ over the current actor mini-batch to obtain $\hat{s}_i \in [0, 1]$, and compute

$$\varphi(\tau_i) = \beta_{\min} + (\beta_{\max}^{\text{dyn}} - \beta_{\min}) \hat{s}_i, \quad (9)$$

where $\beta_{\max}^{\text{dyn}} = \min(\beta_{\min} + \frac{1 - \beta_{\min}}{\bar{\mu}_{\text{norm}} + \epsilon}, \beta_{\text{cap}})$ and $\bar{\mu}_{\text{norm}}$ is the mean normalized score in the current update micro-batch. This expands the scaling range when visual dependency is low and contracts it when visual engagement is already high. Details are in Appendix B.

3.6 Training Objective and Procedure

Let $\bar{r}_t = \text{clip}(r_t, 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}})$ denote the asymmetrically clipped importance ratio. The VIGOR objective to be maximized combines all components:

$$\mathcal{L}^{\text{VIGOR}} = \mathbb{E} \left[\frac{1}{N_{\mathcal{B}}} \sum_{i,t} m_{i,t} \min(r_t \hat{A}'_i, \bar{r}_t \hat{A}'_i) \right] - \lambda_{\text{nll}} \mathcal{R}_{\text{nll}}. \quad (10)$$

Here $N_{\mathcal{B}} = \sum_i |o_i|$ is the number of valid response tokens in the actor batch, matching token-level loss averaging; $\hat{A}'_i = \varphi(\tau_i) w_i^{\text{oc}} \hat{A}_i$ is the shaped advantage; and $m_{i,t}$ is the discriminative TGF mask with adaptive budget k_i . The VPPO-inherited penalty is $\mathcal{R}_{\text{nll}} = -N_{\mathcal{B}}^{-1} \sum_{i,t} \log \pi_{\theta}(o_{i,t} | s_t)$. For non-mixed groups, $w_i^{\text{oc}}=1$, token ranking is unchanged, and $k_i=k_{\text{base}}$, recovering the base VPPO behavior. Thus, the three OC mechanisms control *how much* each rollout contributes, *where* token gradients focus, and *how many* perception-relevant tokens are retained. Algorithm 1 summarizes the update; implementation and boundary details are in Appendices C and C.6.

4 Experiments

4.1 Experimental Setup

Models and training data. Our main experiments train Qwen2.5-VL-7B-Instruct (Bai et al., 2025) on ViRL39K (Wang et al., 2026b). We additionally evaluate Qwen2.5-VL-3B-Instruct under the same two-epoch training and Avg.@8 evaluation protocol to study transfer across model scales; results are reported in Table 2. Both Qwen models are trained on $8 \times \text{A100-80GB}$ GPUs with the vision encoder unfrozen, and the detailed 3B protocol is provided in Appendix E. We further conduct a controlled one-epoch experiment on InternVL2.5-4B (Chen et al., 2025), whose InternViT encoder

Algorithm 1 One VIGOR policy-update step.

Require: Prompt batch, policy π_θ , old policy $\pi_{\theta_{\text{old}}}$, group size G

- 1: **for** each prompt group g **do**
- 2: Sample G rollouts; obtain rewards R_i and advantages $\hat{A}_i$
- 3: **end for**
- 4: Compute $\mathcal{S}_{i,t}$ from original/perturbed-image forward passes
- 5: Initialize $w_i^{\text{oc}}=1$, $\tilde{\mathcal{S}}_{i,t}=\mathcal{S}_{i,t}$, and $k_i=k_{\text{base}}$ for all rollouts
- 6: **for** each mixed-accuracy group g **do**
- 7: Build fingerprints $\mathbf{f}_i$ and purified prototypes $\mathbf{p}_g^\pm$
- 8: Compute OC weights w_i^{oc} via leave-one-out contrast
- 9: Compute $\delta_g[s]$ and rerank $\mathcal{S}_{i,t} \rightarrow \tilde{\mathcal{S}}_{i,t}$
- 10: Set adaptive visual budget k_i
- 11: **end for**
- 12: **for** each rollout o_i **do**
- 13: $\hat{A}_i' \leftarrow \varphi(\tau_i) w_i^{\text{oc}} \hat{A}_i$; select top- k_i mask $m_{i,t}$
- 14: **end for**
- 15: Update π_θ using Eq. 10

and connector differ from Qwen2.5-VL. InternVL GRPO, VPPO, and VIGOR share the same initialization, ViRL39K split, global batch/update budget, and $G=8$; full protocol details are in Appendix D.

Optimization. We use AdamW with a constant learning rate of $1\text{e-}6$ and follow the DAPO recipe (Yu et al., 2025), including dynamic sampling, asymmetric clipping, and token-level loss averaging. We sample $G=8$ rollouts per prompt. VIGOR uses $k=0.4$, $\beta_{\text{oc}}=0.3$, $\alpha=0.2$, and $\Delta=0.1$; the complete optimizer, VPPO, and VIGOR settings are listed in Appendix C.9.

Evaluation. We report Avg.@8 on MathVerse, DynaMath, MMK12, Geo3K, MathVision, WeMath, LogicVista, and MMMU-Pro. Qwen experiments follow PAPO-Eval (Wang et al., 2026d) at temperature 1.0. The controlled InternVL rows use identical prompts, decoding, and format-normalized scoring within that family; details are in Appendix D. Absolute scores are not compared across the two evaluator families.

Baselines. The controlled Qwen2.5-VL-7B block includes GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and VPPO (Huang et al., 2025a), all starting from the same base checkpoint without additional SFT. The 3B scale study additionally includes matched PAPO and PAPO-D baselines (Wang et al., 2026d). Published systems in the upper block of Table 1 are included only for context because their training data or recipes differ. Full descriptions are in Appendix C.8.

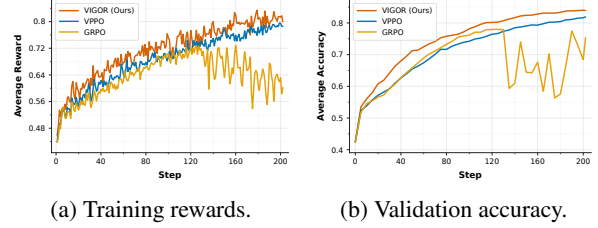


Figure 3: Training dynamics on MMK12: (a) training rewards and (b) validation accuracy over training steps. VIGOR reaches higher late-stage reward and validation accuracy than VPPO and GRPO; the validation advantage emerges early and persists over later checkpoints.

4.2 Main results.

Superior performance over visually-aware baselines. Table 1 shows that VIGOR reaches 57.5% average accuracy, compared with 55.9% for the matched VPPO baseline, 54.1% for DAPO, and 52.6% for GRPO. Relative to VPPO, VIGOR improves seven benchmarks and ties on DynaMath, with the largest gains on LogicVista (+4.1), MMK12 (+2.9), and Geo3K (+1.8). On the controlled 7B setup, VIGOR adds 11.6% training time over VPPO, while peak memory remains comparable and inference cost is unchanged (Appendix C.1). Published systems in the upper block provide broader context rather than a controlled comparison.

Transfer across model scales and architectures. On Qwen2.5-VL-3B, VIGOR improves over VPPO on all six evaluated benchmarks and by 0.9 points on average (Table 2), showing that the OC mechanisms remain effective on a smaller backbone. Under a separate matched one-epoch InternVL2.5-4B protocol, VIGOR improves VPPO on all eight tasks and by 0.96 points on average (Table 3). Together, these controlled results support transfer across model scale within Qwen2.5-VL and across the two evaluated vision-language architectures, without implying absolute score comparability between evaluator families.

Training dynamics. Figure 3 shows the training dynamics on MMK12. Under this training protocol, VIGOR reaches higher late-stage reward and validation accuracy than VPPO and GRPO, with the validation gap emerging within the first 50 steps and persisting over later checkpoints. This pattern is consistent with the OC mechanisms providing a more selective update signal, although the curves alone do not establish a causal explanation.

Table 1: Performance comparison across multimodal reasoning benchmarks (Avg.@8, %). The lower block reports controlled comparisons starting from Qwen2.5-VL-7B-Instruct and trained on ViRL39K without additional supervised fine-tuning; published systems in the upper block are included for context and may use different training data or recipes. †: evaluated using our reproduction with official author-provided prompts where original results were unavailable. *: trained on the Geo3K training set (in-domain advantage). Best results within the controlled lower block are **bolded**. Underlined published scores exceed the best controlled score but are not directly comparable.

Model	Mathematical & Geometric						Logical	Multi-discipline	Avg.
	MathVerse	DynaMath	MMK12	Geo3K	MathVision	We-Math	LogicVista	MMMU-Pro	
Open-Source Models (Trained via Pure RL)									
MM-Eureka-7B [†]	67.1	65.4	67.5	40.3	31.1	65.5	46.3	30.3	51.7
ThinkLite-7B [†]	64.2	64.6	62.6	37.6	<u>32.0</u>	66.5	39.4	28.0	49.4
VL-Rethinker-7B [†]	68.8	65.7	68.3	40.7	<u>31.9</u>	68.9	46.3	37.0	53.5
NoisyRollout-7B [†]	67.8	65.5	50.0	<u>51.8*</u>	22.1	<u>71.0</u>	47.3	34.5	51.3
R1-ShareVL-7B [†]	68.0	65.1	70.9	41.2	30.1	69.9	45.6	35.1	53.2
PAPO-D-7B	68.6	66.8 [†]	80.6	44.1	30.6 [†]	68.3	46.7	36.3	55.3
Qwen2.5-VL-7B	38.8	55.9	43.6	35.8	18.5	46.0	41.6	25.8	38.3
+ GRPO	65.2	63.1	77.3	38.6	28.2	67.3	45.2	35.5	52.6
+ DAPO	66.8	65.5	81.8	40.2	29.3	67.5	46.2	35.6	54.1
+ VPPO	70.1	67.0	81.9	45.1	30.8	68.6	46.4	37.3	55.9
+ VIGOR (Ours)	71.0	67.0	84.8	46.9	31.2	70.1	50.5	38.2	57.5

Table 2: Controlled Qwen2.5-VL-3B results across six multimodal reasoning benchmarks (Avg.@8, %). All methods share the same base checkpoint, ViRL39K split, two-epoch budget, $G=8$, and evaluator. Best results are **bolded**.

Model(Qwen2.5-VL-3B)	MathVerse	MMK12	Geo3k	LogicVista	MMMU-Pro	We-Math	Avg.
+ GRPO	55.3	57.2	28.7	38.1	25.7	58.9	44.0
+ PAPO	57.1	57.4	31.0	38.7	27.1	60.1	45.2
+ DAPO	55.6	66.8	31.2	40.7	28.4	60.0	47.1
+ PAPO-D	60.5	64.0	35.7	41.7	28.8	62.7	48.9
+ VPPO	61.8	66.5	36.8	44.5	29.2	64.6	50.6
+ VIGOR	62.7	67.0	38.4	45.0	31.1	65.0	51.5

VIGOR also maintains the logged negative-log-probability metric within a narrow range similar to VPPO, whereas GRPO shows a sharp increase after step 100 (Appendix C.5).

4.3 Ablation Studies

Component ablations. We remove one OC module at a time from the full system (Table 4). OC Trajectory Weighting and Adaptive Visual Budget contribute comparably, while Discriminative TGF provides a smaller but targeted effect on spatially demanding benchmarks. Note that VPPO already includes TAS; the +1.6 gap is entirely attributable to the OC mechanisms. Their combination achieves optimal performance, confirming complementary interactions across the three granularities: trajectory weighting selects *which* rollouts to amplify, discriminative TGF determines *where* within each rollout to focus, and adaptive budgeting controls *how much* visual capacity to allocate

per group.

Sensitivity to β_{oc} . We investigate how performance varies with the contrastive scaling coefficient β_{oc} , which controls the strength of OC trajectory weighting. As shown in Figure 4, performance peaks at $\beta_{oc}=0.3$ and degrades with both smaller and larger values. This highlights a clear trade-off: too small a β_{oc} provides insufficient contrastive modulation, while too large a value over-amplifies noisy contrast estimates from small groups, destabilizing the advantage signal.

Sensitivity to Discriminative strength α . We vary the discriminative modulation strength α in $\{0.1, 0.2, 0.4\}$. As shown in Table 7 (Appendix F), $\alpha=0.2$ achieves the best performance. A smaller $\alpha=0.1$ under-utilizes the discriminative signal, with token reranking deviating minimally from the original visual dependency ordering. Conversely,

Table 3: Controlled InternVL2.5-4B results (format-normalized Avg.@8, %). All trained methods share the base checkpoint, ViRL39K split, one-epoch budget, $G=8$, and evaluator. Best trained result in each column is bolded.

Method	MathVerse	DynaMath	MMK12	Geo3K	MathVision	We-Math	LogicVista	MMMU-Pro	Avg.
InternVL2.5-4B	36.3	42.9	32.2	29.8	13.8	32.8	31.3	22.2	30.2
+GRPO	55.4	52.1	48.2	36.1	23.4	54.0	37.1	27.7	41.7
+VPPO	57.4	55.2	49.6	38.5	24.5	54.5	39.3	28.5	43.4
+VIGOR	58.6	55.4	50.1	40.1	25.7	54.9	40.9	29.5	44.4

Table 4: Ablation study on VIGOR components. Starting from the full model, we remove one **OC** mechanism at a time. OC Trajectory Weighting and Adaptive Visual Budget contribute comparably, while Discriminative TGF provides targeted gains on spatially demanding benchmarks.

Model Configuration	MathVerse	DynaMath	MMK12	Geo3k	MathVision	We-Math	LogicVista	MMMU-Pro	Avg.
VIGOR(Ours)	71.0	67.0	84.8	46.9	31.2	70.1	50.5	38.2	57.5
+ w/o OC Weighting	70.4	66.1	81.7	43.5	31.1	69.0	48.8	37.8	56.1
+ w/o Discriminative TGF	70.9	66.1	82.3	45.1	30.9	70.5	50.2	37.1	56.6
+ w/o Adaptive Budget	69.1	66.5	82.4	45.7	31.1	69.5	48.0	37.2	56.2
Baseline (VPPO)	70.1	67.0	81.9	45.1	30.8	68.6	46.4	37.3	55.9

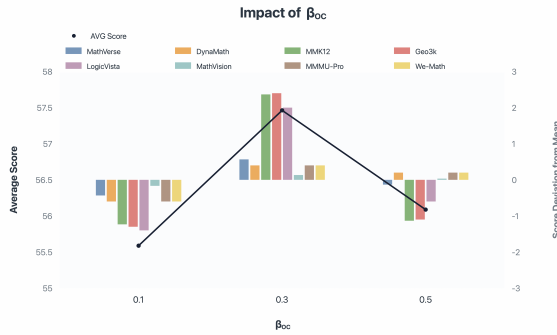


Figure 4: Sensitivity to contrastive scaling β_{oc} . The line shows average accuracy across eight benchmarks; bars show per-benchmark deviation from the mean. Performance peaks at $\beta_{oc}=0.3$, with both smaller and larger values degrading performance—indicating that moderate contrastive modulation is optimal.

$\alpha=0.4$ over-amplifies noisy estimates from groups with weak correct/incorrect contrast, distorting token selection and degrading gradient quality.

4.4 Analysis

Qualitative analysis and boundary cases. Figure 5 shows a geometry problem from a mixed-accuracy group where the correct rollout exhibits $6.3\times$ higher mean visual dependency than the incorrect one, with sustained engagement across reasoning segments. VIGOR exploits this group-relative difference through OC weighting, discriminative TGF, and adaptive budget allocation. However, substantial measured dependency does not guarantee a correct final answer. In a program-tracing case, the model correctly executes the assignments and obtains $A=2, B=1$, but maps $(2, 1)$ to option

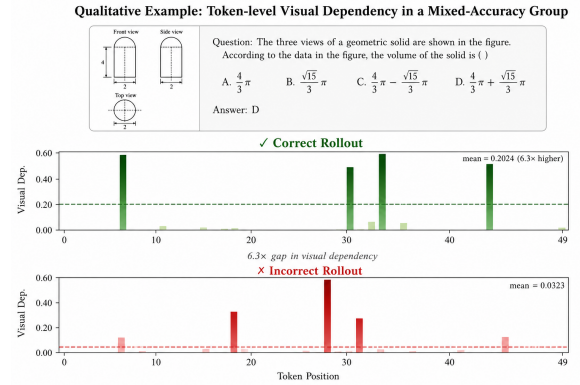


Figure 5: Token-level visual dependency for a correct and an incorrect rollout within the same prompt group. The correct rollout exhibits $6.3\times$ higher mean visual dependency (0.2024 vs. 0.0323), with sustained engagement across multiple reasoning segments, while the incorrect rollout shows sparse, localized measured-dependency spikes.

B rather than the correct option A (mean/top-20% dependency: 0.300/1.476). In a geometry case, it extracts $AD \parallel BC$ but applies the wrong angle relation, selecting $\angle 1 = \angle 4$ instead of $\angle 2 = \angle 3$ (0.137/0.681). The first observable failures therefore occur in downstream answer mapping and relation reasoning, respectively, clarifying that VIGOR improves visual credit assignment without eliminating subsequent reasoning errors.

Measured visual dependency after training.

Using identical decoding settings, we generate responses from the trained VIGOR and VPPO checkpoints on 10% of the MMK12 validation set and recompute the patch-blackening dependency proxy in Eq. 2. Figure 6 shows a 26.7% higher mean score

Table 5: Measured visual dependency under four perturbations on the same 128 MMK12 prompts. Relative gain is VIGOR over VPPO.

Perturbation	VIGOR	VPPO	Rel. gain
Patch blackening	0.0393	0.0331	+18.7%
Gaussian blur	0.0388	0.0362	+7.2%
Full-image mask	0.0544	0.0483	+12.6%
Gaussian noise	0.0045	0.0038	+18.4%

for VIGOR, with a broader rightward shift rather than a change driven by a few outliers. Thus, the accuracy improvement is accompanied by higher measured dependency under this perturbation; this relative diagnostic does not establish causal grounding.

Robustness across visual-dependency proxies.

The training-time score uses random patch blackening, so we test whether the inference-time VIGOR–VPPO trend is specific to this perturbation. On the same 128 MMK12 prompts, we recompute measured dependency with patch blackening, Gaussian blur, full-image masking, and Gaussian noise. As shown in Table 5, VIGOR has higher measured dependency under all four diagnostics. Blur degrades texture and fine detail, full masking removes holistic scene context, and pixel noise perturbs fine attributes; thus, these variants probe visual information beyond spatial patch position. Because the perturbations remove information at different scales, absolute scores are not directly comparable; we therefore examine the within-proxy VIGOR–VPPO difference on matched prompts. Its consistent direction across all four perturbations shows that the shift is not specific to patch geometry, although the varying magnitudes reflect corruption type. This agreement supports a perturbation-robust relative trend, not complete or causal grounding measurement.

5 Conclusion

We present VIGOR, an outcome-contrastive method that uses within-group differences in measured visual dependency to supervise trajectories, tokens, and prompt groups in multimodal RLVR. Its three mechanisms reweight trajectories, prioritize perception-relevant token gradients, and adapt the visual budget without additional annotations, reward models, or inference-time overhead. Across eight benchmarks, VIGOR improves average performance over matched baselines, while ablations

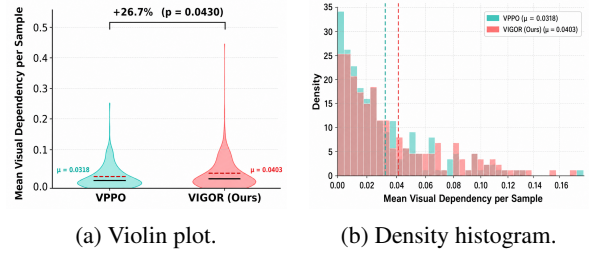


Figure 6: Per-sample measured visual dependency under the patch-blackening proxy on 10% of MMK12 validation. VIGOR has a 26.7% higher mean than VPPO; red dashed lines denote means. These values are relative perturbation diagnostics rather than causal grounding attribution.

and sensitivity analyses support the complementary and stable behavior of its components. Controlled experiments on Qwen2.5-VL at 3B and 7B scales and InternVL2.5-4B further support transfer across the evaluated model scales and vision-language architectures.

Acknowledgments

This study was partially funded by National Natural Science Foundation of China (62502340, 72471165), supported by Fundamental Research Funds for the Central Universities of China (Grant No. PA2025IISL0097), Natural Science Foundation of Anhui Province (2508085MF176), Beijing-Tianjin-Hebei Natural Science Foundation Cooperation Project (25JJJC0045), and Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center.

Limitations

VIGOR derives its OC signal from mixed-accuracy prompt groups; for uniformly correct or incorrect groups it falls back to VPPO, so its additional benefit may diminish on extremely easy or difficult prompts. The perturbation-based dependency score is a relative diagnostic rather than a causal grounding measure and may not fully capture attribute-, color-, texture-, or scene-level reliance, although consistent VIGOR–VPPO trends across four perturbations support the robustness of the comparison. Our experiments cover Qwen2.5-VL at 3B/7B and InternVL2.5-4B; evaluation across additional architectures, scales, and training distributions remains future work.

References

- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zhen Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Alex J Chan, Hao Sun, Samuel Holt, and Mihaela Van Der Schaar. 2024. Dense reward for free in reinforcement learning from human feedback. *arXiv preprint arXiv:2402.00782*.
- T Chen, S Kornblith, M Norouzi, and G Hinton. 2020. Simclr: A simple framework for contrastive learning of visual representations [c]. *International Conference on Learning Representations*, 2.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.
- Jiwan Chung, Neel Joshi, Pratyusha Sharma, Youngjae Yu, and Vibhav Vineet. 2025. What mllms learn about when they learn about multimodal reasoning: Perception, reasoning, or their integration? *arXiv preprint arXiv:2510.01719*.
- Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, and 6 others. 2025. [Process reinforcement through implicit rewards](#). *Preprint*, arXiv:2502.01456.
- Sijia Cui, Pengyu Cheng, Jiajun Song, Yongbo Gai, Guojun Zhang, Zhechao Yu, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. 2026. Clipo: Contrastive learning in policy optimization generalizes rlvr. *arXiv preprint arXiv:2603.10101*.
- Yannis Flet-Berliac, Nathan Grinsztajn, Florian Strub, Bill Wu, Eugene Choi, Chris Cremer, Arash Ahmadian, Yash Chandak, Mohammad Gheshlaghi Azar, Olivier Pietquin, and Matthieu Geist. 2025. [Contrastive policy gradient: Aligning llms on sequence-level scores in a supervised-friendly fashion](#). *Preprint*, arXiv:2406.19185.
- Shujian Gao, Yuan Wang, Jiangtao Yan, Zuxuan Wu, and Yu-Gang Jiang. 2026. Thinking with deltas: Incentivizing reinforcement learning via differential visual reasoning policy. *arXiv preprint arXiv:2601.06801*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.
- Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. 2024. Contrastive preference learning: Learning from human feedback without reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pages 18770–18798.
- Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv e-prints*, pages arXiv–2501.
- Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2025a. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. 2025b. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Bingqing Jiang and Difan Zou. 2026. Structured role-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2605.07274*.
- Junji Jiang, Likang Wu, Zhipeng Hu, Runze Wu, Xudong Shen, and Hongke Zhao. 2025. Knowledge enhanced graph contrastive learning for match outcome prediction. *Information Processing & Management*, 62(3):104010.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.
- Ang Li, Zhihang Yuan, Yang Zhang, Shouda Liu, and Yisen Wang. 2025a. Know when to explore: Difficulty-aware certainty as a guide for llm reinforcement learning. *arXiv preprint arXiv:2509.00125*.

- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. [Llava-onevision: Easy visual task transfer](#). *Preprint*, arXiv:2408.03326.
- Jinghan Li, Junfeng Fang, Jinda Lu, Yuan Wang, Xiaoyan Guo, Tianyu Zhang, Xiang Wang, and Xiangnan He. 2026a. Enhancing multi-modal llms reasoning via difficulty-aware group normalization. *arXiv preprint arXiv:2602.21743*.
- Xinyu Li, Chuang Zhao, Hongke Zhao, Likang Wu, Ming He, and Jianping Fan. 2025b. Ganprompt: Improving llm-based recommendations with gan-enhanced diversity prompts. *ACM Transactions on Information Systems*, 44(1):1–36.
- Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025c. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*.
- Yunheng Li, Hangyi Kuang, Hengrui Zhang, Jiangxia Cao, Zhaojie Liu, Qibin Hou, and Ming-Ming Cheng. 2026b. Rethinking token-level policy optimization for multimodal chain-of-thought. *arXiv preprint arXiv:2603.22847*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Liping Liu, Chunhong Zhang, Likang Wu, Chuang Zhao, Zheng Hu, Ming He, and Jianping Fan. 2025a. Instruct-of-reflection: Enhancing large language models iterative reflection capabilities via dynamic-meta instruction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9956–9978.
- Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025b. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. Understanding rl-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Zipeng Liu, Likang Wu, Ming He, Zhong Guan, Hongke Zhao, and Nan Feng. 2025d. Multi-view empowered structural graph wordification for language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24714–24722.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025e. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044.
- Jinda Lu, Junkang Wu, Jinghan Li, Kexin Huang, Shuo Yang, Guoyin Wang, Jiancan Wu, Xiang Wang, and Xiangnan He. 2026. Bridging perception and reasoning: Token reweighting for rlvr in multimodal llms. *arXiv preprint arXiv:2603.25077*.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.
- Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinze Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv e-prints*, pages arXiv–2501.
- Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer.
- 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. 2025. [Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning](#). *Preprint*, arXiv:2503.07365.
- Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. 2026. Improving vision-language models with perception-centric process reward models. *arXiv preprint arXiv:2604.24583*.
- Youssef Mroueh. 2025. Reinforcement learning with verifiable rewards: Grpo’s effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*.
- Brandon Ong, Tej Deep Pala, Vernon Toh, William Chandra Tjhi, and Soujanya Poria. 2025. Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned. *arXiv preprint arXiv:2509.23250*.
- Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. 2025. [Grpo- \$\lambda\$: Credit assignment improves llm reasoning](#). *Preprint*, arXiv:2510.00194.
- Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. [We-math: Does your large multimodal model achieve human-like mathematical reasoning?](#) *Preprint*, arXiv:2407.01284.

- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14100–14115.
- Hongze Tan, Zihan Wang, Jianfei Pan, Jinghao Lin, Hao Wang, Yifan Wu, Tao Chen, Zhihang Zheng, Zhihao Tang, and Haihua Yang. 2025. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*.
- Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2026. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *Advances in neural information processing systems*, 38:5772–5822.
- Hieu Tran, Zonghai Yao, and Hong Yu. 2025. Exploiting tree structure for credit assignment in rl training of llms. *arXiv preprint arXiv:2509.18314*.
- Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. 2026a. Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9738–9746.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2026b. V1-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.
- Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhui Wang. 2025a. [Visualprm: An effective process reward model for multimodal reasoning](#). *Preprint*, arXiv:2503.10291.
- Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025c. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.
- Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xi-angxiang Chu. 2026c. Visually-guided policy optimization for multimodal reasoning. *arXiv preprint arXiv:2604.09349*.
- Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiuxi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. 2026d. [Perception-aware policy optimization for multimodal reasoning](#). *Preprint*, arXiv:2507.06448.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. 2025. F-llm: Grounding frozen large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24710–24721. IEEE.
- Xingyu Wu, Yu Zhou, and Kay Chen Tan. 2026a. Building llms like lego: Two-dimensional architecture re-assembly of large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44939–44956.
- Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, Mark Coates, and Jian-Yun Nie. 2026b. [It takes two: Your grpo is secretly dpo](#). *Preprint*, arXiv:2510.00977.

- Nan Xiang, Lifeng Xing, and Dequan Jin. 2025. Ronfa: Robust neural field-based approach for few-shot image classification with noisy labels. *Pattern Recognition Letters*.
- Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2025. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*.
- Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.
- Tianze Xu, Yanzhao Zheng, Pengrui Lu, Lyuman-shan Ye, Yong Wu, Zhentao Zhang, Yuanqiang Yu, Chao Ma, Jihuai Zhu, Pengfei Liu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, and Gang Yu. 2026. Rubrics to tokens: Bridging response-level rubrics and token-level rewards in instruction following tasks. *Preprint*, arXiv:2604.02795.
- Hao Yan, Xingchen Liu, Hao Wang, Zhenbiao Cao, Handong Zheng, Liang Yin, Xinxing Su, Zihao Chen, Jihao Wu, Minghui Liao, Chao Weng, Wei Chen, Yuliang Liu, and Xiang Bai. 2025. Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning. *Preprint*, arXiv:2506.02537.
- Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and Jiaying Huang. 2025. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *Preprint*, arXiv:2505.16673.
- Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. 2021. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 11920–11931. PMLR.
- Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. 2026. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models. *arXiv preprint arXiv:2604.01840*.
- 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, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2025a. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2409.02813.
- Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, and 8 others. 2025b. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *Preprint*, arXiv:2504.05118.
- Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Zhixiong Zeng, Siqi Yang, Peng Shi, Lin Ma, and Jing Zhang. 2025a. Perceptual-evidence anchored reinforced learning for multimodal reasoning. *arXiv preprint arXiv:2511.18437*.
- Liang Zhang, Yang Zhang, Ziyao Lu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025b. A self-denoising model for robust few-shot relation extraction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26782–26797.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *Preprint*, arXiv:2403.14624.
- Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pages 48337–48383.

A Visual Fingerprint Details

For each rollout τ_i , the visual fingerprint $\mathbf{f}_i \in \mathbb{R}^{2S+7}$ is constructed as follows. Let $\{\mathcal{S}_{i,t}\}_{t=1}^{T_i}$ be the valid (non-padding) token visual dependency values.

Segment-level features (2S dimensions). We divide the T_i tokens into S equal segments via evenly spaced boundaries. For each segment $s \in \{1, \dots, S\}$ (using 1-based indexing for readability):

- $\mathbf{f}_i[2s-2]$: mean of $\mathcal{S}_{i,t}$ within segment s ,
- $\mathbf{f}_i[2s-1]$: max of $\mathcal{S}_{i,t}$ within segment s .

These encode the temporal structure of visual engagement—whether the model relies on the image early (during problem parsing), in the middle (during reasoning), or late (during answer generation).

Global statistics (7 dimensions). Let $\mathbf{v} = (\mathcal{S}_{i,1}, \dots, \mathcal{S}_{i,T_i})$ denote the sequence of valid token visual dependency values in their original (positional) order:

1. **Overall mean:** $\bar{\mathcal{S}}(\tau_i) = \frac{1}{T_i} \sum_t \mathcal{S}_{i,t}$
2. **Top-20% mean:** mean of the 20% highest $\mathcal{S}_{i,t}$ values (by magnitude, not position), capturing peak visual dependency intensity
3. **Tail-20% mean:** mean of $\mathcal{S}_{i,t}$ for the last 20% of tokens (by position), capturing answer-phase visual reliance
4. **Tail-10% mean:** mean of $\mathcal{S}_{i,t}$ for the last 10% of tokens, a finer answer-phase signal
5. **Early-third mean:** mean over the first $\lfloor T_i/3 \rfloor$ tokens
6. **Mid-third mean:** mean over the middle $\lfloor T_i/3 \rfloor$ tokens
7. **Late-third mean:** mean over the last $\lfloor T_i/3 \rfloor$ tokens

The tail-position features (items 3–4) are motivated by the structure of Chain-of-Thought reasoning: models typically produce extended reasoning chains followed by a short final answer. A trajectory that maintains high measured dependency into the answer phase exhibits stronger image sensitivity during answer generation, whereas one whose dependency is confined to initial parsing may rely more heavily on textual priors thereafter.

B Trajectory Advantage Shaping: Full Derivation

Normalization. For each rollout τ_i , we compute the mean visual dependency $\bar{\mathcal{S}}(\tau_i) = \frac{1}{|\tau_i|} \sum_t \mathcal{S}_{i,t}$.

On each data-parallel worker, we normalize it over the current actor mini-batch $\mathcal{B}_{\text{mb}}$:

$$\hat{s}_i = \frac{\bar{\mathcal{S}}(\tau_i) - \mathcal{S}_{\min}}{\mathcal{S}_{\max} - \mathcal{S}_{\min}}, \quad (11)$$

where $\mathcal{S}_{\min}$ and $\mathcal{S}_{\max}$ are the minimum and maximum rollout-level scores in $\mathcal{B}_{\text{mb}}$. These bounds are computed before splitting the actor mini-batch and are reused by its update micro-batches.

Shaping factor. The normalized score $\hat{s}_i \in [0, 1]$ is mapped to a scaling factor:

$$\varphi(\tau_i) = \beta_{\min} + (\beta_{\max}^{\text{dyn}} - \beta_{\min})\hat{s}_i. \quad (12)$$

We set $\beta_{\min}=0.9$, limiting downward scaling to 10%. For the current update micro-batch $\mathcal{B}_\mu$, we compute

$$\bar{\mu}_{\text{norm}} = \frac{1}{|\mathcal{B}_\mu|} \sum_{i \in \mathcal{B}_\mu} \hat{s}_i, \quad (13)$$

and dynamically set

$$\beta_{\max}^{\text{dyn}} = \min\left(\beta_{\min} + \frac{1 - \beta_{\min}}{\bar{\mu}_{\text{norm}} + \epsilon}, \beta_{\text{cap}}\right), \quad (14)$$

where $\epsilon=10^{-8}$ provides numerical stability and $\beta_{\text{cap}}=1.5$ prevents runaway scaling.

Intuition behind dynamic $\beta_{\max}$. The dynamic upper bound adapts to visual dependency in the current update micro-batch:

- **Low visual dependency** ($\bar{\mu}_{\text{norm}} \rightarrow 0$): The scaling range expands, increasing the contribution of the few rollouts with relatively high measured dependency.
- **High visual dependency** ($\bar{\mu}_{\text{norm}} \rightarrow 1$): The range contracts toward $[\beta_{\min}, 1]$, avoiding unnecessary amplification when most rollouts already engage with the image.
- **Interaction with OC weighting:** Because OC trajectory weighting (Eq. 5) also modulates advantages, β_{cap} bounds the compounded scaling.

C Implementation Details

C.1 Computational Overhead

We measure per-step wall-clock time on $8 \times \text{A100-80GB}$ with identical batch configurations (Table 6).

GPU-memory entries are taken from separate W&B resource diagnostics (the VIGOR entry averages two main-configuration runs); they show comparable reserved memory and only 0.49 GB higher

Table 6: Per-step training cost on $8 \times \text{A100-80GB}$ (batch 256, $G=8$). GPU memory is peak allocated/reserved memory per device.

Method	Step (s)	Δ vs. GRPO	GPU mem. (GB)
GRPO	512.5	—	53.38 / 72.25
VPPO	638.7	+24.6%	47.75 / 72.34
VIGOR	712.9	+39.1%	48.24 / 72.16

allocated memory than VPPO, so we do not claim a memory saving. The overhead decomposes into two parts. The augmented forward pass for computing visual dependency $\mathcal{S}_{i,t}$ (shared with VPPO) accounts for +126.2 s per step (+24.6% over GRPO). VIGOR’s own additions (fingerprint extraction, prototype construction, contrastive scoring, discriminative profiling, and adaptive budget computation) add +74.2 s over VPPO (+11.6%), primarily due to synchronization between the CPU-side analytics and the subsequent GPU gradient step. Over the full 202-step ViRL39K run, VIGOR requires approximately 40.0 hours versus 35.8 hours for VPPO, a one-time training difference of 4.2 hours with no inference overhead.

C.2 Across-Benchmark Consistency

VIGOR improves over the matched VPPO baseline on seven of the eight benchmarks and ties on DynaMath. The per-benchmark differences are shown below:

	MV	DM	MMK	Geo	MVis	WM	LV	MP
Δ	+0.9	0.0	+2.9	+1.8	+0.4	+1.5	+4.1	+0.9

The mean improvement is 1.56 points, with gains spanning mathematical, geometric, logical, and multi-discipline reasoning. These across-benchmark comparisons are descriptive: the benchmarks represent heterogeneous tasks rather than independent samples from a common task distribution.

C.3 Mixed-Accuracy Group Statistics

The outcome-contrastive mechanisms in VIGOR (§3.2–§3.4 in the main text) are active only for mixed-accuracy groups—prompt groups that produce both correct and incorrect rollouts. For all-correct or all-incorrect groups, all contrastive weights default to 1 and the visual budget defaults to k_{base} , gracefully degenerating to the base algorithm.

A natural concern is whether sufficient mixed-accuracy groups exist throughout training, especially as the policy improves and accuracy

rises. In practice, online filtering in GRPO-style training (Shao et al., 2024) removes groups with extreme accuracy (all-correct or all-incorrect, corresponding to $\text{filter_low} = 0.01$ and $\text{filter_high} = 0.99$ on the group mean reward), so retained groups are mixed-accuracy by construction. Under our training protocol, online filtering retains 192 of the 256 candidate prompt groups per update (75%) as mixed-accuracy groups, so the OC mechanisms remain active throughout training.

C.4 OC Weight Dynamics

Figure 7 shows the distribution of OC advantage weights w_i^{oc} over the course of training. Several observations confirm the stability and intended behavior of the mechanism:

- **Warmup phase (steps 0–20):** Weights gradually expand from 1.0 to the operational range, following the linear warmup schedule $\beta_{\text{eff}} = \beta \cdot \min(1, t/T_{\text{warm}})$. This prevents abrupt modulation in early training when prototypes are unreliable.
- **Stable operation (steps 20+):** The min-max range stabilizes at approximately $[0.8, 1.2]$, well within the configured bounds $[w_{\text{min}}, w_{\text{max}}] = [0.8, 1.25]$. The mechanism does not saturate at the bounds, indicating that the contrastive signal provides moderate rather than extreme modulation.
- **Mean ≈ 1.02 :** The mean weight stabilizes slightly above 1.0 (≈ 1.02) throughout training, indicating a mild upward bias toward higher-dependency correct rollouts. This confirms that OC weighting primarily *redistributes* advantage across rollouts—the 2% elevation is negligible in magnitude but directionally consistent with amplifying visually-engaged trajectories over textual shortcuts.

C.5 Training Stability

Figure 8 shows the logged negative-log-probability metric over training. GRPO exhibits a sharp increase after step 100, rising from approximately 1.0 to over 8.0, whereas VPPO and VIGOR remain within a narrow range of approximately 0.2–0.3. Because this sampled negative-log-probability is an entropy-related training diagnostic rather than an exact estimate of policy entropy, we interpret the GRPO increase as optimization instability instead of direct evidence of entropy collapse. The comparable trajectories of VPPO and VIGOR in-

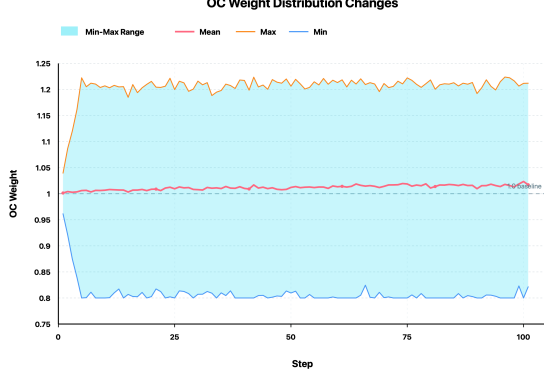


Figure 7: OC advantage weight distribution over training steps. The shaded region shows the min-max range. After the 20-step warmup, weights stabilize in $[0.8, 1.2]$ with mean ≈ 1.02 , indicating moderate and stable contrastive modulation that redistributes rather than amplifies the advantage signal.

dicates that the additional OC mechanisms do not introduce observable instability under this training protocol.

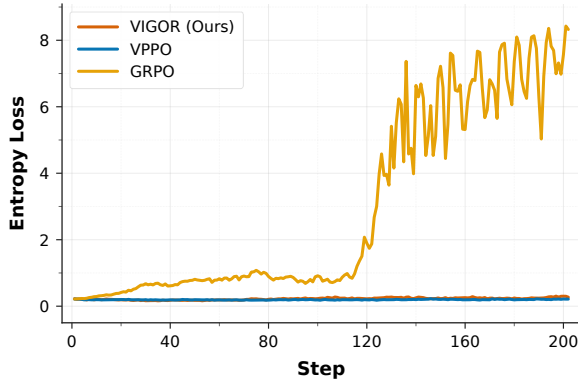


Figure 8: Logged entropy-related training metric. GRPO becomes unstable after step 100, whereas VPPO and VIGOR remain in a narrow range.

C.6 Edge Cases

Singleton correct groups ($|\mathcal{C}_g| = 1$). When a group has only one correct rollout, the leave-one-out correct prototype is undefined. We therefore set $\mathbf{p}_{g \setminus i}^+ = \mathbf{f}_i$ as a fallback, which makes $\cos(\mathbf{f}_i, \mathbf{p}_{g \setminus i}^+) = 1$ and can overestimate alignment. To mitigate this, we halve the effective beta, $\beta_{\text{eff}} \leftarrow \beta_{\text{eff}}/2$, reducing the influence of this less reliable contrast estimate.

Weak discriminative signal. When $\max_s |\bar{\mathcal{S}}_{\mathcal{C}_g}^{(s)} - \bar{\mathcal{S}}_{\mathcal{W}_g}^{(s)}| < 10^{-4}$, the normalization in Eq. 6 would amplify noise. We therefore skip discriminative reranking for that group

and retain the standard perception-based token ordering.

Degenerate visual dependency. When all tokens in a rollout have near-zero visual dependency (e.g., for text-only reasoning), the fingerprint $\mathbf{f}_i$ is approximately zero. Cosine similarity with any prototype returns 0, yielding $c_i \approx 0$ and $w_i^{\text{oc}} \approx 1$. The framework gracefully handles this case without special-casing.

C.7 Benchmark Details

We evaluate on eight diverse multimodal reasoning benchmarks spanning four categories:

Mathematical & Geometric Reasoning (6 benchmarks).

- **MathVerse** (Zhang et al., 2024): 2,612 problems testing mathematical reasoning with visual contexts including geometry diagrams, function plots, and statistical charts. Problems are categorized by the degree of visual information required.
- **MathVision** (Wang et al., 2024a): 3,040 math problems collected from real mathematical competitions, requiring multi-step reasoning over complex visual inputs such as 3D geometry and combinatorial diagrams.
- **DynaMath** (Zou et al., 2025): A dynamic benchmark with programmatically generated variants of math problems, testing robustness to visual perturbations (e.g., rotated figures, changed numerical labels).
- **MMK12** (Meng et al., 2025): 2,000 K-12 level multimodal math problems spanning arithmetic, algebra, geometry, and statistics, with diverse visual formats including handwritten diagrams and textbook figures.
- **Geo3k** (Lu et al., 2021): 3,002 geometry problems requiring formal geometric reasoning (angle computation, area calculation, congruence proofs) from diagrams with labeled points and measurements.
- **We-Math** (Qiao et al., 2024): A hierarchical benchmark decomposing mathematical reasoning into fine-grained knowledge concepts, enabling diagnosis of specific capability gaps.

Logical Reasoning (1 benchmark).

- **LogicVista** (Xiao et al., 2024): 448 problems testing logical reasoning over visual inputs including Venn diagrams, truth tables, flowcharts,

and pattern recognition tasks across 9 logical reasoning categories.

Multi-discipline Reasoning (1 benchmark).

- **MMMU-Pro** (Yue et al., 2025a): A filtered and augmented version of MMMU covering 30+ disciplines (science, engineering, art, humanities), with vision-essential questions that cannot be answered without the image.

The Qwen experiments use exact-match scoring on verifiable answers and adopt the PAPO-Eval filtered subsets (Wang et al., 2026d) where available. The controlled InternVL evaluation uses the format-normalized scorer described in Appendix D.

C.8 Baseline Details

Standard RLVR methods.

- **GRPO** (Shao et al., 2024): Group Relative Policy Optimization. Computes group-normalized advantages from G rollouts per prompt and optimizes a clipped surrogate objective. No visual awareness mechanism.
- **DAPO** (Yu et al., 2025): Dynamic sampling with online filtering, asymmetric clipping ($\epsilon_{\text{low}}=0.2$, $\epsilon_{\text{high}}=0.28$), and token-level loss averaging. Filters out all-correct and all-incorrect groups to maintain training signal quality.

Visually-aware methods.

- **VPPO** (Huang et al., 2025a): Visually-Perceptive Policy Optimization. Introduces token-level visual dependency via image perturbation (random patch blackening), then applies Trajectory-level Advantage Shaping (TAS) and Token-level Gradient Filtering (TGF) to focus learning on tokens with high measured visual dependency.
- **PAPO** (Wang et al., 2026d): Perception-Aware Policy Optimization. Uses a different perturbation strategy (Gaussian noise on image embeddings) to identify perception tokens, with separate treatment of perception vs. reasoning phases. We report PAPO-D (with dynamic filtering) variants.

Open-source models trained on different data (†).

- **MM-Eureka** (Meng et al., 2025): Trained on a curated dataset of domain-specific mathematical problems with difficulty-adaptive sampling.
- **R1-ShareVL** (Yao et al., 2025): Trained on shared visual-language reasoning data with multi-task objectives.

- **VL-Rethinker** (Wang et al., 2026b): Uses multi-round self-correction data where the model iteratively refines its reasoning.
- **ThinkLite** (Wang et al., 2025b): Lightweight reasoning enhancement through selective thinking token insertion.
- **NoisyRollout** (Liu et al., 2025b): Augments rollout diversity by injecting controlled noise during generation to improve exploration.

All methods use Qwen2.5-VL-7B-Instruct as the base model. Methods marked with † are trained on different datasets; all others use ViRL39K. All models are evaluated under the same protocol (accuracy@8, temperature 1.0, exact-match scoring via PAPO-Eval).

C.9 Full Hyperparameter Table

We train all models for two epochs on ViRL39K (Wang et al., 2026b) starting from Qwen2.5-VL-7B-Instruct (Bai et al., 2025). The vision encoder is fully unfrozen during training. For the online RL process, we sample 8 rollouts per prompt and use binary accuracy as the reward signal (1 for correct, 0 for incorrect). Following VPPO (Huang et al., 2025a), our training objective adopts the DAPO recipe (Yu et al., 2025)—incorporating dynamic sampling with online filtering, asymmetric clipping, and token-level policy gradient loss—without a KL divergence penalty. We additionally apply the VPPO-inherited components (TAS and TGF with image perturbation) as well as VIGOR’s three OC mechanisms. All key hyperparameters for the optimizer, RL process, VPPO components, and VIGOR-specific settings are detailed in Table 8.

D InternVL2.5-4B Experimental Protocol

Controlled training setup. All methods use the same InternVL2.5-4B initialization. GRPO, VPPO, and VIGOR are trained on the same ViRL39K split (38,870 examples) for one epoch, corresponding to 202 actor updates, on 4×A100-80GB GPUs. They use the same rollout batch size of 192, actor global batch size of 64, rollout group size $G=8$, and maximum prompt/response lengths of 4,096/8,192 tokens. Optimization uses AdamW in bfloat16 with a constant learning rate of $1e-6$, weight decay 0.01, and asymmetric clipping bounds of 0.20/0.28. Training-time generation uses temperature 1.0 and top- p 0.99. The same InternVL image preprocessing and visual-token layout are used

Table 7: Sensitivity to discriminative modulation strength α . Performance peaks at $\alpha=0.2$; both weaker and stronger modulation degrade results.

Discriminative strength α	MathVerse	DynaMath	MMK12	Geo3k	MathVision	We-Math	LogicVista	MMMU-Pro	Avg.
0.1	71.8	66.7	81.8	44.0	31.0	70.8	47.8	37.1	56.4
0.2	71.0	67.0	84.8	46.9	31.2	70.1	50.5	38.2	57.5
0.4	69.1	66.2	82.4	44.0	31.4	69.5	47.5	37.7	56.0

for all three RL methods. We evaluate the final one-epoch checkpoint of each method rather than selecting method-specific checkpoints by test performance.

Evaluation and answer normalization. Base, GRPO, VPPO, and VIGOR are evaluated on the same eight benchmarks with identical prompts, de-

coding settings, and scoring rules. For each test item, we sample eight responses with temperature 1.0, top- p 1.0, and at most 2,048 new tokens, and report the mean correctness across the eight responses (Avg.@8). Here, *deterministic evaluator* refers to the answer extraction and scoring procedure, not to greedy generation. The evaluator extracts only explicit final answers and canonicalizes numerical values, symbolic expressions, boxed answers, and option letters. For multiple-choice items, an explicitly generated answer value or text may be mapped to an exactly matching listed option; we do not infer an option from ambiguous free-form reasoning or from image-only option content. The same normalization is applied to all four rows in Table 3. Thus, the reported differences constitute a controlled comparison within the InternVL family and are not used for direct absolute-score comparison with Qwen2.5-VL.

Table 8: Complete hyperparameter settings for VIGOR.

Hyperparameter	Value
<i>General Training</i>	
Optimizer	AdamW
Learning rate	1×10^{-6}
LR schedule	Constant
Weight decay	0.01
Max gradient norm	1.0
Epochs	2
Mixed precision	bf16 (FSDP)
Vision encoder	Unfrozen
<i>RL Process</i>	
Global batch size	256
Rollout batch size	512
Rollouts per prompt (G)	8
Rollout temperature	1.0
Rollout top- p	0.99
Max prompt length	4096
Max response length	8192
PPO clip ε_{low}	0.2
PPO clip $\varepsilon_{\text{high}}$	0.28
Online filtering range	[0.01, 0.99]
<i>VPPO Components</i>	
Gradient filtering ratio (k)	0.4
Sampled NLL penalty coefficient	0.06
TAS β_{min}	0.9
TAS β_{cap}	1.5
Image perturbation	14×14 patch, $p=0.5$
<i>VIGOR Components</i>	
OC scaling β_{oc}	0.3
OC weight bounds $[w_{\text{min}}, w_{\text{max}}]$	[0.8, 1.25]
OC warmup steps	20
Fingerprint segments (S)	5
Purification η	0.3
Discriminative α	0.2
Disc. rerank clip	[0.75, 1.25]
Adaptive budget Δ	0.1
<i>Evaluation</i>	
Eval rollouts	8 (accuracy@8)
Eval temperature	1.0
Eval top- p	1.0
Scoring	Exact match

E Qwen2.5-VL-3B Experimental Protocol

All Qwen2.5-VL-3B methods start from the same Qwen2.5-VL-3B-Instruct checkpoint and use the same ViRL39K split. Unless otherwise specified, they follow the Qwen2.5-VL-7B recipe in Table 8: two training epochs, rollout group size $G=8$, an unfrozen vision encoder, and the same optimizer and online-filtering settings. All rows in Table 2 use the same Avg.@8 evaluation prompts, decoding configuration, and answer scorer. We report the final two-epoch checkpoint for every method without method-specific test-set checkpoint selection.

F Additional Sensitivity Analyses

Discriminative Strength Table 7 shows the effect of discriminative modulation strength α , which controls how much the segment-level outcome-contrastive signal adjusts the original token-level visual dependency ordering in Eq. 7. Performance peaks at $\alpha=0.2$; both weaker and stronger modulation degrade results.

When $\alpha=0.1$, the discriminative reranking factor $1 + \alpha\delta_g[s(t)]$ stays close to 1.0 across all seg-

ments (within $[0.9, 1.1]$), causing $\tilde{S}_{i,t}$ to remain nearly identical to the original $S_{i,t}$. The resulting gradient mask thus reduces to standard perception-based filtering, forgoing the benefit of contrasting correct and incorrect rollouts at the segment level.

When $\alpha=0.4$, the unclipped reranking factor can span $[0.6, 1.4]$ and is clipped more frequently to $[0.75, 1.25]$. For groups with a small correct–incorrect visual-dependency gap, the normalized score $\delta_g[s]$ can be sensitive to estimation noise (Eq. 6). Stronger modulation amplifies such variation, while saturation maps distinct segment scores to the same boundary and reduces ranking resolution. Consequently, token priorities may become less reliable than under the moderate default.

The default $\alpha=0.2$ yields factors in $[0.8, 1.2]$, entirely inside the $[0.75, 1.25]$ clip range. It therefore shifts gradient focus while preserving per-token visual dependency $S_{i,t}$ as the dominant ordering criterion, so outcome contrast refines rather than replaces the base perception signal.

Rollout group size. In group-based RLVR, the number of rollouts sampled for each prompt is the group size G . We vary training G under the same protocol; the representative benchmark columns are shown together with the average over all eight tasks.

Table 9: Sensitivity to rollout group size G .

G	MMK12	Geo3K	LogicVista	Avg. (8 tasks)
5	82.3	44.8	51.0	56.3
8	84.8	46.9	50.5	57.5
16	85.5	46.8	51.6	57.9

$G=5$ provides fewer same-prompt trajectories for estimating outcome-conditioned structure. Although $G=16$ reaches the highest average, it doubles the rollout samples relative to $G=8$ for only a 0.4-point improvement; hence, we use $G=8$ in the main experiments.

Adaptive-budget margin. With $k_{\text{base}}=0.4$, the maximum retained fraction at group accuracy 0.5 is $k_{\text{max}} = k_{\text{base}} + \Delta$.

All three settings outperform the fixed-budget component ablation (56.2) and VPPO (55.9) on the eight-task average. The default $\Delta=0.10$ balances additional token retention with filtering selectivity.

Table 10: Sensitivity to adaptive-budget margin Δ .

Δ (k_{max})	MMK12	Geo3K	LogicVista	Avg. (8 tasks)
0.05 (0.45)	83.9	47.0	50.2	57.0
0.10 (0.50)	84.8	46.9	50.5	57.5
0.20 (0.60)	83.5	46.5	49.0	56.7

G More Related Work

Multimodal Reasoning and Robust LLM Adaptation. Beyond the core RLVR methods discussed in §2, recent work adapts language models to broad visual reasoning, affordance understanding, and visual grounding (Huang et al., 2025b; Tan et al., 2026; Wang et al., 2026a; Liu et al., 2025e), as well as structured graph inputs through multi-view wordification (Liu et al., 2025d). Robustness to training and input variation is another recurring theme: curated high-quality subsets can match models trained on $5\times$ larger corpora (Li et al., 2025c), noisy rollouts harm convergence (Liu et al., 2025b), diversity-aware prompt generation reduces prompt sensitivity in LLM recommendation (Li et al., 2025b), and reward-centric techniques such as format penalties and length shaping provide complementary signals (Xiao et al., 2025; Zhang et al., 2025a; Yu et al., 2025). At inference time, tree search (Snell et al., 2024), process reward models (Wang et al., 2025a; Luo et al., 2025), and multi-step verification (Sun et al., 2025) further improve intermediate reasoning. VIGOR instead focuses on how an existing binary reward is distributed across tokens and trajectories through outcome-contrastive visual-dependency patterns, and is methodologically orthogonal to representation, data, prompt, and reward-design strategies.

Dense Credit Assignment and Advantage Estimation. Distributing a scalar outcome reward to individual tokens or steps is a fundamental challenge in language RL (Wu et al., 2026b). Process reward models (PRMs) provide step-level supervision by training a separate verifier on annotated reasoning traces (Lightman et al., 2024; Wang et al., 2024b; Luo et al., 2025; Cui et al., 2025), though they require auxiliary model training or step-level annotations. Model-free alternatives avoid this overhead: RLOO (Ahmadian et al., 2024) uses a leave-one-out group baseline for variance-reduced outcome gradients; REINFORCE++ (Hu, 2025) adds token-level KL penalties; VAPO (Yue et al., 2025b) addresses advantage collapse in long-chain reasoning via length-

adaptive GAE; and Dr.GRPO (Liu et al., 2025c) revises response- and group-level normalization to remove optimization bias in GRPO. Difficulty-aware training is a related thread: curriculum and self-paced strategies (Li et al., 2025a) dynamically adjust problem selection or sample weighting to stabilize convergence on hard examples. VIGOR’s adaptive visual budget (Eq. 8) shares this spirit—allocating more perceptual capacity to medium-difficulty groups where contrastive signal is richest—but requires no external difficulty labels, deriving difficulty directly from group accuracy within the standard GRPO rollout batch.

Contrastive Objectives and Prototype Representations in RL. Contrastive learning aligns complementary views in visual representation learning and graph-based outcome prediction (Chen et al., 2020; Khosla et al., 2020; He et al., 2020; Jiang et al., 2025), and has recently been extended to policy optimization and alignment. Preference-based methods including DPO (Rafailov et al., 2023) and CPL (Hejna et al., 2024) frame alignment as contrastive sequence-level objectives without an explicit reward model; Wu et al. (2026b) show that GRPO with verifiable rewards is equivalent to a KL-regularized contrastive loss; and CLIPO (Cui et al., 2026) and CoPG (Flet-Berliac et al., 2025) incorporate explicit contrastive terms over successful rollouts. Prototype-based representations provide a complementary perspective: proto-typical state representations improve sample efficiency in RL exploration (Yarats et al., 2021), while purified or denoised prototypes enhance few-shot classification robustness (Zhang et al., 2025b; Xiang et al., 2025). VIGOR draws on both threads—applying leave-one-out contrastive scoring over purified correct/incorrect prototypes—but targets *visual dependency structures* specifically, making the resulting credit signal based on outcome-conditioned perturbation responses rather than task performance or token embeddings alone.