FAITHFUL-MR1: FAITHFUL MULTIMODAL REASONING VIA ANCHORING AND REINFORCING VISUAL ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has emerged as a promising paradigm for advancing complex reasoning in large language models, and recent work extends RLVR to multimodal large language models (MLLMs). This transfer, however, surfaces a faithfulness challenge: faithful perception of task-relevant visual evidence and faithful use of that evidence during reasoning. Existing perception supervision often operates on textual descriptions rather than natively on image regions, while answer-level rewards provide limited guidance on when to attend to visual evidence during reasoning. We introduce Faithful-MR1, an anchor-then-reinforce training framework that guides where and when to attend to visual evidence through question-relevant spatial supervision and intervention-guided attention reinforcement. The Anchoring stage turns perception into an explicit pre-reasoning subtask, supervising a dedicated `<Focus>` token's attention directly against image regions rather than through textual descriptions. The Reinforcing stage uses counterfactual image intervention to identify vision-dependent response tokens, rewarding answer-correct trajectories that concentrate visual attention at those positions. Extensive experiments demonstrate that Faithful-MR1 outperforms recent multimodal reasoning baselines on both Qwen2.5-VL-Instruct 3B and 7B backbones while using substantially less training data.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has emerged as a promising paradigm for advancing reasoning in large language models (Guo et al., 2025; Shao et al., 2024; Jaech et al., 2024; Team et al., 2025; Lu et al., 2026). Recent work extends this paradigm to multimodal large language models (MLLMs), using answer-level rewards to improve performance on vision-language tasks (Huang et al., 2026b; Meng et al., 2025; Chen et al., 2025; Shen et al., 2025; Liu et al., 2025b; Yang et al., 2025; Xu et al., 2025a). In multimodal reasoning, however, reaching the correct answer also requires identifying question-relevant visual evidence and attending to it during the derivation. Final-answer rewards provide limited direct guidance for these processes.

Recent work has sought to strengthen the connection between visual evidence and multimodal reasoning. Perception-R1 (Xiao et al., 2026) and Vision-SR1 (Li et al., 2026) introduce supervision over visual descriptions, encouraging models to correctly identify and articulate information from the image. VPPO (Huang et al., 2026a) uses visual dependency to shape trajectory-level advantages and restrict policy updates to visually dependent tokens. These efforts demonstrate the value of incorporating visual evidence into the training signal, through supervision of generated descriptions or modulation of policy updates.

However, description supervision operates indirectly through generated text, while directly rewarding the allocation of attention to visual evidence when drawing on it during reasoning remains underexplored. This motivates more direct guidance of visual attention, both when identifying question-relevant evidence and when directly drawing on it during reasoning. Figure 1 illustrates why both matter. In a *perception failure*, the model reasons from an incorrectly read visual fact. In a *perception–reasoning disconnect* (PRD), it states a visual fact correctly but later ignores or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

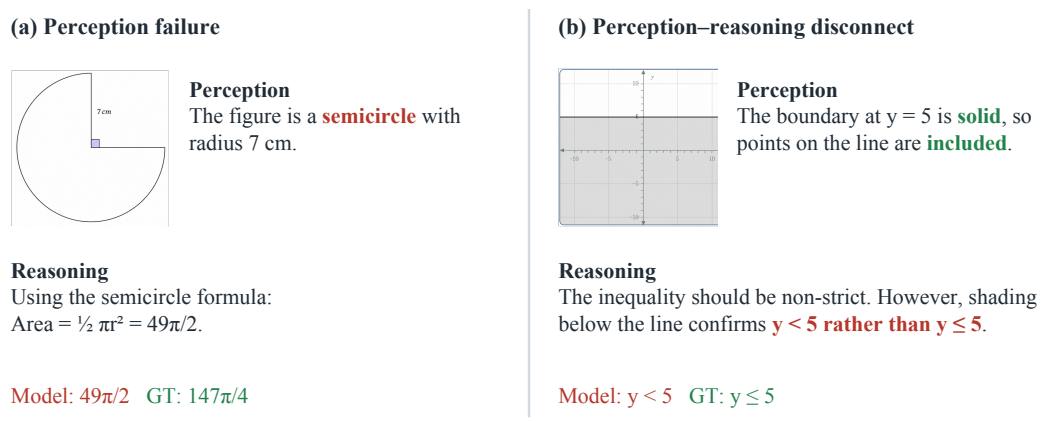


Figure 1: Two illustrative failures in multimodal reasoning. **Left, perception failure:** the model misreads the shape as a semicircle and computes its area from that incorrect premise. **Right, perception–reasoning disconnect (PRD):** the model correctly identifies that the solid boundary includes equality, but then gives a strict inequality.

contradicts it in the derivation. Correct visual descriptions alone therefore do not ensure consistent evidence use throughout reasoning.

We introduce Faithful-MR1, an anchor-then-reinforce training framework that guides where and when to attend to visual evidence through question-relevant spatial supervision and intervention-guided attention reinforcement. The **Anchoring stage** makes identifying question-relevant visual evidence an explicit pre-reasoning subtask. A dedicated `<Focus>` token precedes the reasoning response, and its attention to visual patch tokens is supervised using bounding boxes of question-relevant regions.

The **Reinforcing stage** initializes from the Anchoring checkpoint and uses the same region annotations to guide image interventions. For a sampled response, we compare next-token distributions under the original image and an image with the annotated regions masked, identifying response positions that are sensitive to the removal of visual evidence. An auxiliary reward favors answer-correct trajectories with stronger visual attention at these positions relative to the remaining response tokens. We combine this signal with answer correctness through separately normalized GRPO advantages. The framework thus connects supervision of where to attend with reinforcement of visual attention at intervention-identified positions during generation.

We evaluate Faithful-MR1 on seven benchmarks using Qwen2.5-VL-Instruct at the 3B and 7B scales. Under matched backbone and training data, it outperforms GRPO and VPPO in overall score at both scales. Ablations show that each stage improves over Vanilla SFT followed by Vanilla GRPO, with their combination achieving the strongest overall result. Attention visualizations and reasoning-trace analysis provide complementary views of the learned attention patterns and how visual facts are used in generated responses.

Our contributions are:

- We introduce Faithful-MR1, an anchor-then-reinforce framework that combines question-relevant spatial supervision with intervention-guided attention reinforcement to guide where and when models attend to visual evidence.
- We construct a training corpus of 19.2K examples with question-relevant bounding-box annotations, used both as spatial supervision for Anchoring and to define image interventions for Reinforcing. We will release the corpus, training code, and trained checkpoints.
- Experiments across seven benchmarks and two model scales establish the framework's effectiveness under matched backbone and data. Component ablations support the contributions of both stages, supplemented by attention visualizations and analysis of visual evidence use in reasoning traces.

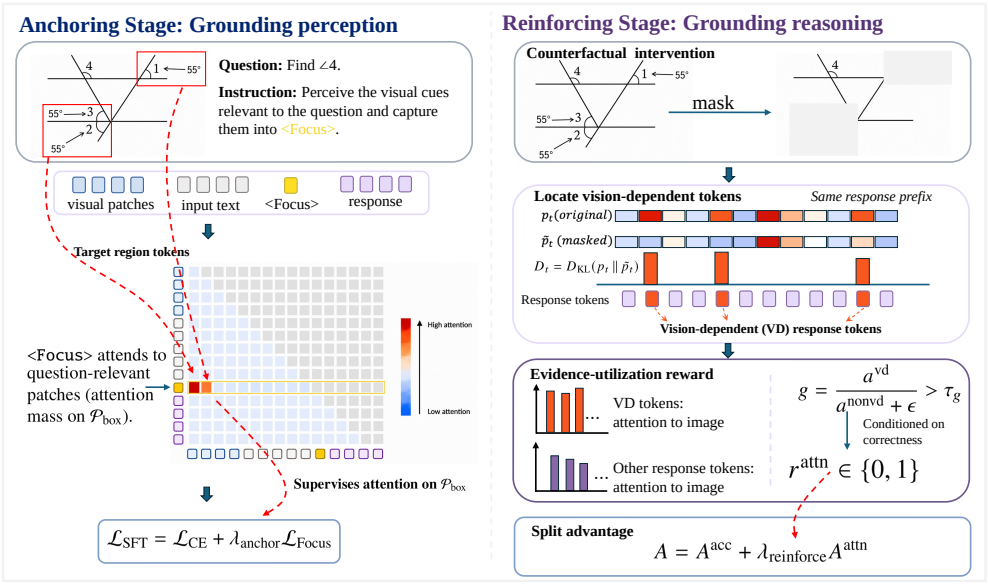


Figure 2: Overview of Faithful-MR1. **Left, Anchoring stage:** question-relevant regions (red boxes) define the target visual tokens for supervising the attention of the <Focus> token; the highlighted heatmap row shows its attention to visual patches. **Right, Reinforcing stage:** the original and region-masked images are evaluated under the same response prefix. Per-token KL divergence identifies vision-dependent response tokens (orange), whose average attention to visual tokens is compared with that of the remaining response tokens. Conditioned on answer correctness, this attention ratio yields an auxiliary reward, incorporated through a separately normalized attention advantage.

2 Related Work

RLVR improves multimodal reasoning through verifiable answer rewards (Huang et al., 2026b; Meng et al., 2025; Liu et al., 2025b), but outcome supervision does not directly constrain how visual evidence is used. Perception-R1 (Xiao et al., 2026) rewards consistency with visual annotations; Vision-SR1 (Li et al., 2026) separates visual descriptions from reasoning with self-rewarding; VPPO (Huang et al., 2026a) reweights optimization using token-level visual dependency; and PAPO (Wang et al., 2025b) adds perception-aware regularization. Faithful-MR1 combines region-grounded supervision of <Focus> attention with a reward for visual attention at vision-dependent response tokens. Appendix B provides the full discussion, including visual hallucination and methods that interleave reasoning with visual operations.

3 Proposed Approach

3.1 Overview

Figure 2 summarizes Faithful-MR1. The Anchoring stage (Section 3.2) augments standard supervised fine-tuning with a region-grounded attention loss on a dedicated <Focus> token, anchoring perception to task-relevant image regions. The Reinforcing stage (Section 3.3) initializes from the Anchoring-stage checkpoint and retains the region-grounded Focus loss during reinforcement learning: counterfactual image intervention localizes vision-dependent response tokens, and a split GRPO advantage rewards answer-correct trajectories that concentrate visual attention at those tokens.

3.2 Anchoring Visual Attention

The Anchoring stage realizes "focus first, then reason" through a dedicated <Focus> token whose attention is supervised directly against task-relevant image regions.

The <Focus> token. We prepend the instruction "Perceive the visual cues relevant to the above question and capture them into <Focus>" to the model's response, introducing <Focus> as a dedicated perceptual anchor for question-relevant visual evidence. <Focus> turns perception into an explicit one-step pre-reasoning subtask: before producing the full reasoning trajectory, the model first localizes and gathers the relevant evidence into <Focus>. Crucially, <Focus> is not a textual supervision target—it is an internal interface whose attention pattern, not its emitted tokens, carries the perceptual signal.

Region-grounded attention supervision. For each sample, let $\mathcal{B} = \{b_1, \ldots, b_M\}$ denote the bounding boxes of question-relevant evidence. Since the VLM consumes images as visual patch tokens, each box maps to a set $\mathcal{P}(b_m)$ of patch tokens; their union

$$\mathcal{P}_{\text{box}} = \bigcup_{m=1}^{M} \mathcal{P}(b_m) \tag{1}$$

defines the target region for <Focus>. Let $a^{(\ell,h)}_{\text{Focus}}$ denote the attention row of <Focus> at layer ℓ and head h, and let

$$s^{(\ell,h)} = \sum_{i \in \mathcal{P}_{\text{box}}} a^{(\ell,h)}_{\text{Focus},i} \tag{2}$$

be its attention mass on target tokens. We supervise $s^{(\ell,h)}$ toward a target ratio τ via cross-entropy

$$\mathcal{L}^{(\ell,h)}_{\text{focus}} = -\tau \log s^{(\ell,h)} - (1-\tau) \log \left(1 - s^{(\ell,h)}\right), \tag{3}$$

averaged over $\mathcal{S}$, comprising all heads in the last four language-model attention layers:

$$\mathcal{L}_{\text{Focus}} = \frac{1}{|\mathcal{S}|} \sum_{(\ell,h) \in \mathcal{S}} \mathcal{L}^{(\ell,h)}_{\text{focus}}. \tag{4}$$

Anchoring-stage supervised fine-tuning objective. The Anchoring stage combines the Focus attention loss with the autoregressive cross-entropy loss on the target answer:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid I, q, y_{<t}) \tag{5}$$

The resulting SFT objective is

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{Focus}}, \tag{6}$$

where λ_{anchor} controls perceptual guidance strength. Answer supervision is preserved while <Focus>'s attention is trained to anchor perception to task-relevant image regions.

3.3 Reinforcing Visual Attention

The Reinforcing stage targets the perception–reasoning disconnect through intervention-guided attention reinforcement. We use counterfactual image intervention: response tokens whose predictions shift when the visual evidence is masked are treated as sensitive to this masking intervention, and answer-correct trajectories that concentrate visual attention at those tokens are rewarded. The policy initializes from the Anchoring-stage checkpoint and retains both its <Focus>-anchored prompt format and the Focus loss defined in Section 3.2. Thus, region-grounded attention supervision continues during RL alongside the reinforcement objective.

Counterfactual localization of vision-dependent tokens. We construct a perturbed image by masking the annotated regions while keeping the rest of the image unchanged. Given a sampled response, the policy performs one forward pass on the original image and another on the perturbed image. Let p_t and $\tilde{p}_t$ denote the next-token distributions at response position t under the original and perturbed images. The visual sensitivity of token t is

$$D_t = D_{\text{KL}}(p_t \,\|\, \tilde{p}_t). \tag{7}$$

For each response with T valid tokens, we rank its tokens by D_t and take the top $\max(1, \lfloor 0.1T \rfloor)$ as candidates. Tokens with high D_t are sensitive to the masking intervention: their predictions change

when the critical visual evidence is removed. This turns evidence utilization from a trajectory-level notion into a token-level attribution problem—we explicitly localize the subset of response positions whose predictions are sensitive to masking the annotated image regions, in line with recent evidence that a small subset of *critical tokens* disproportionately drives reasoning quality (Lin et al., 2025).

Evidence-utilization reward. After identifying vision-dependent tokens, we reward trajectories that place stronger attention on visual content at those positions than at the remaining response tokens. Let a^{vd} be the average attention from vision-dependent response tokens to visual tokens, and a^{nonvd} the corresponding average from non-vision-dependent response tokens. For this reward, the visual-token columns include both image patch tokens and the <Focus> token. Both average attention weights over all heads and visual-token columns, then over the respective response-token group and the last four language-model layers. The evidence-utilization score

$$g = \frac{a^{\mathrm{vd}}}{a^{\mathrm{nonvd}} + \epsilon} \tag{8}$$

uses $\epsilon = 10^{-6}$ for numerical stability and is thresholded into a binary auxiliary reward $r^{\mathrm{attn}} = \mathbf{1}[g > \tau_g]$, with τ_g=2 by default. We condition the attention reward on answer correctness to reinforce visual evidence use within successful solutions. The correctness reward encourages successful task completion, while the attention reward favors answer-correct trajectories that allocate more visual attention to vision-dependent response tokens.

Split GRPO advantage. Let $r^{\mathrm{acc}} \in \{0, 1\}$ be the binary answer-correctness reward. Rather than collapsing r^{acc} and r^{attn} before normalization, we compute GRPO (Shao et al., 2024; Schulman et al., 2017) advantages separately,

$$A^{\mathrm{acc}} = \mathrm{GRPO}(r^{\mathrm{acc}}), \qquad A^{\mathrm{attn}} = \mathrm{GRPO}(r^{\mathrm{attn}}), \tag{9}$$

and combine them as $A = A^{\mathrm{acc}} + \lambda_{\mathrm{reinforce}} A^{\mathrm{attn}}$. Using the standard clipped surrogate, the GRPO-style objective is

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{(I,q),o}\left[\frac{1}{T}\sum_{t=1}^{T}\min\big(r_t(\theta)A,\ \mathrm{clip}(r_t(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})A\big)\right], \tag{10}$$

where $r_t(\theta) = \pi_\theta(o_t \mid I, q, o_{<t}) / \pi_{\theta_{\mathrm{old}}}(o_t \mid I, q, o_{<t})$ is the token-level importance ratio, and $\epsilon_{\mathrm{clip}} = 0.2$ is the base clipping parameter. The full Reinforcing-stage loss is

$$\mathcal{L}_{\mathrm{RL}}(\theta) = -\mathcal{J}_{\mathrm{GRPO}}(\theta) + \lambda_{\mathrm{anchor}} \mathcal{L}_{\mathrm{Focus}}(\theta). \tag{11}$$

Here, $\mathcal{L}_{\mathrm{Focus}}$ is the same region-grounded attention loss used during SFT, applied to the <Focus> token during RL updates. It directly maintains supervision toward the annotated visual regions, while $\lambda_{\mathrm{reinforce}}$ weights the separate attention advantage within $\mathcal{J}_{\mathrm{GRPO}}$. Answer correctness remains the primary reward signal; the attention reward additionally guides visual attention at intervention-identified response positions.

4 Experiments

We evaluate our method on both mathematical reasoning and general multimodal understanding benchmarks. We first describe the experimental setup, then present the main results.

4.1 Experimental Setup

Training data, models, and baselines. Our training data is constructed from the open-source Vision-SR1-47K dataset (Li et al., 2026), with key visual regions annotated as bounding boxes; after filtering, we obtain 6K SFT examples and 13.2K RL examples (19.2K total). The bounding boxes supervise the <Focus> token in both training stages and define the masked regions for the Reinforcing stage's counterfactual image intervention. We use Qwen2.5-VL-Instruct (Bai et al., 2025) as the backbone at the 3B and 7B scales. Under the same data setting we compare with GRPO (Shao et al., 2024) and VPPO (Huang et al., 2026a); we additionally compare with recent public multimodal reasoning checkpoints Vision-R1-7B (Huang et al., 2026b), Perception-R1-7B (Xiao et al., 2026), and Vision-SR1-7B (Li et al., 2026), which are trained at different data scales.

Table 1: Main results across multimodal benchmarks. Higher is better. Best results are shown in bold and second-best results are underlined. DynaMath is reported as Worst Case Accuracy (WCA), the benchmark's robustness metric, where a question is counted correct only if the model is correct on all 10 visual variants. "–" indicates the benchmark overlaps with the model's training set and is therefore invalid for comparison.

Method	Training Data	Math Reasoning						General Visual Question Answering			Overall
		Math Vision	Math Verse	Math Vista	We Math	Dyna Math	Math Avg	MMMU-Pro	Hallusion Bench	General Avg	
Public model comparison											
Vision-R1-7B	210K	–	48.1	69.4	–	24.6	–	34.7	68.1	51.4	–
Perception-R1-7B	1.4K	27.6	48.7	<u>73.2</u>	<u>68.3</u>	<u>25.2</u>	<u>48.6</u>	35.4	<u>70.0</u>	52.7	<u>49.8</u>
Vision-SR1-7B	56K	<u>28.3</u>	47.8	69.8	64.2	23.4	46.7	**41.1**	68.1	<u>54.6</u>	49.0
Backbone model: Qwen2.5-VL-3B-Instruct											
Qwen2.5-VL-3B-Instruct	–	20.4	36.4	56.8	51.7	13.6	35.8	28.0	64.8	46.4	38.8
+ GRPO	19.2K	20.9	33.8	60.9	52.8	11.6	36.0	<u>30.2</u>	67.2	48.7	39.6
+ VPPO	19.2K	<u>21.4</u>	<u>36.7</u>	<u>61.7</u>	<u>53.8</u>	<u>14.8</u>	<u>37.7</u>	29.8	<u>67.7</u>	48.7	<u>40.8</u>
+ **Faithful-MR1**	19.2K	**23.6**	**40.9**	**61.9**	**62.8**	**18.6**	**41.6**	**31.7**	**68.0**	**49.9**	**43.9**
Backbone model: Qwen2.5-VL-7B-Instruct											
Qwen2.5-VL-7B-Instruct	–	23.8	45.6	68.8	61.0	20.4	43.9	34.3	65.0	49.7	45.5
+ GRPO	19.2K	26.5	48.7	70.9	60.9	22.2	45.8	36.6	69.3	53.0	47.9
+ VPPO	19.2K	27.1	<u>49.5</u>	70.2	65.7	<u>25.2</u>	47.5	37.8	**70.8**	54.3	49.5
+ **Faithful-MR1**	19.2K	**28.3**	**51.9**	**73.5**	**68.9**	**26.8**	**49.9**	<u>39.7</u>	69.8	**54.8**	**51.3**

Benchmarks and metrics. We evaluate on seven benchmarks covering math reasoning (MathVision (Wang et al., 2024), MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2024a), WeMath (Qiao et al., 2025), DynaMath (Zou et al., 2025)) and general visual question answering (MMMU-Pro (Yue et al., 2025), HallusionBench (Guan et al., 2024)). For DynaMath we report Worst Case Accuracy (WCA), the benchmark's headline robustness metric where a question is counted correct only if the model is correct on all 10 visual variants. We report per-benchmark accuracy and average the available scores as the overall result.

Implementation and evaluation. The Anchoring stage trains for one epoch with learning rate 1×10^{-5}; the auxiliary <Focus> attention loss uses empirically chosen weight λ_{anchor}=0.1 and target attention mass τ=0.8 over the last 4 attention layers. The Reinforcing stage retains this Focus loss, initializes from the Anchoring checkpoint, and trains with 8 rollouts per prompt at temperature 1.0, rollout batch size 128, and learning rate 1×10^{-6}; the task reward is binary answer correctness; the evidence-utilization reward uses threshold τ_g=2 and coefficient $\lambda_{\text{reinforce}}$=0.3 at the 7B scale and 0.1 at the 3B scale. All evaluations use greedy decoding and are run with the EvalScope framework (Team, 2024); for reliability, we use Qwen3-235B-A22B-Instruct (Team, 2025) as the LLM-as-a-judge across all benchmarks. Full data-construction and supplementary training details are in Appendix A.

4.2 Main Results

Table 1 reports results across seven multimodal reasoning and perception benchmarks, organized into two complementary comparisons: the lower blocks compare training methods under matched Qwen2.5-VL-Instruct backbones (Bai et al., 2025) and 19.2K training examples, while the upper block compares publicly released multimodal reasoning checkpoints trained at a range of data scales. Under matched backbone and data, Faithful-MR1 attains the best overall score at both the 3B and 7B scales (43.9 and 51.3), and the lead extends across most individual benchmarks. Against publicly released checkpoints, Faithful-MR1-7B surpasses Vision-SR1-7B in overall score and Vision-R1-7B on the reported individual benchmarks, despite using substantially less training data. This consistent advantage demonstrates the effectiveness of Faithful-MR1's two-stage design, which

anchors perception to task-relevant image regions and reinforces the faithful use of that evidence in the reasoning chain.

Faithful-MR1's lead is most pronounced on math benchmarks, where the reasoning chain depends most heavily on faithful use of visual evidence; on General VQA, where tasks lean more on broad textual descriptions of the image, Faithful-MR1 still attains the best average. The same pattern carries over to the 3B scale. DynaMath, in particular, offers the strictest test of whether reasoning actually relies on the image: under its Worst Case Accuracy protocol, a question is counted correct only if the model answers all ten visual variants of the same problem correctly, so reasoning chains that bypass the image are penalized directly. Faithful-MR1 leads here, consistent with improvements arising from how visual evidence is propagated through the reasoning chain rather than from polishing perception text alone.

4.3 Ablation Study

Table 2 compares six configurations on Qwen2.5-VL-7B-Instruct to examine the Anchoring and Reinforcing stages. Beyond the base model and Vanilla GRPO, rows (3)–(6) combine two SFT objectives (Vanilla SFT or Anchoring) with two RL objectives (Vanilla GRPO or Reinforcing). Vanilla SFT $\rightarrow$ Reinforcing tests Reinforcing without SFT attention supervision.

Region-grounded attention supervision contributes beyond standard SFT. Holding the RL stage fixed to Vanilla GRPO, replacing Vanilla SFT with Anchoring raises the overall score from 48.6 to 49.2 (rows 3–4). With Reinforcing as the RL stage, the same replacement raises the score from 49.7 to 51.3 (rows 5–6). Thus, explicitly supervising attention toward annotated evidence regions during SFT improves downstream performance under both RL settings, with a larger gain when followed by Reinforcing (+1.6 versus +0.6 points).

Anchoring and Reinforcing provide complementary gains. Replacing Vanilla GRPO with Reinforcing improves the score by 1.1 points after Vanilla SFT (48.6 to 49.7) and by 2.1 points after Anchoring (49.2 to 51.3). The full method improves over Vanilla SFT $\rightarrow$ Vanilla GRPO by 2.7 points and achieves the highest overall score. Together, these controlled comparisons support a concrete contribution from training mechanisms that use annotated evidence regions: region-grounded attention supervision benefits the SFT initialization, and its combination with Reinforcing yields the strongest downstream performance.

Table 2: Ablation of Faithful-MR1. All rows use Qwen2.5-VL-7B-Instruct; "—" denotes no training in that stage. Vanilla SFT and GRPO omit the auxiliary attention loss and evidence-utilization reward, respectively; row (6) combines Anchoring and Reinforcing as Faithful-MR1. Best results in bold, second-best underlined.

#	SFT	RL	Math Reasoning						General Visual Question Answering			Overall
			Math Vision	Math Verse	Math Vista	We Math	Dyna Math	Math Avg	MMMU-Pro	Hallusion Bench	General Avg	
(1)	—	—	23.8	45.6	68.8	61.0	20.4	43.9	34.3	65.0	49.7	45.5
(2)	—	Vanilla GRPO	26.5	48.7	70.9	60.9	22.2	45.8	36.6	69.3	53.0	47.9
(3)	Vanilla SFT	Vanilla GRPO	26.0	50.0	<u>71.1</u>	64.3	22.0	46.7	37.8	69.1	53.4	48.6
(4)	Anchoring	Vanilla GRPO	27.5	49.5	70.9	64.5	<u>24.2</u>	47.3	<u>38.2</u>	69.6	53.9	49.2
(5)	Vanilla SFT	Reinforcing	**28.4**	<u>51.0</u>	70.2	<u>66.2</u>	24.0	<u>48.0</u>	37.9	**70.1**	<u>54.0</u>	<u>49.7</u>
(6)	**Anchoring**	**Reinforcing**	<u>28.3</u>	**51.9**	**73.5**	**68.9**	**26.8**	**49.9**	**39.7**	<u>69.8</u>	**54.8**	**51.3**

Effect of Anchoring and Reinforcing weights. We study two weights on visual-attention guidance: λ_{anchor} on the Focus loss, which is retained in both SFT and RL, and $\lambda_{\text{reinforce}}$ on the attention advantage in the Reinforcing split-GRPO objective. Both control the strength of visual-attention guidance during training. We sweep each independently on Qwen2.5-VL-3B-Instruct. Unlike Table 1, this sensitivity analysis reports DynaMath's Reasoning Robustness ($\mathrm{WCA}/\mathrm{AvgAcc}$; Figure 3), normalizing worst-case accuracy by average accuracy to assess robustness across visual variants.

Both sweeps trace an inverted-U with peaks at moderate values: the Anchoring sweep peaks at λ_{anchor}=0.1 (35.9, +3.3 over λ_{anchor}=0), and the Reinforcing sweep at $\lambda_{\text{reinforce}}$=0.1 (37.7, +1.6 over $\lambda_{\text{reinforce}}$=0); pushing either weight too high (λ_{anchor}=0.3, $\lambda_{\text{reinforce}}$=1) drops robustness below the corresponding zero-weight baseline. That both stages produce the same inverted-U is consistent with their acting on a shared visual-attention mechanism. We accordingly adopt the sweep optima λ_{anchor}=0.1 and $\lambda_{\text{reinforce}}$=0.1 for our 3B main results.

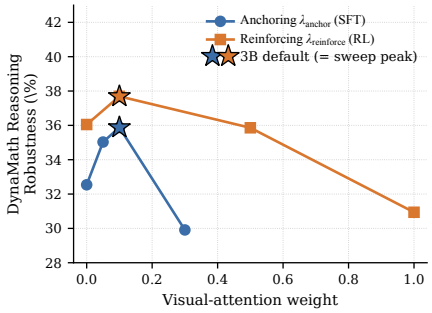


Figure 3: Effect of the Anchoring and Reinforcing weights on DynaMath Reasoning Robustness, sweeping λ_{anchor} and $\lambda_{\text{reinforce}}$ on Qwen2.5-VL-3B-Instruct. Stars mark the 3B defaults, set to the sweep peaks.

Visual anchors and evidence attention. We examine whether the visual anchors learned during Anchoring persist after Reinforcing, and compare attention to visual evidence at vision-dependent tokens with and without Anchoring on the 3B model (Figure 4). Panel (a) compares <Focus> attention before and after Reinforcing: attention to labeled geometric points learned during Anchoring remains evident afterward. Panel (b) compares attention from vision-dependent tokens after Reinforcing with and without Anchoring. On this example, Anchoring + Reinforcing assigns more attention to the image than Vanilla SFT + Reinforcing (11.42% versus 4.72%), with more pronounced attention around the labeled geometric points. Thus, this case shows both retained visual anchors and greater attention to the image beyond the <Focus> token itself.

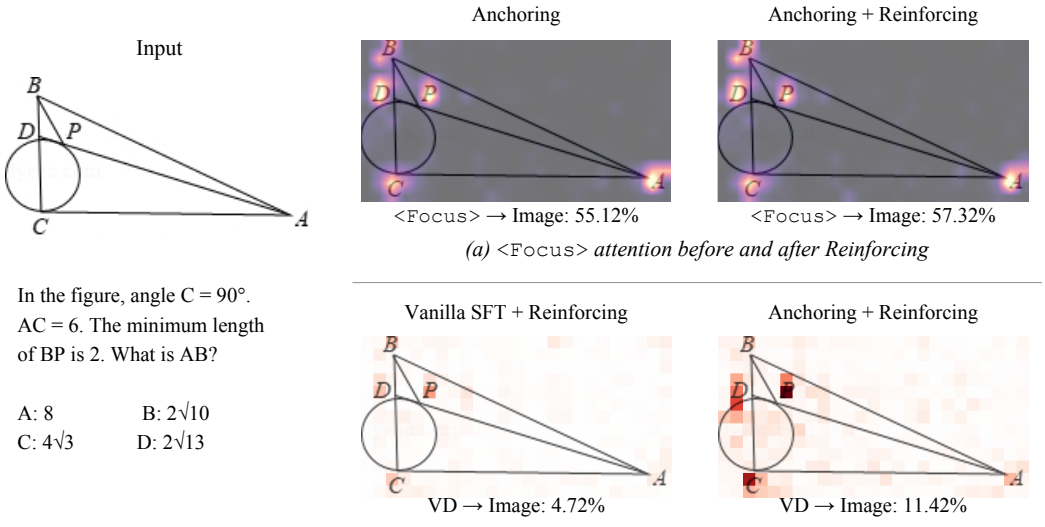


Figure 4: **Visual anchors and downstream evidence attention.** (a) <Focus> attention before and after Reinforcing, starting from Anchoring. (b) Mean VD attention after Reinforcing, starting from Vanilla SFT or Anchoring, with independently generated responses. Paired heatmaps share a scale within each panel; (b) uses absolute attention. Labels report total attention mass assigned to image tokens.

4.4 Quantifying the Perception-Reasoning Disconnect

To characterize the perception–reasoning disconnect in generated reasoning traces, we evaluate responses on the MathVerse Vision-Only split (788 responses), which removes figure-related text from the question and reduces the availability of textual visual cues. For each problem, we pre-extract

gold visual facts using Gemini-3-Flash conditioned on the image and ground-truth answer. Gemini-3-Flash then scores each critical fact using two indicators: $P = 1$ if the fact is verbalized correctly, and $U = 1$ if it serves as a load-bearing premise in the written derivation, with a value consistent with the gold fact and no later contradiction. We report Perception Coverage $= \Pr(P = 1)$, Faithful Use $= \Pr(U = 1 \mid P = 1)$, and Answer Accuracy $= \Pr(A = 1)$. These are behavioral measures of explicit fact reporting and evidence use in the response, rather than direct measurements of internal perception or computation. Appendix A.2 provides an illustrative fact-level scoring example.

Table 3 shows that higher Perception Coverage does not necessarily accompany higher answer accuracy. Vision-SR1-7B has the highest coverage (74.6%) but the lowest Faithful Use (49.4%). Faithful-MR1 achieves the highest Faithful Use (66.7%) and answer accuracy (45.6%), while producing the shortest responses (325 tokens on average). Compared with VPPO under matched backbone and training data, it improves Faithful Use and accuracy by 3.8 and 3.2 percentage points, respectively, despite lower Perception Coverage (61.3% versus 65.1%). This pattern is consistent with our framework's emphasis on directly guiding visual attention: supervision and rewards act on attention to image tokens, without requiring more extensive verbal descriptions of the image. The results suggest that more faithful use of explicitly stated evidence can accompany stronger reasoning performance even when fewer visual facts are verbalized.

Table 3: Behavioral analysis on the MathVerse Vision-Only split (788 responses; percentages except Length). Perception Coverage $= \Pr(P{=}1)$; Faithful Use $= \Pr(U{=}1 \mid P{=}1)$; Length: mean tokens per response under the Qwen2.5-VL tokenizer; Answer Acc. $= \Pr(A{=}1)$. Best **bold**, second-best underlined.

Method	Perception Coverage	Faithful Use	Length (tok)	Answer Acc.
Vision-R1-7B	70.2	54.3	526	41.9
Vision-SR1-7B	**74.6**	49.4	597	42.4
Qwen2.5-VL-7B-Instruct	65.2	63.3	401	26.9
+ VPPO	65.1	62.9	382	42.4
+ **Faithful-MR1**	61.3	**66.7**	**325**	**45.6**

5 LIMITATIONS

Faithful-MR1 relies on bounding-box annotations of question-relevant regions, obtained from Gemini-3-Flash, and incurs additional training compute from one masked-image forward pass per rollout. The method adds no inference-time computation; its generated responses are also shorter than those of the compared reasoning baselines (Table 3).

6 CONCLUSION

Faithful-MR1 addresses faithful perception and visual evidence use through two training stages: Anchoring supervises <Focus> attention against image regions, and Reinforcing retains this supervision while rewarding visual attention at vision-dependent response tokens in answer-correct trajectories. Experiments on Qwen2.5-VL 3B and 7B show gains across seven benchmarks, supported by component ablations and attention analyses. The fact-level analysis further distinguishes correct visual-fact reporting from faithful use of those facts in the written derivation. Under matched backbone and training data, the full framework achieves the strongest overall performance at both model scales. The ablations further show that region-grounded supervision improves the SFT initialization and complements attention reinforcement during RL. Together, these findings motivate training signals that address both evidence selection and evidence use: supervising where attention is allocated before reasoning, and rewarding its allocation at visually sensitive positions during reasoning. This provides a practical route to improving multimodal reasoning through visual-attention guidance.

AI USE STATEMENT

We used generative AI tools for synthetic data generation and annotation, method implementation, translation, dataset cleaning and filtering, and model-based evaluation, as well as manuscript drafting, language polishing, and LaTeX formatting. The authors take full responsibility for the final content of this work, including all AI-assisted text, claims, code, data, and results.

REFERENCES

Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025.

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.

Google DeepMind. Gemini 3 flash. https://deepmind.google/models/gemini/flash/, 2026. Accessed: 2026-04-26.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 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 (CVPR)*, pp. 14375–14385, June 2024.

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, 2025.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=bRA4lVWJVQ.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=UZIjskfbfU.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Lee Boyd-Graber, Haitao Mi, and Dong Yu. Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=C1M4ETatgM.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances llm's reasoning capability. In *Forty-second International Conference on Machine Learning*, 2025.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025b.

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 *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=KUNzEQMWU7.

Zhicong Lu, Li Jin, Ziwei Chen, Changyuan Tian, Xian Sun, Xiaoyu Li, Yi Zhang, Qi Li, and Guangluan Xu. Relation-aware multi-pass comparison deconfounded network for change captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13349–13363, 2024b.

Zhicong Lu, Changyuan Tian, PeiguangLi PeiguangLi, Li Jin, Sirui Wang, Wei Jia, Ying Shen, and Guangluan Xu. Piper: Benchmarking and prompting event reasoning boundary of llms via debiasing-distillation enhanced tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28591–28613, 2025.

Zhicong Lu, Zichuan Lin, Wei Jia, Changyuan Tian, Deheng Ye, Peiguang Li, Li Jin, Nayu Liu, Guangluan Xu, and Wei Feng. Hisr: Hindsight information modulated segmental process rewards for multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2603.18683*, 2026.

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

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

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. *arXiv preprint arXiv:2409.19256*, 2024.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

ModelScope Team. EvalScope: Evaluation framework for large models, 2024. URL https://github.com/modelscope/evalscope.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24290–24315, 2025.

Changyuan Tian, Zhicong Lu, Shuang Qian, Nayu Liu, Peiguang Li, Li Jin, Leiyi Hu, Zhizhao Zeng, Sirui Wang, Ke Zeng, et al. Rectify evaluation preference: Improving llms' critique on math reasoning via perplexity-aware reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33241–33249, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025a.

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*, 2025b.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of MLLMs via visual perception reward. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KttCXdjj4w.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025a.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=KzU33wR875.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2376–2385, 2025.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, pp. 169–186, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73241-6. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025a.

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

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VOAMTA8jKu.

A Additional Experimental Details

A.1 Training and Evaluation Protocol

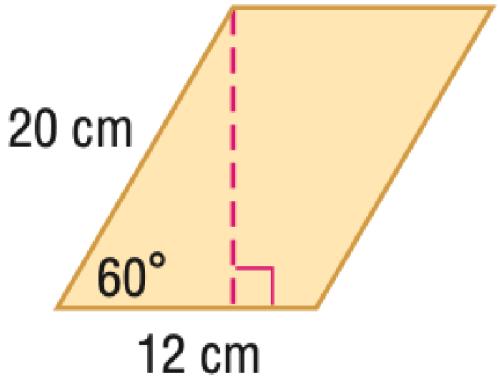
Training data construction. We construct our training data from the open-source Vision-SR1-47K dataset (Li et al., 2026). Key visual regions are annotated as bounding boxes by Gemini-3-Flash (Google DeepMind, 2026) conditioned on the image, query, and ground-truth answer (full prompt in Appendix A.3). For the SFT split, responses are generated by rollouts from Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and filtered by rules and an LLM-as-a-judge for answer correctness and reasoning quality. After filtering we obtain 6K SFT examples and 13.2K RL examples (19.2K total). All matched-backbone baselines use this same split; GRPO uses only the final-answer reward, while VPPO and our method use the same answer reward with their respective faithfulness training signals.

Implementation details. We implement the Anchoring stage with LlamaFactory (Zheng et al., 2024) and the Reinforcing stage with EasyR1 (Zheng et al., 2025a; Sheng et al., 2024), both with bf16 precision. GRPO and VPPO baselines are reproduced from their official open-source implementations on our training data. All training runs are conducted on 32 AMD MI308X GPUs under the ROCm/HIP runtime.

A.2 Paired Case Studies

This case asks for the perimeter of a parallelogram with adjacent sides labeled 20 cm and 12 cm. VPPO correctly reads the 12 cm side but substitutes a 10 cm projection in the perimeter calculation ($P = 1, U = 0$ for this fact). Faithful-MR1 consistently uses the stated side lengths and obtains the correct perimeter of 64 cm ($P = 1, U = 1$).

Find the perimeter of the parallelogram. Round to the nearest tenth if necessary.
Choices: A:60 B:62 C:64 D:66



Target visual fact: The adjacent side is labeled 12 cm. **Ground truth:** C (64 cm).

VPPO: original response

To find the perimeter of the parallelogram, we need to determine the lengths of all four sides. We know two sides are 20 cm and 12 cm, and we need to find the lengths of the other two sides.
First, let's use the given angle and the height to find the length of the side that is 12 cm. The height of the parallelogram is perpendicular to the base, and we can use trigonometry to find the length of the side that is 20 cm.
The height h of the parallelogram can be found using the sine of the angle:

$$h = 20 \sin(60^\circ) = 20 \cdot \frac{\sqrt{3}}{2} = 10\sqrt{3} \approx 17.32 \text{ cm}$$

Now, we can use the Pythagorean theorem to find the length of the side that is 12 cm. The side that is 12 cm is the base of a right triangle with the height $10\sqrt{3}$ and the hypotenuse 20 cm. The other side of the right triangle is:

$$\sqrt{20^2 - (10\sqrt{3})^2} = \sqrt{400 - 300} = \sqrt{100} = 10 \text{ cm}$$

So, the side that is 12 cm is the base of the parallelogram, and the other side is 10 cm. Therefore, the perimeter of the parallelogram is:

$$2 \times (20 + 10) = 2 \times 30 = 60 \text{ cm}$$

Thus, the perimeter of the parallelogram is $\boxed{A}$.

Faithful-MR1: original response

To find the perimeter of the parallelogram, we need to determine the lengths of all four sides. We already know two sides: one is 20 cm and the other is 12 cm. Since opposite sides of a parallelogram are equal, the other two sides will also be 20 cm and 12 cm.
The perimeter P of a parallelogram is given by the formula:

$$P = 2 \times (\text{side}_1 + \text{side}_2)$$

Substitute the known side lengths:

$$P = 2 \times (20 + 12)$$

Calculate the sum inside the parentheses:

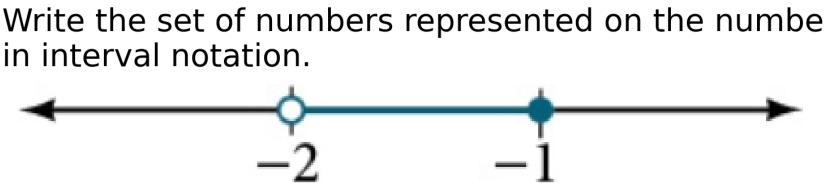
$$20 + 12 = 32$$

Now, multiply by 2:

$$P = 2 \times 32 = 64$$

The perimeter of the parallelogram is: $\boxed{C}$

This case asks for the interval represented by a shaded number-line segment with an open endpoint at -2 and a closed endpoint at -1. Vision-SR1 correctly identifies the endpoints but reverses their order, returning $[-1, -2)$. Faithful-MR1 preserves their order and inclusion conditions, yielding the correct interval $(-2, -1]$.



Target visual fact: The shaded segment has an open left endpoint at -2 and a closed right endpoint at -1. **Correct interval:** $(-2, -1]$.

Vision-SR1: original response

To determine the set of numbers represented on the number line in interval notation, we need to analyze the markings and the type of endpoints used on the number line.

1. **Identify the number line markings:**
   - The number line starts at -2 and goes to -1.
   - There is an open circle at -2 and a closed circle at -1.
2. **Interpret the markings:**
   - An open circle at -2 means that -2 is not included in the set.
   - A closed circle at -1 means that -1 is included in the set.
3. **Translate to interval notation:**
   - An open interval does not include the endpoints, denoted by parentheses.
   - A closed interval includes the endpoint, denoted by brackets.
4. **Combine the information:**
   - The interval includes all numbers from -1 up to, but not including, -2.
5. **Write in interval notation:**
   - The interval notation for this set is: $[-1, -2)$.

Verification:
- The open circle at -2 ensures the interval is not inclusive of -2.
- The closed circle at -1 ensures the interval is inclusive of -1.

Thus, the set of numbers represented on the number line in interval notation is $[-1, -2)$.
<description>
The image shows a number line with two markers: an open circle at –2 and a closed (filled) circle at –1. This indicates that the set of numbers includes all values from (and including) –1 up to (but not including) –2. In interval notation, this is written as [–1, –2)./n</description>
$\boxed{[-1, -2)}$

Faithful-MR1: original response

To determine the set of numbers represented on the number line in interval notation:

1. Identify the endpoints:
   - The left endpoint is -2, which is open (indicated by the hollow circle).
   - The right endpoint is -1, which is closed (indicated by the filled circle).
2. Write the interval notation:
   - Since the interval includes all numbers from -2 up to and including -1, we use a bracket for the closed endpoint and a parenthesis for the open endpoint.

The final answer is: $\boxed{(-2, -1]}$

A.3 PROMPTS

This section lists the prompts used at every stage where Faithful-MR1 invokes an LLM or a VLM as a pipeline component. Placeholders are written as `{NAME}` and are filled in at call time.

(P1) Bounding-box region annotation (Gemini-3-Flash). Conditioned on the image, query, and ground-truth answer, we ask Gemini-3-Flash to localize the visual entities required to verify the answer as bounding boxes (normalized to $[0, 1000]$).

Role: You are a Lead Visual Grounding Engineer specializing in Evidence-Based Localization. Your goal is to identify and ground only the essential visual anchors that constitute direct proof for the `{QUERY}` and `{ANSWER}`.

Task Objective: Analyze the `{IMAGE}`, `{QUERY}`, and `{ANSWER}`. Your mission is to provide precise bounding boxes for the essential visual entities strictly required to verify the correctness of the `{ANSWER}`. Do not label background clutter or entities that do not directly support the logical derivation from the `{QUERY}` to the `{ANSWER}`.

Strict Annotation Principles:

1. **Evidence-Centricity (Primary Rule)**: Only annotate objects that act as "smoking guns" for the answer.
2. **Atomic Localization**: Label individual, specific objects. Avoid "scene-level" or "group" boxes.
3. **Tightness**: Boxes should closely fit the visible object boundaries with reasonable margins that respect the object's natural contours. Avoid excessive padding, but allow slight breathing room to preserve the object's context and readability.
4. **Exclusion of Contextual Noise**: Do not annotate:
   - Large environmental or structural elements (e.g., "ground", "wall", "sky", "full figure", or "full table") unless they are the direct subject of the query.
   - Implied objects that are not clearly visible.
   - Objects mentioned in the answer for purely descriptive/decorative purposes that don't assist in the logical deduction.
5. **Vision-Only Scenarios**: In cases where the query and options are rendered directly within the `{IMAGE}`, you must explicitly provide bounding boxes for the query text and every individual option, regardless of which one is the correct `{ANSWER}`.

Reasoning Process (Chain-of-Thought): Before generating coordinates, perform the following mental steps:

1. **Core Subject Identification**: What are the primary subjects in the `{IMAGE}` required to justify the `{ANSWER}` for the `{QUERY}`?
2. **Causal Filter**: Does this specific entity provide direct supporting or refuting evidence to justify the `{ANSWER}` for the `{QUERY}`? (e.g., If the query asks if a car is "red" and the answer is "Incorrect", the car itself is the refuting evidence to show its actual color, while the road it sits on is irrelevant.)
3. **Spatial Mapping**: Precisely locate the regions of the remaining "Essential Anchors."

Output Format: Return a valid JSON object only. Use normalized coordinates `[ymin, xmin, ymax, xmax]` in the range of `[0, 1000]`.

```
{
  "reasoning_process": "Provide a concise step-by-step analysis
  following the 3-step Reasoning Process: 1) Core Subject
  Identification, 2) Causal Filter, and 3) Spatial Mapping.",
  "objects": [
    {
      "name": "specific_object_name",
      "box_2d": [ymin, xmin, ymax, xmax]
    }
  ]
}
```

(P2) Training and inference prompt template. The shared prompt template used by our method for both training stages and evaluation inference.

```
<image>

{PROBLEM}

Perceive the visual cues relevant to the above question and capture
them into <Focus>.

Think step by step, then provide your final answer enclosed within
`\boxed{}` at the end of your response. For multiple-choice
questions, place only the option letter (e.g., A) in `\boxed{}`. For
other questions, place the final numerical or text answer in
`\boxed{}`.
```

(P3) LLM-as-a-judge (Qwen3-235B-A22B-Instruct). We use Qwen3-235B-A22B-Instruct via EvalScope to compare each model's predicted answer against the ground truth and return a single letter `A` (correct) or `B` (incorrect).

```
Your job is to look at a question, a gold target, and a predicted
answer, and return a letter "A" or "B" to indicate whether the
predicted answer is correct or incorrect.

[Question]
{QUESTION}

[Reference Answer]
{GOLD}

[Predicted Answer]
{PRED}

Evaluate the model's answer based on correctness compared to the
reference answer.
Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT

Just return the letters "A" or "B", with no text around it.
```

(P4) Gold visual facts extraction. Conditioned on the image and ground-truth answer, we ask Gemini-3-Flash to enumerate the minimal set of visual facts on which any correct solution must rely.

You are extracting the visual facts required to solve a multimodal math problem.

Question
`{PROBLEM}`

Ground-truth answer
`{GROUND_TRUTH}`

Task: Identify the *minimal set* of visual facts (numeric labels, geometric relations, spatial configurations, table entries, chart values, ...) that any correct solution must rely on.

For each fact output:

- `"id"`: short unique identifier (`"F1"`, `"F2"`, ...)
- `"fact"`: one canonical sentence stating the fact
- `"criticality"`: `"critical"` if the answer is unreachable without it, else `"supporting"`
- `"fact_type"`: one of `"numeric_label"`, `"geometric_relation"`, `"spatial_position"`, `"object_count"`, `"chart_value"`, `"table_entry"`, `"other"`

Rules:

1. List ONLY facts that the IMAGE provides; do not list facts that are already stated in the question text.

2. Be conservative: if a fact is implied or redundant, mark it `"supporting"`.
3. Use the ground-truth answer ONLY to decide which image facts are needed; never copy the answer or any derivation conclusion into the fact list.
4. Aim for 1–5 critical facts. Do not over-list.
5. Keep each fact one short, declarative sentence.
6. **STRICT NO-ANSWER-LEAK RULE**: Every listed fact must be DIRECTLY OBSERVABLE from the image (a label you can read, a line/region you can see, a count you can make). DO NOT include any of the following, no matter how true:
   - the ground-truth answer itself or any restatement,
   - the result of a derivation (e.g., "the maximum occurs at X", "the side equals Y", "$f(x)$ has Z roots"),
   - properties that require solving the problem to know (e.g., "triangle ABC is isosceles" when isoscelesness is the conclusion, not a marked label).

   If a candidate fact is something a solver PROVES rather than READS, drop it.
7. Sanity check before emitting: for each fact, ask "could a person who has NOT solved this problem look at the image and verify this fact in seconds?" If no, the fact is leaking the solution; remove it.

Return only one JSON object, no markdown, no extra text:

```
{
  "gold_facts": [
    {"id": "F1", "fact": "...", "criticality": "critical",
    "fact_type": "..."},
    ...
  ]
}
```

(P5) Per-fact PRD scoring. For each gold fact, we ask Gemini-3-Flash to independently decide whether it is (i) explicitly *stated* in the response and (ii) *instrumentally used* as a load-bearing premise in the derivation.

You are an expert evaluator for multimodal mathematical reasoning. Your goal is to detect Perception-Reasoning Disconnect (PRD): cases where the model *states* a correct visual fact but its *derivation of the final answer* does not actually rely on that fact (drops it, substitutes a different value, or contradicts it).

Question
`{PROBLEM}`

Ground-truth answer
`{GROUND_TRUTH}`

Gold visual facts (anchor for evaluation; pre-extracted from image and solution)
`{GOLD_FACTS_JSON}`

Model response (verbatim, do NOT split or reformat)

```
---
{MODEL_RESPONSE}
---
```

Critical Distinction.
Two questions you must keep STRICTLY SEPARATE:

(A) "Did the model STATE this visual fact?" **Step 1.** Narrow scope: only counts an explicit verbal mention or paraphrase of the fact in the response.

(B) "Did the model's DERIVATION of the final answer ACTUALLY USE this fact as a load-bearing premise (i.e., the stated value/relation is plugged into a calculation, equation, or logical deduction that propagates to the answer)?" **Step 2.** Strictly stronger than (A): a fact that appears only as decoration, in an unrelated step, or that is later bypassed by a different value DOES NOT count as load-bearing.

A model can mention a fact and then ignore it, contradict it, or substitute a different value when computing.

Calibration example (study before answering).
Suppose Gold fact F1 = "central angle AOB is labeled 140 in the diagram", and the model writes: "I

observe that angle AOB is 140. To find angle ACB, I recall that the inscribed angle is typically half of 90, which is 45. So ACB = 45." Correct judgment for F1:

- **Step 1 (STATEMENT)**: `stated_in_response=true`, `stated_correctly=true`, `statement_quote="angle AOB is 140"`.
- **Step 2 (INSTRUMENTAL USE)**: `instrumental_in_derivation=false` (140 is mentioned, then ignored), `value_consistency="na"`, `contradicted_in_derivation=true` ("half of 90, which is 45" implicitly uses a different angle than 140), `derivation_quote="half of 90, which is 45"`.

This is a textbook PRD: stated correctly but not load-bearing in the derivation.

Evaluation Procedure.
Step 1 – STATEMENT (narrow scope: explicit mention only). For EACH gold fact, output:

- `"id"`, `"stated_in_response"` (true | false),
- `"stated_correctly"` (true | false | `"na"`; `"na"` only if not stated),
- `"statement_quote"`: short verbatim quote, `""` if not stated.

Step 2 – INSTRUMENTAL USE (load-bearing in derivation). INDEPENDENT of Step 1; trace the derivation and decide whether the fact's value/relation actually carries through to the answer.

- `"id"`, `"instrumental_in_derivation"` (true | false): true iff the stated value/relation is plugged into a calculation, equation, or deductive step that leads to the final answer; false if it is decoration, never referenced again, or a different value is used instead.
- `"value_consistency"` (`"consistent"` | `"substituted"` | `"na"`): consistent if numeric/symbolic value used in derivation matches the stated one; substituted if a different value is used; na if non-numeric or `instrumental_in_derivation=false`.
- `"contradicted_in_derivation"` (true | false): true if any derivation step overrides, ignores-and-substitutes, or otherwise contradicts the stated fact.
- `"derivation_quote"`: short verbatim quote of the derivation step where the fact is used/substituted/contradicted, or `""`.

Step 3 – Reasoning quality note (audit only; one sentence). Briefly describe any logic or arithmetic errors. NOT used in P/U computation; recorded for analysis.
Step 4 – Final answer.

- `"predicted_answer"`: model's final answer as it appears, `""` if missing.
- `"answer_matches_ground_truth"`: true | false | `"unclear"` (semantic match allowed).
- `"answer_match_reason"`: one sentence.

Output (single JSON object, no markdown, no extra text):

```
{
  "fact_statement": [
    {"id": "F1", "stated_in_response": true, "stated_correctly":
    true, "statement_quote": "..."}
  ],
  "fact_instrumental_use": [
    {"id": "F1", "instrumental_in_derivation": true,
    "value_consistency": "consistent", "contradicted_in_derivation":
    false, "derivation_quote": "..."}
  ],
  "reasoning_quality_note": "...",
  "predicted_answer": "...",
  "answer_matches_ground_truth": true,
  "answer_match_reason": "..."
}
```

B Extended Related Work

B.1 RLVR for Multimodal Reasoning

Reinforcement Learning with Verifiable Rewards (RLVR) has become a powerful paradigm for improving language-model reasoning in domains where final answers can be checked automatically (Guo et al., 2025; Jaech et al., 2024; Team et al., 2025; Tian et al., 2026). RLVR is typically instantiated on top of PPO (Schulman et al., 2017); its group-relative variant GRPO (Shao et al., 2024) replaces per-sample value estimation with group-based advantage normalization, and DAPO (Yu

et al., 2025) further refines GRPO with techniques that stabilize training at scale. Recent work adapts this paradigm to multimodal large language models (MLLMs) by defining verifiable rewards over image-text tasks. Vision-R1 (Huang et al., 2026b) demonstrates the effectiveness of R1-style answer-level RLVR for multimodal reasoning, and a broad line of follow-up work explores rule-based RL across diverse task families, training scales, rollout strategies, and shortcut-mitigation designs (Meng et al., 2025; Chen et al., 2025; Shen et al., 2025; Liu et al., 2025b; Yang et al., 2025; Wang et al., 2025a; Liu et al., 2025a; Xia et al., 2025). A parallel direction scales long-chain visual reasoning through structured supervision (Xu et al., 2025a; Thawakar et al., 2025). These methods establish outcome-verifiable reinforcement learning as an effective post-training strategy for MLLMs, but final-answer rewards remain sparse with respect to whether the reasoning process is actually grounded in visual evidence.

B.2 Faithful Multimodal Reasoning

Faithful multimodal reasoning requires both faithful perception of task-relevant visual evidence and faithful use of that evidence in the reasoning chain. Prior work shows that MLLMs frequently hallucinate visual content or rely on insufficient visual evidence (Guan et al., 2024; Liu et al., 2023; 2024; Lu et al., 2024b), and that longer reasoning chains may amplify hallucination when not properly grounded (Xu et al., 2025b; Yao et al., 2025; Lu et al., 2025). Recent RL methods introduce perception- and grounding-aware signals: Perception-R1 (Xiao et al., 2026) rewards consistency with visual annotations, Vision-SR1 (Li et al., 2026) decomposes responses into visual descriptions and language reasoning with self-rewarding, VPPO (Huang et al., 2026a) uses token-level visual dependency to reweight optimization toward visually pivotal tokens, and PAPO (Wang et al., 2025b) adds a perception-aware regularizer to the policy objective. A complementary line couples reasoning with explicit visual operations: Ground-R1 (Cao et al., 2025) rewards region selection, and DeepEyes (Zheng et al., 2025b) learns to "think with images" by interleaving visual operations into the chain. Our work addresses both halves jointly: we strengthen perception through supervision native to the image and explicitly target the perception–reasoning disconnect, a failure mode that cuts across existing efforts in which correctly perceived evidence is dropped or contradicted during reasoning.