PERCEPTION OR REASONING? DIAGNOSING AND TACKLING RL BOTTLENECKS IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has brought strong gains to language model reasoning, but its improvements on visual tasks remain less consistent. One difficulty is that visual task failures can arise from different sources: the same benchmark may be limited by perception for one model and by reasoning for another. We introduce Perception and Reasoning Attribution (PARA), which measures the remaining perceptual and reasoning headroom of each model–benchmark pair by comparing how many current errors are recovered with more accurate visual information versus a stronger reasoner. Across 4 VLMs and 17 multimodal benchmarks, PARA broadly agrees with the expected characteristics of existing benchmarks while revealing substantial model-specific variation. More importantly, it shows that RL gains become smaller as perceptual headroom increases, but larger when reasoning explains more of the recoverable errors, suggesting that current RL exploits reasoning headroom more effectively than perceptual headroom. Motivated by this gap, we analyze how existing RL objectives affect visual perception and find that reward- and token-level signals influence visual attention only indirectly. We therefore propose GRAD, a group-relative attention distillation objective that uses the best rollout in each group to directly guide the visual attention of weaker rollouts. GRAD consistently improves performance across two model scales and seven benchmarks, and further complements existing RL methods.

1 INTRODUCTION

Reinforcement learning (RL) has become a key post-training paradigm for large language models (LLMs) and vision-language models (VLMs), producing substantial gains in mathematical, coding, and multimodal reasoning tasks (Shao et al., 2024; DeepSeek-AI et al., 2025; Huang et al., 2026b; Shen et al., 2025; Liu et al., 2025; Yu et al., 2025). Yet these gains remain less consistent on visual tasks. A prevailing explanation is that RL improves reasoning more readily than perception, leaving access to visual evidence as a persistent bottleneck (Xiao et al., 2025; Li et al., 2026c; Wu et al., 2026b). However, this observation does not reveal what limits a particular model on a particular task. As illustrated in Fig. 1, the same problem may fail because a model misperceives the necessary evidence or because it fails to reason from correctly perceived evidence. Existing analyses largely rely on benchmark characteristics, controlled interventions, or manual inspection (Wang et al., 2025; Li et al., 2026c; Wu et al., 2026b), leaving no general way to quantify how much improvement remains available from perception versus reasoning at the model–benchmark level.

Motivated by this gap, we introduce *Perception and Reasoning Attribution* (PARA), an intervention-based framework that measures how much of a model's current errors is recoverable through stronger perception versus stronger reasoning. It estimates these two headroom by selectively strengthening visual information or reasoning while preserving the other one. Across 4 VLMs and 17 benchmarks, PARA reveals broad benchmark-level trends alongside substantial model-specific variation.

Leveraging PARA, we further reveal a consistent asymmetry in visual RL: gains decrease with perceptual headroom but increase with reasoning-related headroom, suggesting that current RL exploits reasoning more effectively than perception. We show that reward- and token-level perceptual signals affect visual attention only indirectly, whereas direct attention supervision responds more systematically to perceptual headroom. Together, these results identify visual attention as a more direct optimization interface for perception.

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.

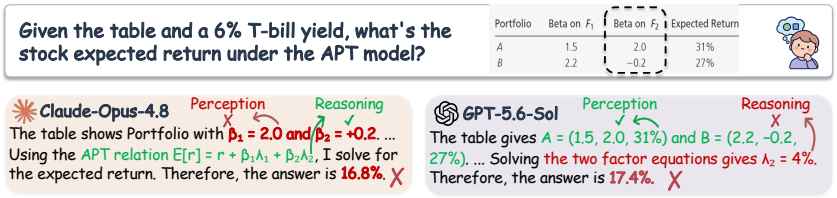


Figure 1: **The same task can expose different failure modes across models.** One model may fail to perceive the relevant evidence, while another perceives it correctly but fails to reason from it.

Direct attention supervision, however, still requires a reliable target because outcome rewards conflate perception and reasoning. We therefore propose GRAD, *Group-Relative Attention Distillation*, which uses the best sibling within each on-policy rollout group as a positive visual-attention target for weaker rollouts. Across two model scales and seven benchmarks, GRAD consistently improves performance, strengthens perception, and remains complementary to existing RL methods, without requiring external teachers or additional annotations. Our contributions are threefold:

- **Diagnosing model-specific perception and reasoning bottlenecks.** We introduce PARA, an intervention-based measure of recoverable perceptual and reasoning headroom. Across four VLMs and 17 benchmarks, PARA reveals substantial model-specific variation that cannot be inferred from benchmark labels alone.
- **Explaining perception–reasoning asymmetry of visual RL.** We show that RL gains decrease with perceptual headroom but increase with reasoning, and connect this asymmetry to indirect optimization path from reward- or token-level signals to visual attention.
- **Tackling perceptual bottlenecks with group-relative attention distillation.** We propose GRAD, which distills visual attention from the best on-policy sibling into weaker rollouts. GRAD improves performance across models and benchmarks without external teachers or additional annotations, while remaining complementary to existing RL objectives.

2 Related Work

Post-Training in VLMs. Recent work extends RL with verifiable rewards to VLMs via GRPO-style optimization with answer-based rewards, task-specific verifiers, multimodal cold starts, or structured visual reasoning traces (Shao et al., 2024; DeepSeek-AI et al., 2025; Yu et al., 2025; Huang et al., 2026b; Shen et al., 2025; Liu et al., 2025; Xia et al., 2025; Yang et al., 2025). Yet outcome feedback cannot distinguish perceptual from reasoning failures, motivating visual-description rewards, perception checklists, visual–textual grounding, and perception-centric process rewards (Xiao et al., 2025; Li et al., 2026d; Ding et al., 2025; Zhang et al., 2026a; Min et al., 2026), perception-first curricula (Wu et al., 2026a), visual-dependency-aware trajectory and token weighting (Huang et al., 2026a; Wang et al., 2026a; Li et al., 2026b), or regularization under altered visual evidence (Wang et al., 2026b; Zhang et al., 2026b; Rezaei et al., 2026). Controlled studies likewise find reasoning improves more readily than perception, linking the gap to weak reward–perception coupling (Wu et al., 2026b). Reward- and token-level methods act through returns, policy gradients, or output distributions, complementing our attention-level objective; we diagnose model–benchmark bottlenecks and directly optimize visual attention. Complementary work intervenes in visual representations or supervises visual-tokens. (Song et al., 2026b; Li et al., 2026c; Song et al., 2026a; Li et al., 2026a; Hu et al., 2026) We instead derive attention targets from on-policy rollout groups.

On-Policy Distillation. Conventional knowledge distillation matches student output distributions or sequences to those of a teacher on fixed data. (Hinton et al., 2015; Kim & Rush, 2016) MiniLLM and GKD instead distill teacher logits on student-generated trajectories to reduce train–inference shift (Gu et al., 2024; Agarwal et al., 2024). Extending to VLMs, VOLD transfers reasoning from a text-only teacher (Bousselham et al., 2026). Vision-OPD uses a crop-conditioned privileged teacher (Yuan et al., 2026), and VA-OPD and ViCuR focus on visually dependent tokens or recoverable visual cues (Liu et al., 2026; Tian et al., 2026). Attention complements logit distillation by exposing how outputs use visual evidence. OPAD matches teacher logits and attention distributions on student rollouts (Li et al., 2026a). Related methods optimize visual-token routing or align attention with annotated evidence regions (Song et al., 2026a; Hu et al., 2026). Most rely on expert teachers or annotated or privileged data, such as reasoning traces, evidence regions, crops, or visual cues. Ours instead uses within-group targets without separate teachers or extra annotations.

3 DIAGNOSING PERCEPTION AND REASONING BOTTLENECKS IN VISION-LANGUAGE MODEL RL

Visual task failures can arise from insufficient perception, insufficient reasoning, or both. Importantly, the dominant limitation is not determined by the benchmark alone: the same task may expose different bottlenecks for different models. We therefore introduce PARA, a model–benchmark-specific framework for measuring how much of a model's current error set is recoverable through stronger perception versus stronger reasoning. We first define these two forms of recoverable headroom, then validate the description-based proxy underlying the measurement, and finally use PARA to characterize how perception and reasoning bottlenecks vary across models and benchmarks.

3.1 PARA: PERCEPTION AND REASONING ATTRIBUTION

Visual Benchmarks are often characterized by the capabilities they are designed to assess, but such labels do not identify what limits a particular model. For example, one model may fail on a visual reasoning problem because it misperceives a relevant object, whereas another may perceive the same evidence correctly but fail to reason from it. We therefore define PARA at the model–benchmark level and measure two intervention-based quantities: *perceptual headroom*, corresponding to failures recoverable with stronger visual information while retaining the target model's own reasoning, and *reasoning headroom*, corresponding to failures recoverable with stronger reasoning while retaining the target model's own perceived visual information. Fig. 2(a) illustrates this construction.

Formally, let $F^{M,B}$ denote the set of examples in benchmark B that model M answers incorrectly, and let $n_F^{M,B} = |F^{M,B}|$. We estimate the two forms of headroom by selectively strengthening one capability while preserving the other as much as possible.

Perceptual Intervention. To estimate perceptual headroom, a stronger reference model $M^{o,i}$ generates a detailed description of the image without seeing the question. The target model M then answers the original question using this additional visual information. The subset of current failures corrected under this intervention is denoted

$$P_i^{M,B} \subseteq F^{M,B}. \tag{1}$$

To reduce sensitivity to information loss or noise in the description, we evaluate the target model both from the description alone and from the original image augmented with the same description, and count a failure as recovered if either condition succeeds.

Reasoning Intervention. To estimate reasoning headroom, we reverse the intervention. The target model M first generates its own question-blind description of the image, which serves as a proxy for the visual information it has retained. The stronger reference model $M^{o,i}$ then answers the original question using this description. The corresponding recovered failures form

$$R_i^{M,B} \subseteq F^{M,B}. \tag{2}$$

We include an example in $R_i^{M,B}$ only when the reference model can also answer it correctly from the original image. This solvability check prevents failures beyond the reference model's own capability from being incorrectly attributed to the target model's reasoning. Using multiple reference models, we aggregate the recovered examples as

$$P^{M,B} = \bigcup_{i=1}^{|\mathcal{M}^o|} P_i^{M,B}, \qquad R^{M,B} = \bigcup_{i=1}^{|\mathcal{M}^o|} R_i^{M,B}. \tag{3}$$

We use the term *reference model* operationally: these models provide stronger intervention capabilities but are not assumed to be infallible. In our experiments, we instantiate $\mathcal{M}^o$ with two frontier APIs, `Claude-Opus-4.8` and `GPT-5.6-Luna` The framework naturally extends to additional or stronger reference models. We then define the perceptual and reasoning headroom as

$$\mathrm{PARA}_{\mathrm{PH}}^{M,B} = \frac{|P^{M,B}|}{n_F^{M,B}}, \qquad \mathrm{PARA}_{\mathrm{RH}}^{M,B} = \frac{|R^{M,B}|}{n_F^{M,B}}. \tag{4}$$

Where $\mathrm{PARA}_{\mathrm{PH}}^{M,B}$ measures the fraction of the target model's current failures recoverable through stronger perceptual information, and $\mathrm{PARA}_{\mathrm{RH}}^{M,B}$ measures the fraction recoverable through stronger reasoning given the visual information retained by the target model. These quantities measure *recoverable headroom*, rather than assigning an intrinsic perception or reasoning label to the benchmark.

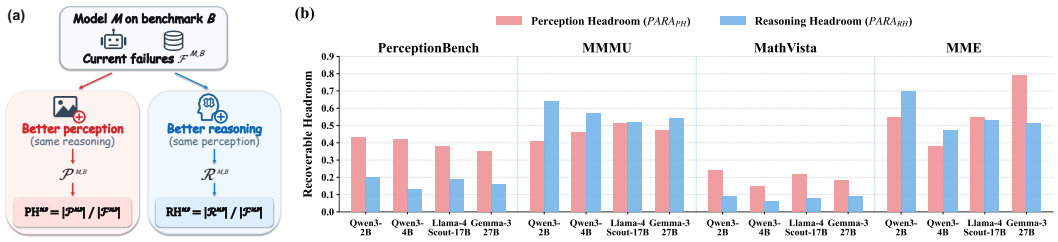


Figure 2: **(a) Illustration of PARA.** Starting from a model's current failures, PARA measures the fractions recoverable by selectively strengthening perception or reasoning. **(b) Representative Headroom Measurements.** Results exhibit broad benchmark-level trends but substantial model-specific variation, showing that effective bottlenecks cannot be inferred by benchmark labels alone.

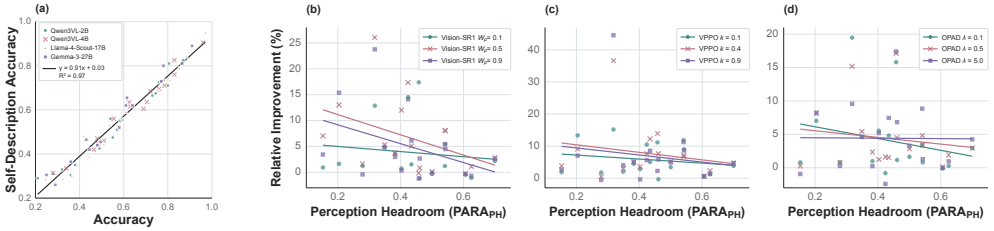


Figure 3: **(a) PARA validation.** Self-description accuracy closely tracks direct image accuracy. **(b)(c)(d) Perceptual supervision comparison.** Only direct attention supervision exhibits a coefficient-dependent response to perceptual headroom $\mathrm{PARA}_{\mathrm{PH}}$.

3.2 Validating Question-Blind Descriptions as a Perception Proxy.

The reasoning intervention relies on the target model's question-blind description as a proxy for the visual information it has extracted from the image. Because textual descriptions may discard visual information and thereby impair downstream reasoning (Wang et al., 2023), we evaluate how faithfully this representation preserves the target model's effective perceptual capability. Specifically, for each model and benchmark, we compare two conditions using the same question-answering instruction. In the *direct-image* condition, the model answers directly from the original image. In the *self-description* condition, the model first generates a question-blind description and subsequently answers the question from that description alone. If the description preserves most of visual information available to the model, performance in two conditions should closely correlate one another.

As shown in Fig. 3(a), self-description accuracy closely follows direct-image accuracy across models and benchmarks. Pooling the model–benchmark pairs yields a fitted slope above 0.90 with $R^2 = 0.97$, while the mean accuracy decrease across benchmarks ranges from only 0.1 to 3.4 points across the four models. These results indicate that question-blind descriptions preserve sufficient visual information to serve as an reliabe proxy for the perceptual capability of the evaluated VLMs.

3.3 Bottlenecks Are Model–Benchmark Specific

PARA reveals broad benchmark-level trends, but also substantial variation across models on the same benchmark. Fig. 2(b) shows representative results, while complete measurements across all 17 benchmarks and four models are provided in Appx. A.

At a coarse level, the measured headroom is consistent with the intended characteristics of several benchmarks. For example, PerceptionBench exhibits greater perceptual than reasoning headroom, whereas MMMU generally exhibits greater reasoning headroom as expected. However, benchmark characteristics do not determine the model-specific bottleneck. On MME, reasoning headroom is larger for both Qwen models, perceptual and reasoning headroom are comparable for Llama-4-Scout-17B, and perceptual headroom is substantially larger for Gemma-3-27B. Thus, even on the same benchmark, different models can leave qualitatively different opportunities for improvement.

The mismatch can be even more pronounced for benchmarks conventionally viewed as reasoning-oriented. On MathVista, perceptual headroom exceeds reasoning headroom for all four evaluated models despite the benchmark's emphasis on multimodal reasoning. This implies that a model's failures on a reasoning benchmark might not primarily reflect its insufficient reasoning capabilities.

Together, these results indicate that perception and reasoning headroom depend jointly on the model and benchmark, and bottlenecks should instead be diagnosed at the model–benchmark level.

4 Understanding the Perception–Reasoning Asymmetry in Visual RL

Having established that perception and reasoning headroom vary substantially across model–benchmark pairs, we next ask whether these diagnosed bottlenecks help explain why reinforcement learning produces uneven gains across visual tasks.

4.1 RL Exploits Reasoning Headroom More Than Perceptual Headroom

Reinforcement learning has produced substantial gains in language-model reasoning (Shao et al., 2024; Yu et al., 2025), yet improvements in visual perception remain more limited and inconsistent (Xiao et al., 2025; Li et al., 2026c; Wu et al., 2026b). Prior work has observed that post-training improves reasoning more readily than perception, but has primarily characterized this asymmetry through manual error analysis or controlled interventions. With PARA, we can instead relate RL gains to model–benchmark-specific perception and reasoning headroom under a common measurement framework. To characterize the relative contribution of reasoning among recoverable failures, we define the directional reasoning bottleneck index

$$\mathrm{PARA}_{\mathrm{BI}_R}^{M,B} = \frac{|R^{M,B}|}{|P^{M,B} \cup R^{M,B}|}. \tag{5}$$

This quantity measures how broadly the reasoning intervention covers the union of failures recoverable through either intervention. Because $P^{M,B}$ and $R^{M,B}$ may overlap, $\mathrm{PARA}_{\mathrm{BI}_R}$ should be interpreted as a directional coverage measure rather than as a decomposition of the error set.

We compare the relative improvement of model after RL over its corresponding base model with two quantities measured on base model: perceptual headroom $\mathrm{PARA}_{\mathrm{PH}}$ and reasoning bottleneck index $\mathrm{PARA}_{\mathrm{BI}_R}$. As in Fig. 4, the same directional trends hold across both RL algorithms and model scales. RL gains decrease as $\mathrm{PARA}_{\mathrm{PH}}$ increases, but increase with $\mathrm{PARA}_{\mathrm{BI}_R}$. In other words, RL tends to yield smaller improvements when a larger fraction of the base model's failures is recoverable through stronger perception, while yielding larger improvements when reasoning intervention covers more of the recoverable failure set.

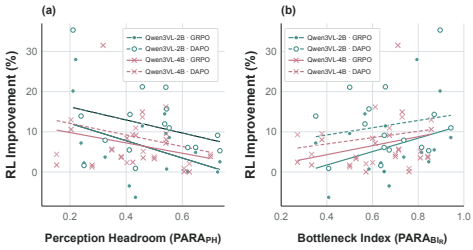


Figure 4: **Asymmetric RL gains.** Across models and algorithms , RL gains **(a)** decrease with perceptual headroom $\mathrm{PARA}_{\mathrm{PH}}$ and **(b)** increase with the reasoning bottleneck index $\mathrm{PARA}_{\mathrm{BI}_R}$.

We interpret these relationships as cross-benchmark directional trends rather than predictions for individual tasks. Nevertheless, their consistency across algorithms and models reveals a clear asymmetry: current RL objectives appear to exploit reasoning headroom more effectively than perceptual.

4.2 Existing RL Objectives Indirectly Optimize Visual Attention

Existing methods that target visual perception during RL primarily modify either the reward assigned to a rollout or the allocation of policy updates across output tokens (Li et al., 2026d; Huang et al., 2026a). Although these approaches differ operationally, both ultimately modify the effective advantage used by the policy-gradient objective. We therefore examine how such advantage-level signals propagate to the model's visual-attention pathway.

Given a rollout τ and output token t, let $\rho_t(\theta)$ denote the importance ratio and $\widetilde{A}_t(\tau)$ the effective advantage. For simplicity, we write the GRPO objective as

$$\mathcal{L}(\theta) = \mathbb{E}_{\tau,t}\left[\ell\Big(\rho_t(\theta), \widetilde{A}_t(\tau)\Big)\right]. \tag{6}$$

Reward-level methods modify the value of $\widetilde{A}_t$, while token-level methods redistribute it across output tokens. In either case, $\widetilde{A}_t$ is treated as fixed during a policy update.

Within a VLM attention block, let visual contribution to the representation of generated token t be

$$h_t^{\mathrm{vis}} = \sum_{j \in \mathcal{V}} a_{t,j} v_j, \tag{7}$$

where v_j denotes the value representation of visual token j and $a_{t,j}$ the attention from output token t to visual token j. Collecting the visual representations and attention weights as h^{vis} and a, respectively, the loss along this pathway has the composition $\mathcal{L}(z(h^{\text{vis}}(a)); \widetilde{A})$. By the chain rule,

$$\nabla_a \mathcal{L} = \left(\frac{\partial h^{\text{vis}}}{\partial a}\right)^\top \left(\frac{\partial z}{\partial h^{\text{vis}}}\right)^\top \nabla_z \mathcal{L}. \tag{8}$$

Eq. 8 makes the optimization path to visual attention explicit. The term $\nabla_z \mathcal{L}$ determines the update to the output logits, $\partial z / \partial h^{\text{vis}}$ propagates this signal to the visual contribution, and $\partial h^{\text{vis}} / \partial a$ further propagates it to the attention allocation. Thus, reward- and token-level perceptual signals can influence visual attention, but only through their downstream effect on the output logits. The objective contains no separate term that directly evaluates or supervises the visual-attention distribution itself.

We examine this distinction empirically using Vision-SR1 (Li et al., 2026d) and VPPO (Huang et al., 2026a) as representative reward-level and token-level methods, respectively. Vision-SR1 rescales the rollout advantage using a perceptual reward weight W_p, whereas VPPO redistributes token-level advantages through a trajectory weight $\alpha(\tau)$ and a visual mask $m_t(k)$. Both therefore alter the strength or allocation of $\widetilde{A}_t$ without directly supervising visual attention.

As shown in Fig. 3(b) and 3(c), strengthening these perceptual modifications, either by increasing W_p for Vision-SR1 or decreasing the mask-token ratio k for VPPO, does not systematically produce larger gains on benchmarks with greater perceptual headroom PARA_{PH}. Meanwile, increasing the strength of either intervention does not consistently rotate the fitted relationship toward a more positive slope. These results do not imply that advantage-level methods fail to improve overall performance; rather, they suggest that increasing their perceptual signal does not systematically translate into greater benefit on tasks with more available perceptual headroom.

4.3 Visual Attention Provides a Direct Optimization Interface

The preceding analysis suggests that directly supervising visual attention provides a more immediate optimization path to the model's visual pathway. Suppose we augment the RL objective with an auxiliary regularizer $\mathcal{R}(a)$ on the attention distribution, weighted by λ. The resulting gradient with respect to attention becomes

$$\nabla_a \mathcal{L}^a = \left(\frac{\partial h^{\text{vis}}}{\partial a}\right)^\top \left(\frac{\partial z}{\partial h^{\text{vis}}}\right)^\top \nabla_z \mathcal{L} + \lambda \nabla_a \mathcal{R}(a). \tag{9}$$

Unlike advantage-level supervision, the additional term $\lambda \nabla_a \mathcal{R}(a)$ acts directly on the attention distribution rather than reaching it only through the output logits.

We examine this direct pathway using On-Policy Attention Distillation (OPAD) from RAL (Li et al., 2026a). Let p_t^θ denote the current visual-attention distribution for token t, and let p_t^{old} denote the corresponding distribution recorded during rollout generation. OPAD uses the regularizer

$$\mathcal{R}(a) = \mathbb{E}_{\tau,t}\left[\widetilde{A}_t(\tau) D_{\text{JS}}\left(p_t^\theta \,\|\, p_t^{\text{old}}\right)\right], \tag{10}$$

where D_{JS} denotes the Jensen–Shannon divergence. Positive advantages encourage preservation of attention patterns associated with successful rollouts, whereas negative advantages push the current policy away from attention patterns associated with unsuccessful rollouts.

As shown in Fig. 3(d), increasing the coefficient λ systematically changes the relationship between performance gain and perceptual headroom: the fitted slope moves from negative toward approximately flat, reducing the performance disadvantage associated with larger PARA_{PH}. Together, the theoretical and empirical results suggest that visual attention provides a more direct interface for addressing the perceptual headroom identified by PARA.

5 GRAD: Group-Relative Attention Distillation for Perception

Sec. 4.3 identifies visual attention as a direct optimization interface for addressing perceptual bottlenecks. However, direct attention supervision alone does not guarantee a reliable update direction. Existing attention-level objectives remain coupled to rollout outcomes: they preserve the old-policy attention of successful rollouts while pushing unsuccessful rollouts away from their previous attention patterns. This coupling is ambiguous because an outcome reward jointly reflects perception and

reasoning. A rollout may fail because of incorrect reasoning despite attending to the relevant visual evidence, in which case pushing away from its attention pattern could penalize valid perception.

Consistent with this limitation, while OPAD changes the perception-dependent trend in Fig. 3(d), it still yields only moderate overall gains relative to other methods in Fig. 3(b)(c). These observations suggest that attention provides the appropriate optimization interface, but effective perceptual supervision requires a more careful objective design.

We therefore propose GRAD, *Group-Relative Attention Distillation*. Rather than using each rollout's own old-policy attention as its reference, GRAD constructs a positive attention target from the best sibling within the same on-policy rollout group. For a rollout group g generated from the same image and prompt, we select

$$b(g) = \arg\max_{j \in g_{\mathrm{v}}} r_j, \tag{11}$$

where r_j is the reward of rollout j and g_{v} denotes the subset of rollouts with valid attention distributions. When multiple siblings attain the same maximal reward, we select the rollout with the shortest trajectory. We then treat negative-advantage rollouts as students and distill the best sibling's visual-attention pattern into them. For rollout i, we define the adaptive supervision weight

$$w_i = \max(-A_i, 0), \tag{12}$$

such that supervision is activated only when $A_i < 0$ and becomes stronger as the advantage decreases. Groups with identical rewards are therefore naturally excluded because they contain no negative-advantage students. Importantly, w_i is always nonnegative, so minimizing the distillation objective consistently pulls a student's attention toward the best sibling rather than pushing it away from its own previous attention. This converts outcome feedback into a positive attention target without requiring an external teacher or additional annotation.

To focus the auxiliary objective specifically on perception, we perform alignment only over visual tokens. For rollout i, let $\vec{p}_i^{\,\theta}$ denote its visual-attention distribution, obtained by renormalizing attention over visual tokens and pooling across response tokens and attention heads. Because sibling rollouts share the same visual input, their resulting visual-attention distributions are directly comparable even when their generated responses have different lengths. Let $\widehat{p}_g$ denote the corresponding visual-attention distribution of the best sibling $b(g)$. The GRAD objective is

$$\mathcal{L}_{\mathrm{GRAD}} = \frac{1}{|\mathcal{I}|} \sum_{(g,i)\in\mathcal{I}} w_i D_{\mathrm{JS}}\left(\vec{p}_i^{\,\theta} \,\middle\|\, \widehat{p}_g\right), \tag{13}$$

where $\mathcal{I}$ denotes the set of valid negative-advantage students.

The complete training objective combines standard GRPO objective with outcome reward and group-relative attention distillation tailored for mitigating perception bottleneck:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{GRPO}} + \lambda \mathcal{L}_{\mathrm{GRAD}}, \tag{14}$$

where λ controls the strength of attention supervision. By constructing perceptual supervision directly from on-policy sibling rollouts, GRAD requires no external teacher, additional annotations, extra rollouts, or inference-time modifications.

6 Experiments

6.1 Experimental Setup

Dataset and Benchmarks. Following VisionSR1 (Li et al., 2026d), we use `Vision-SR1-47K` as the training dataset for RL. We evaluate on 7 visual benchmarks, including MME (Fu et al., 2023), MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), MATH-Vision (Wang et al., 2024), PerceptionBench (Lin et al., 2026), ScienceQA (Lu et al., 2022), and VisNumBench (Weng et al., 2025). We report MME on its original scale and normalize it to a 100-point scale only when computing the average score. We follow the official evaluation protocol and metric for each benchmark.

Baselines. Beyond GRPO (Shao et al., 2024), we compare against baselines spanning diverse visual supervision. VisionSR1 (Li et al., 2026d) introduces a perception reward, VPPO (Huang et al., 2026a) reallocates policy updates across response tokens, and OPAD (Li et al., 2026a) directly regularizes visual attention with old policy. We additionally use DAPO (Yu et al., 2025) to test whether our method complements other RL algorithms.

Table 1: **Results across seven visual benchmarks.** MME is reported on its native scale, while Avg. rescales it with $1/28$. The best and second-best results are shown in **bold** and underline, respectively.

	Method	MME	MMMU	MMMU-Pro	MathVis	PercepBench	ScienceQA	VisNum.	Avg.
Qwen3-VL-2B	Base	2044.0	46.6	30.8	13.5	22.7	86.1	35.1	44.0
	GRPO	2217.6	49.1	34.3	18.1	21.3	88.2	35.3	46.5
	Vision-SR1	2223.2	48.3	34.8	20.7	23.7	87.5	**41.1**	47.9
	VPPO	2212.0	50.8	33.8	19.1	22.8	88.2	37.6	47.3
	OPAD	2256.8	49.6	32.9	16.1	20.3	87.1	36.3	46.1
	Ours	**2290.4**	**52.1**	**35.1**	**21.1**	**24.4**	**90.3**	40.8	**49.4**
Qwen3-VL-4B	Base	2248.4	60.1	45.6	30.3	24.8	91.7	49.7	54.6
	GRPO	2240.0	62.1	48.4	37.1	25.7	93.9	52.1	57.0
	Vision-SR1	2332.4	60.6	47.9	38.2	**29.1**	94.3	53.7	58.2
	VPPO	2315.6	62.9	48.6	35.5	25.9	94.3	52.9	57.5
	OPAD	2279.2	63.1	48.0	37.8	25.1	94.4	51.4	57.3
	Ours	**2349.2**	**64.3**	**49.8**	**40.5**	27.9	**95.0**	**55.0**	**59.5**

Table 2: **Complementarity with existing methods.** Each entry reports the performance integrating GRAD to corresponding baseline on Qwen3-VL-2B. The largest gain at each benchmark is **bolded**.

Method	MME	MMMU	MMMU-Pro	MathVis	PercepBench	ScienceQA	VisNum.	Avg.
Vision-SR1 w/ Ours	+44.8	**+2.8**	+0.2	**+4.3**	+0.8	**+2.4**	**+2.6**	**+2.1**
VPPO w/ Ours	+19.6	+0.5	+0.5	+1.3	-0.9	+1.6	+1.3	+0.7
DAPO w/ Ours	**+123.2**	-0.2	**+0.6**	+2.3	**+1.7**	+2.3	+1.8	+1.9

Implementation Details. We use *Qwen3-VL-2B-Instruct* and *Qwen3-VL-4B-Instruct* (Bai et al., 2025) and perform full fine-tuning on four NVIDIA H200 GPUs with random seed 2026. We use a global batch size of 128 and sample eight rollouts per prompt. Details are provided in Appx. B.

6.2 Experimental Results

Main Results. In Tab. 1, we compare GRAD with baselines across seven visual benchmarks. GRAD achieves the highest average performance for both models, outperforming GRPO by 2.9 and 2.5 points, respectively. On MME, GRAD improves over GRPO by 72.8 and 109.2 points, respectively. OPAD performs comparably to GRPO overall, whereas GRAD improves upon OPAD by 3.3 and 2.2 points, highlighting the benefit of our group-relative design. Moreover, although VPPO and Vision-SR1 improve through objectives orthogonal to ours, GRAD exceeds both by at least 1.5 points at 2B and 1.3 points at 4B on average, demonstrating our overall effectiveness.

Analysis of Perceptual Improvement. In Fig. 5(b), we further visualize inference attention in representative cases from PerceptionBench and VisNumBench. In the upper case, GRAD shifts attention from the fish's eye to its tail, and in the lower case, it focuses on Object A rather than the irrelevant Object D, enabling the correct volume comparison. These cases provide qualitative evidence that GRAD improves perception by direct visual attention supervision.

Given GRAD aiming to improve perception, we further investigate how it changes the residual headroom. As shown in Fig. 5(a), GRAD reduces $\mathrm{PARA}_{\mathrm{PH}}$ on five benchmarks, while GRPO reduces only two. On MMMU-Pro, PerceptionBench, and ScienceQA, GRAD reverses the increase into a reduction. Across the remaining four benchmarks, GRAD further increases $\mathrm{PARA}_{\mathrm{RH}}$ by a substantially larger margin, indicating that stronger perception preserves more usable visual evidence and allows the same reasoning intervention to recover a larger fraction of failures. Together, the reduction in perceptual headroom and increase in reasoning headroom provide complementary evidence that GRAD directly strengthens perception during RL.

Complementarity with Existing Methods. Furthermore, we evaluate whether GRAD can be combined with existing methods. As shown in Tab. 2, integrating our objective improves the average scores of Vision-SR1, VPPO, and DAPO by 2.1, 0.7, and 1.9 points, respectively. The combinations improve all seven benchmarks for Vision-SR1 and six of seven for both VPPO and DAPO. These gains demonstrate that our attention-level objective complements methods at the reward, token, and policy levels, and can be readily integrated with existing methods.

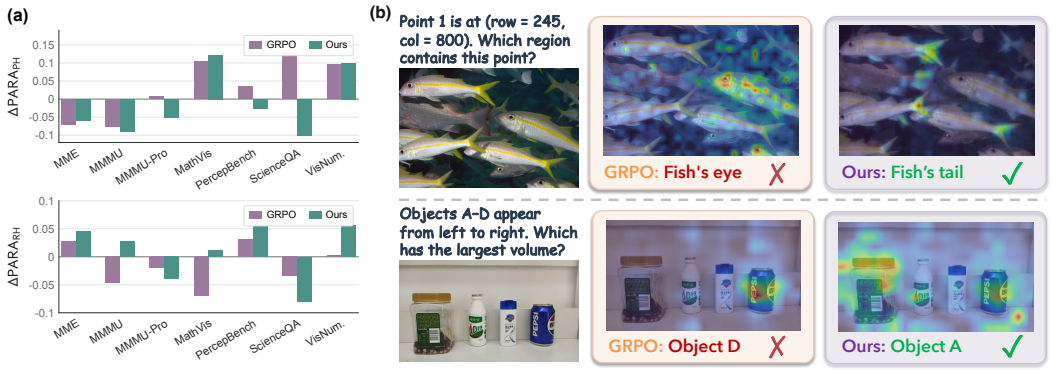


Figure 5: **(a) Changes in perception and reasoning headroom relative to Qwen3-VL-4B.** GRAD reduces perceptual headroom consistently over GRPO while making more residual errors recoverable through reasoning intervention. **(b) Visualization of visual attention map.** GRAD focuses on the correct visual evidence while GRPO fails.

Table 3: **Ablation study results.** Both components work jointly to achieve better performance.

Method	MME	MMMU	MMMU-Pro	MathVis	PercepBench	ScienceQA	VisNum.	Avg.
Ours	**2290.4**	52.1	**35.1**	21.1	**24.4**	**90.3**	**40.8**	**49.4**
- w/ Attention Mass.	2265.2	50.7	34.6	**21.2**	21.7	88.8	40.2	48.3
- w/o Percep. Only	2228.8	**52.7**	34.2	19.1	21.2	90.1	38.2	47.9
- w/o Best Sibling	2284.8	50.7	32.4	13.5	19.1	87.3	33.6	45.5

6.3 Ablation Studies

Method Components. Tab. 3 examines each component of GRAD. On Qwen3-VL-2B, removing the perception-only restriction lowers the average score by 1.5 points, yet the resulting variant remains above GRPO, indicating that our objective supplies the perceptual signal missing from plain RL. Further removing best sibling supervision, randomly choosing a target sibling instead, reduces the score to 45.5, confirming the importance of a reliable group-relative target. In addition, we evaluate the performance of keeping only the tokens attribute to top 50% attention in the best sibling. Adding this constraints has only modest effects on most benchmarks and lowers the average by 1.1 points. Together, these results support our perceptual and group-relative attention design.

Hyper-parameters. Following Li et al. (2026a), we set $\lambda = 1.0$ and supervise only the last decoder layer throughout our experiments. Setting $\lambda = 0.2$ or $\lambda = 5.0$ lowers the average score by about 2 points, although both configurations remain better than GRPO. Extending supervision to the top 4 or 8 layers also underperforms the last layer setting, while applying it to the top 16 or all layers leads to substantially larger performance drops. These denser targets appear more difficult to optimize, suggesting that effective multi-layer supervision requires a dedicated design. Detailed results of ablation study are provided in Appx. B.2.

7 Conclusion

In this work, we examine how much perceptual and reasoning headroom remains for VLMs across visual benchmarks, and how these two forms of headroom relate to the gains from RL post-training. We introduce PARA, a quantitative measure that estimates, for each model–benchmark pair, how many current errors can be recovered with better visual information versus better reasoning. Across diverse models and benchmarks, PARA reveals substantial model-specific variation and a consistent asymmetry in RL: gains decrease with greater perceptual headroom but increase as reasoning accounts for more recoverable errors. We further show that reward- and token-level objectives affect visual attention only indirectly, motivating more direct attention-level supervision. Accordingly, we propose GRAD, a group-relative attention distillation objective that consistently improves performance and remains complementary to existing RL methods. Together, our results connect model-specific bottleneck measurement with targeted optimization for improving visual RL.

AI USE STATEMENT

We used generative AI tools solely for language polishing and improving the clarity and readability of the manuscript. We did not use generative AI tools for automated research discovery, formulation of research ideas, design of experiments, analysis of results, or generation of scientific claims.

Additionally, generative AI tools were used to assist with code writing during implementation. All AI-assisted code was manually reviewed line by line by the authors, tested for correctness, and verified against the intended algorithms and experimental settings.

We have carefully reviewed all AI-assisted content and take full responsibility for the final content of this work, including all text, claims, implementations, and experimental artifacts.

ETHICS STATEMENT

This work does not involve human subjects, sensitive personal data, or applications that raise additional ethical concerns. We follow standard research integrity practices and use publicly available benchmarks for evaluation.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed PARA analysis framework and GRAD training algorithm in Sec. 5. Additional implementation details, training configurations, ablation settings, and experimental protocols are provided in the Appendix. We will release the source code and configuration files to facilitate reproduction of our results.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: Reasoning transfer from LLMs to vision-language models via on-policy distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yizhuo Ding, Mingkang Chen, Zhibang Feng, Tong Xiao, Wanying Qu, Wenqi Shao, and Yanwei Fu. VTPerception-R1: Enhancing multimodal reasoning via explicit visual and textual perceptual grounding. *arXiv preprint arXiv:2509.24776*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-policy distillation of large language models. In *International Conference on Learning Representations*, 2024.

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*, pp. 14375–14385, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ruina Hu, Chen Wang, Lai Wei, Jionghao Bai, Bin Yu, Weiran Huang, Kai Wang, and Yue Wang. Attend to evidence: Evidence-anchored spatial attention supervision for multimodal RLVR. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, 2026a.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, 2026b.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Bangzheng Li, Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Xingyu Fu, Muhao Chen, and Derek Zhiyuan Cheng. Reinforced attention learning. *arXiv preprint arXiv:2602.04884*, 2026a.

Qiming Li, Tianlun Li, Xiaolong Cheng, Hangyu Li, Ruiyan Gong, Kangning Niu, Kaitao Jiang, and Mu Xu. PRPO: Perception-reinforced policy optimization via token-level dynamic advantage reshaping. *arXiv preprint arXiv:2606.08708*, 2026b.

Xirui Li, Ming Li, and Tianyi Zhou. What does RL improve for visual reasoning? a frankenstein-style analysis. *arXiv preprint arXiv:2602.12395*, 2026c.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. Self-rewarding vision-language model via reasoning decomposition. In *International Conference on Learning Representations*, 2026d.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. PerceptionBench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision*, 2024.

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, 2025.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

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

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Meta AI Blog, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving vision-language models with perception-centric process reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Ahmad Rezaei, Mohsen Gholami, Saeed Ranjbar Alvar, Kevin Cannons, Mohammad Asiful Hossain, Weimin Zhou, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. CPPO: Contrastive perception for vision language policy optimization. *arXiv preprint arXiv:2601.00501*, 2026.

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. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Junha Song, Byeongho Heo, Geonmo Gu, Jaegul Choo, Dongyoon Han, and Sangdoo Yun. Learning to see what you need: Gaze attention for multimodal large language models. *arXiv preprint arXiv:2605.13080*, 2026a.

Junha Song, Sangdoo Yun, Dongyoon Han, Jaegul Choo, and Byeongho Heo. RL makes MLLMs see better than SFT. In *International Conference on Learning Representations*, 2026b.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual cues as recoverable privilege for multimodal on-policy distillation. *arXiv preprint arXiv:2606.05718*, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with the MATH-Vision dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Xinhe Wang, Jin Huang, Xingjian Zhang, Tianhao Wang, and Jiaqi W. Ma. Your reasoning benchmark may not test reasoning: Revealing perception bottleneck in abstract reasoning benchmarks. *arXiv preprint arXiv:2512.21329*, 2025.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-guided policy optimization for multimodal reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 6646–6664, 2026a.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026b.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. Filling the image information gap for VQA: Prompting large language models to proactively ask questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2874–2890, 2023.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. VisNumBench: Evaluating number sense of multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Juncheng Wu, Hardy Chen, Haoqin Tu, Xianfeng Tang, Freda Shi, Hui Liu, Hanqing Lu, Cihang Xie, and Yuyin Zhou. From seeing to thinking: Decoupling perception and reasoning improves post-training of vision-language models. In *International Conference on Machine Learning*, 2026a.

Xueqing Wu, Yu-Chi Lin, Kai-Wei Chang, and Nanyun Peng. On asymmetric optimization of reasoning and perception in vision-language model post-training. *arXiv preprint arXiv:2605.29496*, 2026b.

xAI. Grok-1.5 Vision Preview. xAI Blog, 2024. URL `https://x.ai/news/grok-1.5v`.

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. *arXiv preprint arXiv:2506.07218*, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. WeThink: Toward general-purpose vision-language reasoning via reinforcement learning. *arXiv preprint arXiv:2506.07905*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Zhixiong Zeng, Siqi Yang, Peng Shi, Lin Ma, and Jing Zhang. Perceptual-evidence anchored reinforced learning for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41111–41120, 2026a.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. Consistent paths lead to truth: Self-rewarding reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.08745*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Proceedings of the European Conference on Computer Vision*, 2024.

Shuoshuo Zhang, Yizhen Zhang, Jingjing Fu, Lei Song, Jiang Bian, Yujiu Yang, and Rui Wang. See less, see right: Bi-directional perceptual shaping for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

A PARA RESULTS

A.1 FULL BENCHMARK RESULTS

Tab. 4 reports the full PARA results for all 17 benchmarks across the four models evaluated in Sec. 3.1, including Qwen3-VL-2B and Qwen3-VL-4B (Bai et al., 2025), Llama-4-Scout-17B (Meta AI, 2025), and Gemma-3-27B (Gemma Team et al., 2025).

Table 4: **Full PARA results across models and benchmarks.** PH and RH denote perception and reasoning headroom, and BI_P and BI_R denote their directional bottleneck indices. The complete results further support the benchmark-level trends alongside distinct patterns across models.

Benchmark	Qwen3-VL-2B				Qwen3-VL-4B				Llama-4-Scout-17B				Gemma-3-27B			
	PH	RH	BI_P	BI_R	PH	RH	BI_P	BI_R	PH	RH	BI_P	BI_R	PH	RH	BI_P	BI_R
BLINK	0.54	0.35	0.77	0.50	0.54	0.24	0.86	0.38	0.62	0.26	0.96	0.40	0.61	0.39	0.85	0.54
VisNumBench	0.54	0.42	0.76	0.58	0.54	0.42	0.79	0.61	0.53	0.46	0.74	0.64	0.60	0.44	0.81	0.59
RealWorldQA	0.54	0.31	0.97	0.57	0.46	0.46	0.73	0.73	0.50	0.33	0.86	0.57	0.67	0.42	0.98	0.61
PerceptionBench	0.43	0.20	0.89	0.41	0.42	0.13	0.87	0.27	0.38	0.19	0.86	0.42	0.35	0.16	0.80	0.37
MMStar	0.62	0.49	0.83	0.65	0.50	0.32	0.83	0.53	0.47	0.31	0.81	0.53	0.58	0.36	0.88	0.56
ChartQA	0.33	0.35	0.68	0.74	0.35	0.35	0.71	0.71	0.23	0.13	1.00	0.57	0.46	0.35	0.91	0.70
HallusionBench	0.73	0.55	0.86	0.65	0.62	0.42	0.86	0.57	0.58	0.30	0.93	0.48	0.68	0.39	0.94	0.53
MME	0.55	0.70	0.74	0.94	0.38	0.47	0.68	0.84	0.55	0.53	0.88	0.83	0.79	0.51	0.98	0.63
MMBench	0.65	0.65	0.81	0.81	0.61	0.58	0.80	0.76	0.48	0.36	0.92	0.69	0.46	0.38	0.85	0.69
ScienceQA	0.73	0.73	0.85	0.85	0.70	0.70	0.70	0.70	0.90	0.60	1.00	0.60	0.46	0.54	0.60	0.70
MM-Vet	0.25	0.17	0.91	0.64	0.28	0.16	0.85	0.50	0.33	0.22	0.78	0.53	0.30	0.25	0.84	0.71
MMMU	0.41	0.64	0.56	0.87	0.46	0.57	0.65	0.81	0.51	0.52	0.75	0.77	0.47	0.54	0.69	0.78
MMMU-Pro	0.46	0.45	0.67	0.66	0.43	0.41	0.72	0.67	0.32	0.40	0.58	0.73	0.31	0.45	0.52	0.77
LogicVista	0.41	0.41	0.67	0.67	0.40	0.31	0.78	0.60	0.55	0.39	0.75	0.54	0.60	0.23	0.86	0.33
MathVista	0.24	0.09	0.92	0.35	0.15	0.06	0.83	0.33	0.22	0.08	0.94	0.32	0.18	0.09	0.90	0.43
MathVerse	0.21	0.34	0.56	0.90	0.20	0.29	0.60	0.86	0.19	0.21	0.59	0.66	0.19	0.23	0.63	0.77
MATH-Vision	0.22	0.43	0.41	0.79	0.32	0.30	0.76	0.71	0.27	0.39	0.48	0.71	0.41	0.34	0.75	0.61
Mean	0.46	0.43	0.76	0.68	0.43	0.36	0.77	0.62	0.45	0.33	0.81	0.59	0.48	0.36	0.81	0.61

A.2 PROMPT TEMPLATES AND EVALUATION PIPELINE

We acknowledge that evaluating every example in each benchmark would yield more accurate measurements. Due to computational resource constraints, in this work, we use 200 examples per benchmark as an estimate and apply exact-match evaluation to the boxed final answer. The nine natively multiple-choice benchmarks use the multiple-choice answer format. We apply the same answer instruction to the base condition and all corresponding intervention conditions. The self-description validation uses the same instruction for both inputs so that the image and description conditions differ only in their visual input modality.

Tab. 5 summarizes the templates used in our evaluation. The oracle and target model descriptions use the same question-blind instruction, and no condition uses a separate system prompt.

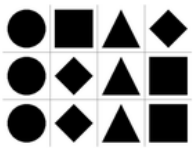
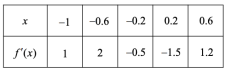
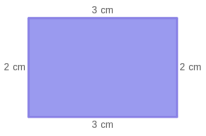
Table 5: **Prompt templates used for PARA.**

Template	Prompt
Question answering	You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think> </think>` tags. The final answer MUST BE put in `\boxed{}`. For multiple-choice questions, put the letter of the correct option in `\boxed{}`, e.g., `\boxed{A}`.
Question-blind description	Describe this image in exhaustive detail for someone who cannot see it. Report every object, text, number, axis label, chart value, spatial relationship, color, and fine visual detail you can discern. Do NOT answer any question or guess what is being asked—only describe what is literally visible.
Description-conditioned answer	`Visual description:` [visual description] `Question:` [question] [Question answering instruction]

Tab. 6 presents representative question-blind self-descriptions from the four benchmarks highlighted in the main text. These examples show that the descriptions capture visible content in detail without

introducing question-conditioned reasoning or answer information. Tabs. 7 and 8 additionally show descriptions from Claude Opus 4.8 and GPT-5.6-Luna on the same benchmarks using distinct examples. Tab. 9 presents a perception-recovery case in which the GPT-5.6-Luna description changes Qwen3-VL-4B's answer from A to the correct answer C.

Table 6: **Representative examples of question-blind descriptions.** Each description is generated by Qwen3-VL-4B without seeing the question.

Visual Input	Question-blind Description
PerceptionBench 	The image is a 3×4 grid composed of 12 distinct black shapes on a plain white background. Each cell contains a single solid black shape. The top row contains a circle, square, upward-pointing triangle, and diamond. The middle row contains a circle, diamond, upward-pointing triangle, and square. The bottom row repeats the middle row: a circle, diamond, upward-pointing triangle, and square. Each shape is centered within its cell, and the shapes do not overlap.
MMMU 	The image displays a table with two rows and six columns. The top row is labeled x, followed by the values -1, -0.6, -0.2, 0.2, and 0.6. The bottom row is labeled $f'(x)$, followed by the corresponding values 1, 2, -0.5, -1.5, and 1.2. The values are centered in their cells, with black text and lines on a white background.
MathVista 	The image displays a single solid-colored rectangle centered on a plain white background. The rectangle is filled with a uniform muted purple color and is wider than it is tall. The top and bottom sides are each labeled "3 cm," while the left and right sides are each labeled "2 cm." The labels are positioned outside the corresponding sides, and the rectangle has sharp right-angled corners.
MME 	The image displays a broom resting on a green grassy surface. It has a long, slender, light-colored wooden handle with a bright yellow rounded cap. The handle connects to a wide, dark-blue rectangular head with a raised central joint. Densely packed, light-colored straw bristles fan outward below the head and appear slightly worn at their tips. The grass contains scattered brown patches, and a partially visible dark object with a small red or orange area appears near the upper-right edge.

For the perception intervention, each oracle first generates a question-blind description. The target model then answers from the image augmented with the same description. A previous failure is included in $P_i^{M,B}$ if the base model succeeds with either oracle model description. For the reasoning intervention, the target model first generates its own question-blind description, with which the oracle answers the original question. A previous failure is included in $R_i^{M,B}$ only if the oracle also answers correctly from the original image. We take the item-level union over Claude Opus 4.8 and GPT-5.6-Luna as defined in Eq. 3.

Descriptions are generated with a 1024 token budget. Target model answers use up to 4096 tokens. Decoding is greedy for the locally served models. Claude Opus 4.8 and GPT-5.6-Luna use the default configurations, with no explicit stop strings used.

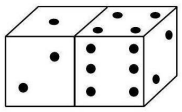
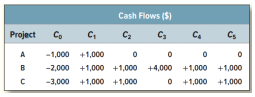
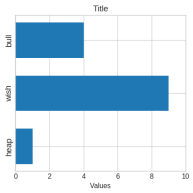
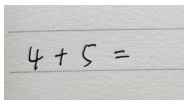
A.3 Reasoning Share

Beyond the directional BI_P and BI_R reported in the main paper, we further report Reasoning Share (RS) in Tab. 10 as an alternative normalized comparison of the two measured headrooms.

$$\mathrm{RS}^{M,B} = \frac{|R^{M,B}|}{|P^{M,B}| + |R^{M,B}|}. \tag{15}$$

RS directly measures their relative balance. $\mathrm{RS} = 0.5$ when $P = R$, while values above 0.5 indicate greater reasoning than perceptual headroom. This symmetric normalization is easy to interpret,

Table 7: **Representative question-blind descriptions from Claude Opus 4.8.**

Visual Input	Claude Opus 4.8 Description
PerceptionBench 	The image is a black-and-white line drawing of two dice positioned side by side. The left die shows a front face with two diagonal dots and a top face with one dot. The right die shows a front face with six dots, a top face with three dots, and a right face with two dots. The dice touch along their adjacent edges on a plain white background.
MMMU 	The image shows a table titled "Cash Flows ($)" in white text on a blue banner. Its columns are Project and C_0 through C_5. Project A contains $-1{,}000$, $+1{,}000$, and four zeros. Project B contains $-2{,}000$, $+1{,}000$, $+1{,}000$, $+4{,}000$, $+1{,}000$, and $+1{,}000$. Project C contains $-3{,}000$, $+1{,}000$, $+1{,}000$, 0, $+1{,}000$, and $+1{,}000$.
MathVista 	The image is a horizontal bar chart titled "Title." Three medium-blue bars correspond to "bull," "wish," and "heap." The x-axis is labeled "Values" and runs from 0 to 10 in increments of 2. The bars reach approximately 4, 9, and 1, respectively. The white plotting area contains faint grid lines and a rectangular border.
MME 	The image shows off-white ruled paper with two gray dashed horizontal guide lines. Centered between them is the handwritten expression "4 + 5 =" in dark ink. The handwriting is informal, and the space to the right of the equals sign is blank. No other text, numbers, or objects are visible.

but represents the two channels as complementary shares and therefore obscures their overlap. In contrast, the bottleneck indices BI_P and BI_R normalize each channel by $|P \cup R|$, measuring how broadly each intervention covers the distinct recoverable error set. Both indices can be high when the two interventions recover many of the same failures, thereby preserving information about their overlap. We therefore use BI in the main paper, while RS provides an alternative normalization that preserves the ordering and exhibits the same broad trends. The values below use the same results over the two oracles as Tab. 4.

B Experimental Details

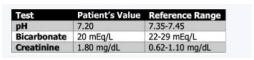
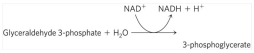
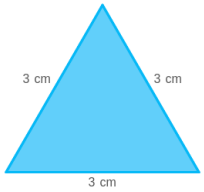
B.1 Datasets and Implementation Details

Training dataset. Following Vision-SR1 (Li et al., 2026d), we train all methods on Vision-SR1-47K, which contains approximately 47K image question–answer pairs collected from 24 open-source datasets. The data span multimodal mathematical reasoning, commonsense knowledge, and general visual understanding. We use the same training split across all methods to isolate the effect of their learning objectives.

Benchmarks. We list all benchmarks evaluated in this work. The first seven benchmarks below are used in our post-training experiments, and we extend extra ten benchmarks for evaluation of PARA.

- **MME** (Fu et al., 2023) evaluates both perceptual and cognitive abilities through a broad collection of visual tasks.
- **MMMU** (Yue et al., 2024) assesses expert-level multimodal understanding and reasoning across academic disciplines.
- **MMMU-Pro** (Yue et al., 2025) strengthens MMMU with more challenging questions and redesigned multiple-choice settings.

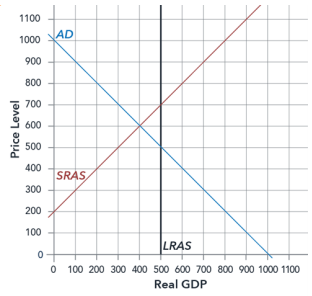
Table 8: **Representative question-blind descriptions from GPT-5.6-Luna.**

Visual Input	GPT-5.6-Luna Description
PerceptionBench 	The image is a rectangular table with a dark-gray header and three data rows. The first and third data rows have light-gray backgrounds, while the middle row is white. Its columns are "Test," "Patient's Value," and "Reference Range." The rows report pH as 7.20 and 7.35–7.45; Bicarbonate as 20 mEq/L and 22–29 mEq/L; and Creatinine as 1.80 mg/dL and 0.62–1.10 mg/dL.
MMMU 	The image is a black-and-white chemical reaction diagram. "Glyceraldehyde 3-phosphate + H_2O" appears to the left of a horizontal reaction arrow, while "3-phosphoglycerate" appears below the product side. Above the arrow, NAD^+ and NADH + H^+ are connected by a curved arrow that bends downward and rises toward NADH + H^+. All text and lines are dark against a plain white background.
MathVista 	The image shows a large upright triangle centered on a white background. Its interior is uniformly light blue, with a brighter blue outline. The triangle is approximately equilateral, with a sharp top vertex and a horizontal base. Each of its three sides is labeled "3 cm" in dark-gray text, and no other objects or chart elements are visible.
MME 	The image is a wide photograph of a lake under an overcast sky. Layered gray and lavender clouds fill the upper half. A tree-covered shoreline and low rounded ridge extend across the center, with autumn foliage in green, brown, rust, orange, and muted red. Blue-gray water with fine horizontal ripples fills the foreground; no people, buildings, or boats are visible.

- **MATH-Vision** (Wang et al., 2024) evaluates mathematical problem solving grounded in diagrams and other visual contexts.
- **PerceptionBench** (Lin et al., 2026) targets fine-grained atomic visual perception across diverse visual attributes.
- **ScienceQA** (Lu et al., 2022) evaluates multimodal science question answering using image and textual context.
- **VisNumBench** (Weng et al., 2025) tests numerical understanding and reasoning grounded in visual content.
- **BLINK** (Fu et al., 2024) evaluates core visual perception skills, including depth, correspondence, and multi-view reasoning.
- **ChartQA** (Masry et al., 2022) measures visual and logical reasoning over real-world charts.
- **HallusionBench** (Guan et al., 2024) diagnoses visual illusion, language hallucination, and image-context reasoning in VLMs.
- **LogicVista** (Xiao et al., 2024) evaluates visually grounded logical reasoning across multiple reasoning types.
- **MathVerse** (Zhang et al., 2024) tests whether multimodal models understand and use diagrams when solving visual mathematics problems.
- **MathVista** (Lu et al., 2024) evaluates mathematical reasoning across charts, diagrams, figures, and other visual contexts.
- **MMBench** (Liu et al., 2024) provides a systematic multiple-choice evaluation of broad multimodal capabilities.
- **MMStar** (Chen et al., 2024) evaluates six core multimodal capabilities using vision-dependent samples with reduced data leakage.
- **MM-Vet** (Yu et al., 2023) evaluates integrated capabilities that require combinations of recognition, OCR, knowledge, generation, mathematics, and spatial understanding.

Table 9: **Example of perception recovery.** Qwen3-VL-4B answers the same MMMU question before and after receiving the question-blind description generated by GPT-5.6-Luna, recovering the perception failure.

Visual Input



What is the size of real GDP in the short-run equilibrium?

(A) 500 (B) 450 (C) 400 (D) 600

Base Model Generation

The short-run equilibrium occurs where the aggregate demand (AD) curve intersects the short-run aggregate supply (SRAS) curve. The intersection appears at $(500, 600)$, so the size of real GDP is 500 billion. Therefore, the answer is $\boxed{\text{A}}$.

GPT-5.6-Luna Description

The image is a square economic graph on a white background. The vertical axis is labeled "Price Level," and the horizontal axis is labeled "Real GDP." Both axes have tick marks and numeric labels at intervals of 100. A blue downward-sloping line labeled "AD" begins at approximately $(0, 1000)$, passes through approximately $(400, 600)$, and crosses the horizontal axis at approximately $(1000, 0)$. A red upward-sloping line labeled "SRAS" begins at approximately $(0, 200)$ and passes through approximately $(400, 600)$. A vertical black line labeled "LRAS" is located at real GDP 500.

Recovered Answer

The short-run equilibrium occurs where the AD and SRAS curves intersect. The description places their intersection at $(400, 600)$, so the size of real GDP is 400 billion. Therefore, the answer is $\boxed{\text{C}}$.

Table 10: **Reasoning Share across models.** RS reports $R/(P + R)$ for the four models and 17 benchmarks in the main PARA evaluation. †Natively multiple-choice.

Benchmark	Qwen3-VL-2B	Qwen3-VL-4B	Llama-4-Scout-17B	Gemma-3-27B
BLINK†	0.39	0.30	0.30	0.39
VisNumBench†	0.43	0.44	0.46	0.42
RealWorldQA	0.37	0.50	0.40	0.38
PerceptionBench	0.31	0.24	0.33	0.32
MMStar†	0.44	0.39	0.40	0.39
ChartQA	0.52	0.50	0.36	0.43
HallusionBench	0.43	0.40	0.34	0.36
MME	0.56	0.55	0.49	0.39
MMBench†	0.50	0.49	0.43	0.45
ScienceQA†	0.50	0.50	0.40	0.54
MM-Vet	0.41	0.37	0.41	0.46
MMMU†	0.61	0.55	0.51	0.53
MMMU-Pro†	0.49	0.48	0.56	0.60
LogicVista†	0.50	0.44	0.42	0.28
MathVista†	0.28	0.29	0.26	0.33
MathVerse	0.61	0.59	0.53	0.55
MATH-Vision	0.66	0.48	0.60	0.45
Mean	0.47	0.44	0.42	0.43

- **RealWorldQA** (xAI, 2024) measures spatial understanding of real world scenes.

We use each benchmark's official metric and evaluation protocol. MME is reported on its original scale in the result tables and normalized to a 100-point scale only when computing the cross-benchmark average.

Training setup. We perform full-parameter fine-tuning of Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct on four NVIDIA H200 GPUs using fully sharded data parallelism and gradient checkpointing. We train with one PPO epoch per batch, a global batch size of 128, and eight rollouts per prompt. We report the results of the 270 step checkpoints. We use AdamW with a learning rate of 1×10^{-6}, weight decay 0.01, and maximum gradient norm 1.0. Following GRPO, we use a standard binary accuracy reward from a rule-based verifier, rather than self-rewarding signals based on model judgments or reasoning consistency (Yuan et al., 2024; Zhang et al., 2025). The final answer is extracted from the generated `\boxed{}` and compared with the ground-truth answer. Following the official Vision-SR1 implementation, we use a low-variance KL penalty with coefficient $\beta = 0.01$, asymmetric clipping ratios $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.3$. During training, rollouts are sampled with temperature 1.0 and top-p $= 0.99$, with a maximum response length of 4096 tokens. We use a constant learning-rate schedule without warmup. The attention loss coefficient λ is 1.0, and supervision is applied to the last decoder layer. All experiments use random seed 2026, and input images are capped at 1,003,520 pixels ($1280 \times 28 \times 28$).

Implementation details and efficiency. The student attention is recomputed with gradients during the current policy forward pass, requiring no additional forward pass. We use generated response tokens as queries and visual tokens as keys. Attention is first normalized over visual keys independently for each head and averaged uniformly across heads. Padding tokens are excluded and different response lengths are handled by response-wise pooling. Siblings share the same image and therefore the same visual token layout.

B.2 Additional Ablation Results

In Sec. 6.3, we discuss the effectiveness of our method components and the robustness of our default coefficient and supervision scope. In this section, we further provide more details.

Grounding penalty. The grounding variant in Tab. 3 derives a visual support set G_g from the best sibling's pooled attention distribution $\widehat{p}_g$, instead of the full attention map. Specifically, we sort the visual tokens by $\widehat{p}_g$ and select the smallest prefix whose cumulative probability above a threshold γ. We then penalize the current rollout attention mass outside this support:

$$\mathcal{L}_{\text{ground}} = \frac{1}{|\mathcal{I}|} \sum_{(g,i)\in\mathcal{I}} w_i \sum_{j \notin G_g} \bar{p}_i^{\theta}(j). \tag{16}$$

The variant adds $\beta\mathcal{L}_{\text{ground}}$ to the sibling alignment objective, using the same configurations as GRAD. G_g is a sample-specific subset of the existing visual tokens. We report the coefficient comparison results in Tab. 11, and layer supervision comparison in Tab. 12.

Table 11: **Sensitivity to the attention-loss coefficient.** We vary λ on Qwen3-VL-2B.

λ	MME	MMMU	MMMU-Pro	MathVis	PercepBench	ScienceQA	VisNum.	Avg.
0.2	<u>2248.4</u>	**52.7**	34.2	19.1	20.6	<u>90.1</u>	<u>38.2</u>	<u>47.9</u>
1.0	**2290.4**	<u>52.1</u>	**35.1**	**21.1**	**24.4**	**90.3**	**40.8**	**49.4**
5.0	2228.8	49.4	<u>34.5</u>	<u>19.7</u>	<u>20.9</u>	88.8	36.1	47.0

Table 12: **Effect of supervised attention layers.** We extend the final-layer objective to the top 4, 8, 16, or all layers of Qwen3-VL-2B.

	MME	MMMU	MMMU-Pro	MathVis	PercepBench	ScienceQA	VisNum.	Avg.
Final layer	**2290.4**	**52.1**	**35.1**	**21.1**	**24.4**	**90.3**	**40.8**	**49.4**
Top 4	2259.6	48.0	34.3	18.4	19.5	88.5	36.3	46.5
Top 8	2242.8	50.9	33.9	20.4	20.1	89.7	36.1	47.3
Top 16	2245.6	47.1	29.9	14.8	18.8	85.9	33.2	44.3