On Asymmetric Optimization of Reasoning and Perception in Vision-Language Model Post-Training

Anonymous ACL submission

Abstract

Post-training has greatly improved reasoning in frontier vision-language models, yet its gains for perception remain comparatively limited, creating a bottleneck for end-to-end visual reasoning. To investigate this gap, we introduce a *controlled diagnostic framework* with two synthetic tasks that disentangle perception from reasoning. Our analysis reveals a consistent **perception-reasoning asymmetry**: post-training improves reasoning more substantially than perception, though the underlying mechanism differs by training paradigm. For supervised fine-tuning (SFT), this asymmetry stems from **token imbalance** in chain-of-thought supervision, where perception occupies fewer tokens and thus receives a weaker training signal. Dynamically reweighting the loss mitigates this imbalance and boosts end-to-end performance by up to 18.2. For reinforcement learning (RL), the asymmetry instead arises from **reward coupling**: outcome rewards correlate more strongly with reasoning than with perception, weakening the signal for perception learning. Adding a perception-aware reward alleviates the imbalance and improves end-to-end accuracy by up to 6.0; even without ground-truth perception rewards, a reliable surrogate reward provide useful signal, yielding gains of 3.2 points. Together, our results comprehensively diagnose asymmetric optimization and suggest concrete interventions to balance perception and reasoning.

1 Introduction

Post-training methods, especially supervised fine-tuning (SFT) and reinforcement learning (RL), have become central to modern vision-language models (VLMs). Frontier VLMs trained with these techniques exhibit increasingly strong reasoning capabilities (Deng et al., 2025; Wang et al., 2025a; Bai et al., 2025a), achieving state-of-the-art performance across a wide range of benchmarks (Lu et al.; Yue et al., 2024, 2025). However, stronger reasoning does not necessarily imply more reliable perception: recent studies suggest that post-training may yield limited gains, or even degradations, in perception (Liu et al., 2025; Tian et al., 2025; Yao et al., 2025). These observations raise a central question: *why does post-training disproportionately favor reasoning over perception?* Understanding the cause of this imbalance is critical, as perception errors can severely bottleneck end-to-end visual reasoning performance.

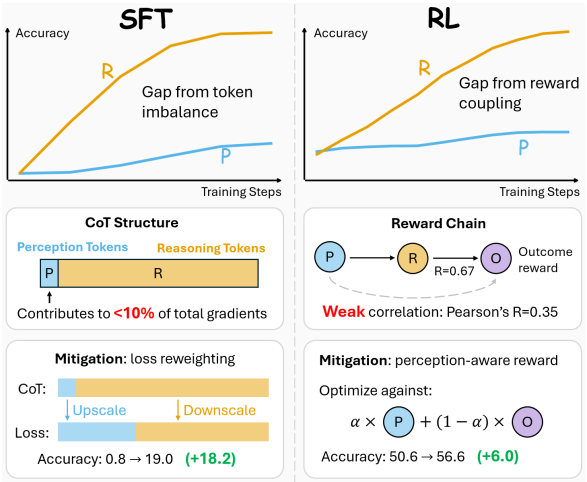

Figure 1: **Asymmetric optimization of perception and reasoning in VLM post-training.** Both SFT (left) and RL (right) improve reasoning (R) substantially more than perception (P), but with different causes and mitigation mechanism.

To systematically study this question, we experiment in a set of **controlled settings** for visual reasoning. Our design is guided by two desiderata: (i) fully controllable visual inputs with ground-truth perception annotations; and (ii) a structured output format that cleanly separates perception and reasoning tokens. This design enables two complementary analyses: 1) direct evaluation of model-predicted perception, and 2) counterfactual evaluation of reasoning, where the model performs reasoning conditioned on ground-truth perception. We instantiate this framework with two synthetic

tasks, graph coloring and Sudoku, using programmatically rendered images.

Within this framework, we conduct controlled experiments with two strong open-source VLM families, Qwen3-VL (Bai et al., 2025a) and InternVL3.5 (Wang et al., 2025c), on graph coloring and Sudoku. We identify **a consistent asymmetric learning behavior**, where both SFT and RL improve reasoning substantially more than perception. As a result, perception remains the dominant bottleneck for end-to-end visual reasoning, even after post-training improves the model's ability to reason over correctly perceived inputs.

To explain this asymmetry, we investigate two mechanisms: *token imbalance*, where perception receives little optimization signal because it occupies a smaller fraction of CoT tokens; and *reward coupling*, where RL outcome rewards align more strongly with reasoning than with perception.

Empirical results in SFT identify **token imbalance** as the primary driver: perception accounts for only 2.5% of tokens in CoT supervision, contributing just 1.3% of the loss and 8.5% of the gradient norm when training Qwen3-VL on graph coloring. A targeted loss-reweighting intervention confirms this mechanism. Increasing the perception weight improves perception accuracy but exposes a smooth *perception-reasoning trade-off*: under a fixed training budget, improving perception can weaken reasoning. Crucially, end-to-end accuracy peaks when perception is *modestly upweighted*, improving over standard SFT by up to 13.8 points. To avoid manual tuning, we further apply dynamic multi-task balancing methods and improve over standard SFT by up to 18.2 points.

RL results, instead, suggest **reward coupling** as the primary driver: outcome rewards correlate strongly with reasoning correctness ($r = 0.65$–1.00), but much more weakly with perception correctness ($r = 0.34$–0.43). Augmenting outcome reward with perception reward confirms this mechanism and again exposes a *perception–reasoning trade-off*: increasing the perception-reward weight improves perception but weakens reasoning. In most settings, end-to-end performance peaks at a *non-zero perception-reward weight*, improving over standard RL by up to 6.0 points. Since ground-truth perception rewards are often unavailable, we further evaluate surrogate reward designs from prior work (Chen et al., 2025; Li et al., 2025c; Xiao et al., 2025). We find that reward-perception correlation serves as an effective *diagnostic* for predicting perception gains, with the strongest surrogate yielding a 3.2-point improvement.

Overall, our contributions are threefold: (1) We introduce a controlled framework for disentangling perception and reasoning in VLM post-training; (2) We identify a consistent perception-reasoning asymmetry and trace it to distinct mechanisms: token imbalance in SFT and reward coupling in RL; (3) We show that targeted mitigations—loss reweighting for SFT and perception rewards for RL—reduce this asymmetry and improve end-to-end performance by up to 18.2 points.

2 Analysis Framework

2.1 Problem Setup

Let π_θ denote a vision-language model parameterized by θ. Given an input sequence $\mathbf{x}$ consisting of both visual and textual tokens, the model generates the output sequence $\mathbf{y} = [y_1, \cdots, y_T]$ in an auto-regressive manner, $y_t \sim \pi_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{<t})$.

Disentangled outputs. Standard VLM CoT entangles perception with reasoning, making it difficult to supervise them separately or attribute errors. In practice, however, CoT traces often exhibit a natural order: models first verbalize the visual information and then reason over it. Recent work starts to explicitly enforce perception-before-reasoning format; this by itself improves final performance, and further enables targeted supervision or rewards on the perception components (Li et al., 2025c; Xia et al., 2025; Xiao et al., 2025; Chen et al., 2025; Wu et al., 2026). Following this line of work, we structure each output $\mathbf{y}$ as two textual sequences, perception $\mathbf{p}$ followed by reasoning $\mathbf{r}$, separated by special tokens. Slightly abusing notation, we write the generation process as $\mathbf{p} \sim \pi_\theta(\cdot \mid \mathbf{x}),\ \mathbf{r} \sim \pi_\theta(\cdot \mid \mathbf{x}, \mathbf{p})$.

Disentangled evaluation. Our evaluation measures three metrics: end-to-end accuracy a, perception accuracy a_p, and reasoning accuracy a_r.

To support this evaluation, we assume the existence of a *canonical textual representation* $\mathbf{p}^*$ that preserves the complete visual information needed for the task. This allows us to directly evaluate **perception accuracy** by comparing the model-generated perception $\mathbf{p}$ against the oracle perception: $a_p = \mathbf{1}(\mathbf{p} = \mathbf{p}^*)$. We further define a task-specific evaluator $Acc(\mathbf{r} \mid \cdot)$ that checks a reasoning output $\mathbf{r}$ over a textual perception representation rather than the raw image $\mathbf{x}$. The **end-to-end**

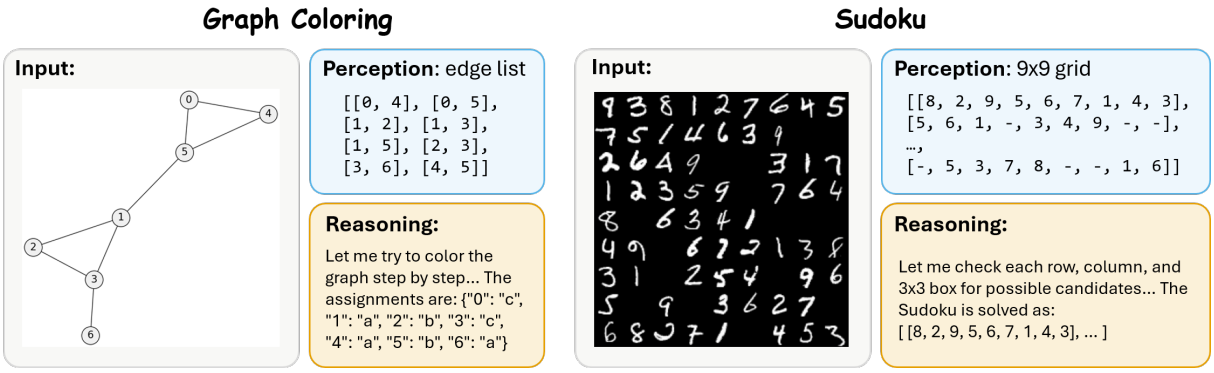

Figure 2: **Task illustration.** We show the graph coloring (GC) and Sudoku tasks with their input images, ground truth perception, and expected reasoning respectively.

accuracy is thus given by $a = Acc(\mathbf{r} \mid \mathbf{p}^*)$.

Evaluating reasoning accuracy is nontrivial because reasoning is conditioned on perception, so reasoning errors are naturally entangled with perception errors. An intuitive metric is **conditional reasoning accuracy**, $\tilde{a}_r = Acc(\mathbf{r} \mid \mathbf{p})$, which evaluates whether the reasoning is correct under the model's own predicted perception. However, since $\mathbf{p}$ is model-generated, this metric does not isolate intrinsic reasoning ability: models may score higher by producing perceptions that are easier to reason from, rather than by reasoning more accurately. To decouple reasoning from perception, we instead use **counterfactual reasoning accuracy**. Specifically, we perform a *counterfactual inference* by conditioning the model on the oracle perception and sample $\mathbf{r}' \sim \pi_\theta(\cdot \mid \mathbf{x}, \mathbf{p}^*)$, and then measure $a_r = \mathrm{Acc}(\mathbf{r}' \mid \mathbf{p}^*)$. Unless otherwise stated, all reported reasoning accuracy refers to this counterfactual metric.

2.2 Tasks Setup

Our framework requires tasks with (i) a canonical perception representation p^* for measuring perception accuracy, and (ii) a programmatic evaluator $Acc(\cdot \mid \cdot)$ defined solely over that representation. Real-world visual reasoning benchmarks, however, rarely meet these requirements: they often lack perception annotations, and canonical perception is ill-defined for realistic images due to ambiguity in faithful scene descriptions. Additionally, many benchmarks lack suitable train/test splits for controlled SFT and RL experiments.

We therefore use **synthetic tasks as a controlled testbed**. Starting from text-only reasoning problems, we programmatically render visual inputs while retaining the original text input as the canonical perception representation p^*. As shown in Figure 2, we instantiate two tasks: (1) **Graph Coloring** (GC), where p^* is the graph edge list and correctness requires coloring all nodes with the graph's chromatic number while assigning different colors to adjacent nodes; and (2) **Sudoku**, where p^* is the partially filled 9×9 grid and correctness requires completing the grid such that every row, column, and 3×3 subgrid contains each digit exactly once. As both tasks are programmatically constructed, we can synthesize large-scale training sets to perform controlled analysis.

2.3 Training Paradigm

Supervised fine-tuning. Given a dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$, where each input $\mathbf{x}$ is paired with a target output $\mathbf{y}$, SFT optimizes the token-averaged cross-entropy objective:

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}\left[\frac{1}{|\mathbf{y}|}\sum_{t=1}^{|\mathbf{y}|} \log \pi_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right].$$

Reinforcement learning. For RL, each input $\mathbf{x}$ is paired with a task reward $R(\mathbf{y} \mid \mathbf{x})$, instantiated as the end-to-end accuracy in our setting. Following GRPO, for each input $\mathbf{x}$, we sample a group of G responses $\{\mathbf{y}^{(i)}\}_{i=1}^G$ from the reference policy $\pi_{\theta_{\mathrm{old}}}$ and compute rewards $r_i = R(\mathbf{y}^{(i)} \mid \mathbf{x})$. Rewards are normalized within the group to obtain a sequence-level advantage:

$$\hat{A}_i = \frac{r_i - \mathrm{mean}\left(\{r_j\}_{j=1}^G\right)}{\mathrm{std}\left(\{r_j\}_{j=1}^G\right)},$$

which is shared across all tokens in the response, i.e., $\hat{A}_{i,t} = \hat{A}_i$. Omitting clipping and KL regularization for clarity, the GRPO objective is:

$$\mathcal{J}_{\mathrm{GRPO}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{y}^{(i)}|}\sum_{t=1}^{|\mathbf{y}^{(i)}|}\rho_{i,t}(\theta)\hat{A}_{i,t}\right],$$

where $\rho_{i,t}(\theta)$ is the token-level importance ratio.

2.4 Candidate Mechanisms and Interventions

The disentangled framework allows us to diagnose not only the presence of perception-reasoning asymmetry, but also its potential causes. We consider two candidate mechanisms: token imbalance in token-level objectives, and reward coupling when end-to-end accuracy is used as the RL reward. For each candidate mechanism, we formulate a testable hypothesis and introduce a targeted intervention, deferring empirical validation to §3.

H1: Token imbalance. A first candidate mechanism is that perception receives a weaker optimization signal because it occupies a much smaller fraction of CoT tokens. In VLM CoT, perception is often a compact transcription of the visual input, while reasoning may span many tokens of verification, reflection, and self-correction. Prior work reports that perception accounts for only 17.6% of tokens in base VLM outputs and drops to 12–14% after RL training (Li et al., 2025b). As in §3.2 and §3.3, a similar imbalance is present in our setting.

Since both SFT and RL use token-averaged objectives, perception and reasoning are implicitly weighted by their token counts. For SFT, for example, writing the output as $\mathbf{y} = (\mathbf{p}, \mathbf{r})$, the standard token-averaged loss can be decomposed as

$$\mathcal{L}_{\text{SFT}} = \frac{1}{|\mathbf{p}| + |\mathbf{r}|} \left(\mathcal{L}_p + \mathcal{L}_r\right),$$
$$\mathcal{L}_p = -\mathbb{E}\left[\sum_{y_t \in \mathbf{p}} \log \pi_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right],$$
$$\mathcal{L}_r = -\mathbb{E}\left[\sum_{y_t \in \mathbf{r}} \log \pi_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right],$$

Thus, even when perception is necessary for solving the task, its shorter span gives it less aggregate influence on the update.

We test this hypothesis with **loss reweighting**. For SFT, we optimize against: $\mathcal{L}_{\text{SFT},\lambda} = \lambda \frac{\mathcal{L}_p}{|\mathbf{p}|} + (1-\lambda)\frac{\mathcal{L}_r}{|\mathbf{r}|}$. Here, λ directly controls the relative emphasis on perception. When λ exceeds the perception token fraction in the output CoT, perception is upweighted relative to standard training. The same intervention can be applied to GRPO analogously, with implementation details in Appendix B.3. If token imbalance is a main driver of the asymmetry, increasing the perception weight should improve perception accuracy and reduce the end-to-end bottleneck.

H2: Reward coupling. For RL, a second candidate mechanism is that the outcome reward, i.e., end-to-end accuracy a, is more weakly coupled to perception accuracy than to reasoning accuracy. Perception influences a only when the downstream reasoning is also correct, whereas reasoning more directly determines the final answer and thus the reward. Consequently, using a as the reward signal can induce noisy or misaligned credit assignment for perception, analogous to credit-assignment challenges for intermediate reasoning steps in text-only LLM RL (Lightman et al., 2024).

We test this hypothesis with **reward augmentation**. Instead of the standard outcome reward $\mathcal{R}(y \mid x) = a$, we train with:

$$\mathcal{R}_\alpha(y \mid x) = \alpha \cdot a_p + (1-\alpha) \cdot a,$$

where $\alpha \in [0, 1]$ controls the strength of the perception signal. All other RL components follow §2.3. If reward coupling drives the RL asymmetry, increasing α should strengthen perception optimization and mitigate a perception-reasoning imbalance.

3 Experimental Results

We empirically investigate three questions: (1) Does post-training optimize perception and reasoning asymmetrically? (2) What mechanism drives the asymmetry? (3) Which interventions best rebalance perception and reasoning, and how do they affect end-to-end accuracy? We first describe the experimental setup in §3.1, and then present results for answering these three questions for SFT and RL in §3.2 and §3.3 respectively.

3.1 Experimental Settings

We experiment with `Qwen3-VL-2B-Instruct` (Bai et al., 2025a) and `InternVL3.5-2B` (Wang et al., 2025c) on the `GC` and `Sudoku` tasks. For each task, we generate $500/2k/16k$ data instances for testing, SFT training, and RL training respectively. To curate output supervision for SFT, we concatenate ground-truth perception with reasoning distilled from `Qwen3-32B`. For loss reweighting, we experiment with $\lambda \in \{0.1, 0.2, 0.5, 0.8, 0.9\}$; for reward coupling, we experiment with $\alpha \in \{0.0, 0.2, 0.5, 0.8, 1.0\}$, where $\alpha = 0$ corresponds to the standard RL. Full details are in Appendix B.1.

3.2 Supervised Fine-Tuning Phase

Asymmetric optimization in SFT. We first investigate how standard SFT optimizes perception and

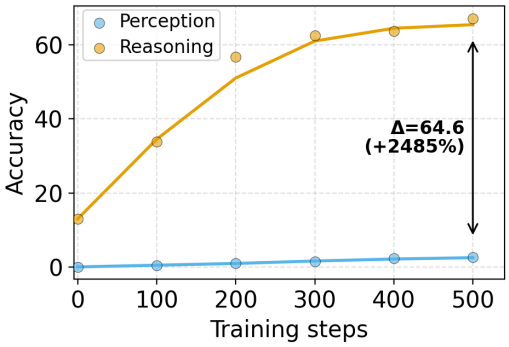

Figure 3: Perception and reasoning accuracy during SFT optimization of Qwen model on GC task, highlighting the asymmetric optimization.

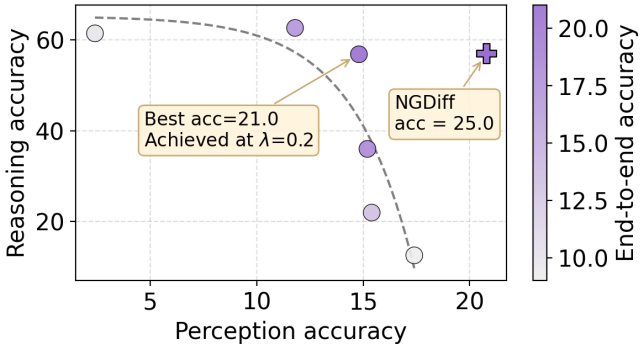

Figure 4: Perception-reasoning trade-off when training Qwen on the GC task with SFT by varying λ. We also show NGDiff achieves a better trade-off and higher end-to-end accuracy.

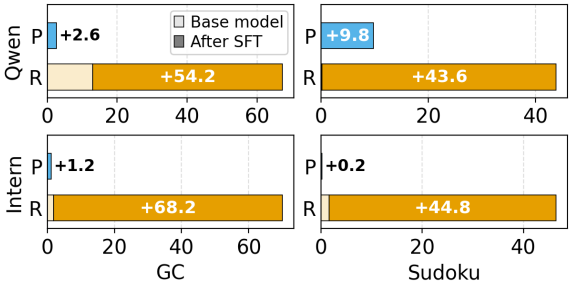

Figure 5: **Perception and reasoning accuracy before and after SFT.** Both start near zero, but SFT substantially improves reasoning while leaving perception largely flat.

	GC		Sudoku	
	Qwen	InternVL	Qwen	InternVL
Standard SFT	9.8	10.6	6.4	0.8
Loss Reweighting	21.0	**20.6**	18.4	14.6
	↑11.2	↑10.0	↑12.0	↑13.8
NGDiff	**25.0**	**20.6**	**22.0**	**19.0**
	↑15.2	↑10.0	↑15.6	↑18.2

Table 1: **End-to-end accuracy after SFT.** We compare standard SFT and two SFT asymmetry mitigation methods: loss reweighting with the best tuned λ and NGDiff.

reasoning performance. Starting from base models that achieve near-zero accuracy in both perception and reasoning on our out-of-distribution tasks, SFT exhibits a pronounced asymmetry: it substantially improves reasoning while yielding only limited gains in perception. This pattern is consistent across all four model-task settings, as summarized in Figure 5; Figure 3 further illustrates the optimization dynamics of Qwen on the GC task as a representative setting.

Testing H1: Token imbalance drives SFT asymmetry. We first test whether SFT asymmetry can be explained by token imbalance in CoT supervision. Since the reasoning traces are distilled from a text-only RL model, following common practice (Deng et al., 2025), they occupy substantially more tokens than the perception component: perception accounts for only 2.2% of tokens in GC and 2.5% in Sudoku. This imbalance yields a much weaker optimization signal: in the Qwen-GC setting, for example, perception contributes only 1.3% of the loss and 8.5% of the gradient norm, supporting token imbalance as a concrete mechanism behind SFT asymmetry.

We then employ the loss reweighting intervention in §2.4 as a mechanism test. As shown in Figure 4, varying λ reveals a Pareto-like frontier: under a fixed training budget, increasing λ improves perception accuracy but weakens reasoning optimization. Importantly, **the best end-to-end visual reasoning accuracy is achieved by moderately upweighting perception**. As reported in Table 1, the best λ improves end-to-end performance over standard SFT by 10.0–13.8 points across all four settings, suggesting loss reweighting as a simple and effective mitigation for SFT asymmetry.

Dynamic weighting further improves the trade-off. While loss reweighting improves end-to-end performance, its optimal λ is model- and task-dependent, requiring an expensive hyperparameter tuning. We thus evaluate NGDiff that dynamically balances multiple objectives based on their gradient norms (Jin et al., 2025). For each mini-batch, let $\mathbf{g}_p$ and $\mathbf{g}_r$ denote the gradients from the perception and reasoning objectives, respectively. NGDiff then sets the perception weight as:

$$\lambda = \frac{1}{||\mathbf{g}_p||} \Big/ \left(\frac{1}{||\mathbf{g}_p||} + \frac{1}{||\mathbf{g}_r||} \right).$$

As shown in Table 1, NGDiff consistently matches or outperforms the manually tuned loss-reweighting in all settings, with gains of up to 4.4 points. This highlights NGDiff as a practical and effective mitigation for SFT asymmetry.

Robustness to stronger perception initialization. One concern is that the trade-off emerges only from

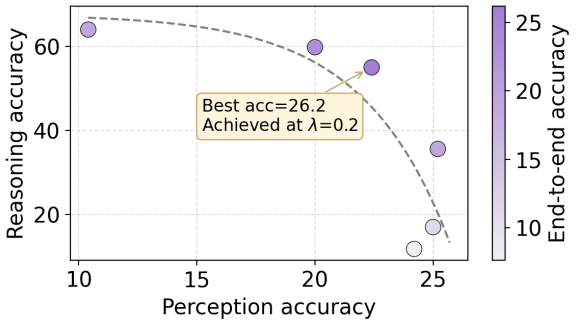

Figure 6: The perception-reasoning trade-off persists with stronger perception initialization.

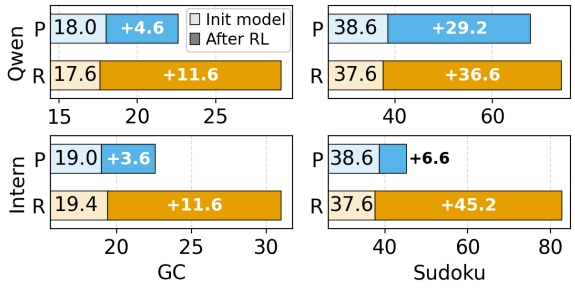

Figure 7: **Perception and reasoning accuracy before and after RL**, where RL improves reasoning more substantially than perception.

weak initial perception, limiting its relevance to practical post-training. We thus add a perception-only SFT stage before loss reweighting to provide a stronger perception initialization. As in Figure 6, a similar perception-reasoning trade-off persists with end-to-end performance peaking at an intermediate perception weight. This suggests that the trade-off stems from joint perception-reasoning optimization rather than an artifact of weak initialization.

Takeaway for SFT

Findings: SFT under-optimizes perception because reasoning tokens dominate the CoT trace, leaving perception weakly optimized and bottlenecking end-to-end performance.

Suggestions: Reweight the SFT loss toward perception. Moderate perception upweighting improves end-to-end performance, while NGDiff provides a tuning-free way to approach a better perception-reasoning balance.

3.3 Reinforcement Learning Phase

Asymmetric optimization in RL. We next investigate how RL optimizes perception and reasoning. To ensure a fair comparison, we initialize RL from SFT checkpoints with matched perception and reasoning accuracies. As shown in Figure 7, RL exhibits a similar asymmetry to SFT: it consistently improves reasoning more than perception across all four settings. Figure 9 further shows the optimization dynamics for the `Qwen-GC` setting as a representative case.

Testing H1: token imbalance does not explain RL asymmetry. Since RL is initialized from SFT checkpoints, the CoT token imbalance persists and even intensifies: in `Qwen-GC`, for example, the perception-token fraction drops from 2.2% to 1.5% after RL training. However, the reweighting intervention from §2.4 no longer mitigates the asymmetry. Standard RL achieves $a/a_p/a_r = 18.8/22.6/31.0$, while reweighting policy-gradient terms with $\lambda = 0.1$, 0.2, and 0.5 degrades performance to $17.8/15.4/29.8$, $17.4/14.2/26.6$, and $13.2/2.2/18.8$, respectively. Thus, although token imbalance remains, directly upweighting perception in the RL objective consistently disrupts training rather than improving perception, indicating that RL asymmetry is not primarily driven by token allocation.

Testing H2: Reward coupling drives RL asymmetry. To quantitatively test reward coupling as a potential cause, we measure how strongly the outcome reward aligns with perception $v.s.$ reasoning correctness. For each setting, we sample 2500 rollouts, evaluate the perception and reasoning correctness for each rollout, and compute Pearson's r between each correctness metric and the outcome reward. For perception, we use perception accuracy a_p. For reasoning, we use conditional reasoning accuracy $\tilde{a}_r = Acc(\mathbf{r} \mid \mathbf{p})$, since counterfactual reasoning accuracy requires resampling reasoning under oracle perception and is therefore not defined within a single rollout. Figure 8 shows that the outcome reward is consistently more correlated with reasoning than with perception, indicating that standard RL supplies a stronger learning signal for reasoning and supporting reward coupling as the primary driver of RL asymmetry.

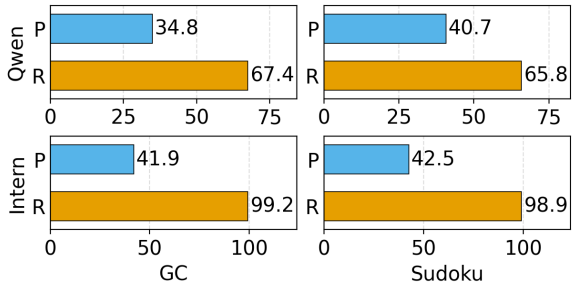

Figure 8: Correlation between outcome reward and perception (P) or reasoning (R) accuracy, measured by Pearson's r.

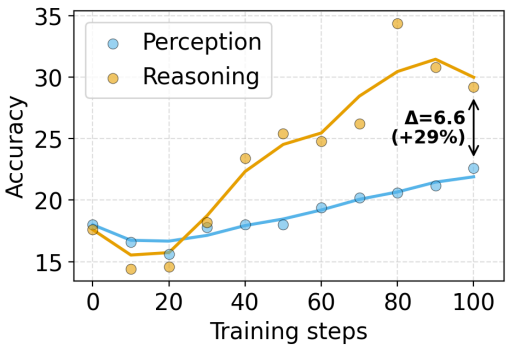

Figure 9: Perception and reasoning accuracy during RL optimization of Qwen model on GC task.

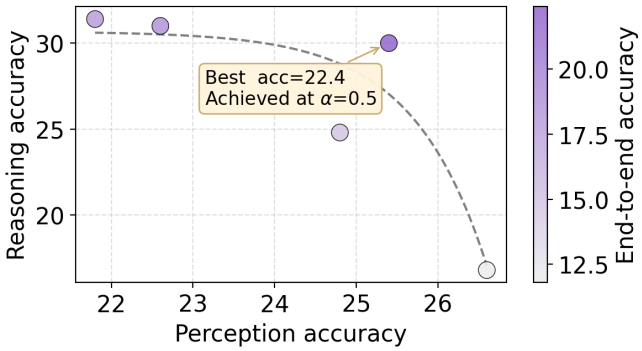

Figure 10: Perception-reasoning trade-off when training Qwen on the GC task with SFT by varying α.

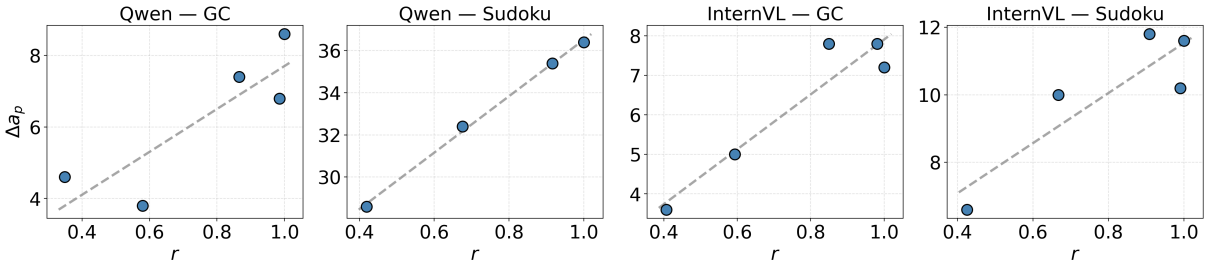

Figure 11: The correlation r between RL reward and perception accuracy reliably predicts the perception gains Δa_p after RL.

	GC		Sudoku	
	Qwen	InternVL	Qwen	InternVL
Init model	10.2	15.8	18.4	14.6
Standard RL ($\alpha = 0$)	18.8	20.6	50.6	**40.2**
Reward Augmentation	**22.4**	**22.6**	**56.6**	38.6
	↑3.6	↑2.0	↑6.0	↓1.6

Table 2: **End-to-end accuracy after RL**. We compare standard RL against the reward augmentation mitigation approach with best tuned non-zero α.

	GC		Sudoku	
	Qwen	InternVL	Qwen	InternVL
Outcome Reward	34.8	40.7	41.9	42.5
CLIP Similarity	-0.1	-0.1	5.8	8.0
Self Reward (N=1)	32.0	37.3	46.6	45.7
Self Reward (N=4)	65.8	67.1	71.3	69.8
Teacher Reward	**96.0**	**95.6**	**74.9**	**72.0**

Table 3: Pearson's correlation r between perception accuracy between and different reward designs.

	r	a	a_p	a_r
Outcome only	34.8	18.8	22.6	31.0
Reward Augmentation ($\alpha = 0.5$)				
Self Reward ($N = 1$)	32.0	18.4	21.8	27.6
Teacher reward	96.0	21.0	25.0	**30.6**
GT perception reward	**100.0**	**22.4**	**25.4**	30.0

Table 4: **RL performance under different reward settings** in the Qwen-GC setting. For each reward, we report its Pearson correlation r with perception accuracy. $a/a_p/a_r$ denote end-to-end, perception, and reasoning accuracies, respectively.

We then employ reward augmentation as a mechanism test. Varying α exposes a perception-reasoning trade-off as shown in Figure 10. Although the frontier is noisier than in SFT due to RL variance, the trend is consistent: larger α improves perception but weakens reasoning under the same training budget. Figure 11 shows thta across different α settings, reward-perception correlation reliably predicts reliably predicts the perception gains after RL, suggesting **reward-perception correlation as a useful diagnostic** for reward design. Finally, Table 2 further shows that end-to-end accuracy is maximized at an intermediate, non-zero α in most settings, indicating **perception-aware reward augmentation is an effective mitigation**.

Surrogate perception rewards offer a practical approximation. Since ground-truth perception rewards are rarely available in realistic settings, we evaluate three surrogate designs from prior work: CLIP-based image-text similarity (Chen et al., 2025), self-rewarding with rollout budgets $N \in \{1, 4\}$ (Li et al., 2025c), and teacher-based rewards using `gemini-3-flash-preview` (Xiao et al., 2025). Implementation details are in Appendix B.4. Using correlation with ground-truth perception accuracy as a diagnostic, Table 4 shows that CLIP provides little useful signal, whereas the teacher reward is consistently the most reliable. Self-rewarding with $N = 1$ only modestly improves over the outcome reward, while increasing the budget to $N = 4$ substantially strengthens the signal at additional computational cost.

We further validate this diagnostic with a proof-of-concept RL run on Qwen-GC. With $\alpha = 0.5$, we augment the outcome reward using self-rewarding

($N = 1$) or teacher-based perception rewards, and compare both against the standard RL as well as oracle perception reward. As shown in Table 3, stronger reward-perception correlation yields larger perception gains: the teacher reward nearly matches the oracle reward and clearly outperforms self-rewarding. These results support **reward-perception correlation as a practical diagnostic** for surrogate design, and suggest **surrogate perception rewards as practical mitigations** when sufficiently aligned with ground-truth perception.

Takeaway for RL

Findings: RL under-optimizes perception because outcome rewards are more tightly coupled with reasoning than with perception, yielding a weak signal for perception optimization.

Suggestions: Add perception-aware rewards. Moderately augmenting the outcome reward with a perception reward improves end-to-end performance. When oracle perception rewards are unavailable, reliable surrogate rewards also yield gains, with reward-perception correlation as a simple diagnostic for reward reliability.

4 Related Work

Post-training in VLMs. Post-training is known to incentivize reasoning capabilities in pre-trained VLMs, traditionally via supervised fine-tuning on long-CoT (Xu et al., 2025; Yang et al., 2025; Yao et al., 2026). More recently, the success of GRPO for LLM reasoning (Shao et al., 2024b; Guo et al., 2025; Jaech et al., 2024) has motivated its adaptation to VLMs (Huang et al., 2025; Peng et al., 2025; Meng et al., 2025; Wang et al., 2026; Deng et al., 2025; Zhang et al., 2025), including some frontier VLMs (Bai et al., 2025b,a; Wang et al., 2025c).

However, stronger reasoning does not translate into more reliable perception. For example, Li et al. (2023); Bai et al. (2024); Liu et al. (2024) have documented the hallucination problem of VLMs. For RL training, recent work reports that RL improves reasoning but leaves perception flat or even degraded (Liu et al., 2025; Tian et al., 2025; Yao et al., 2025; Xiao et al., 2025; Wang et al., 2025d). Together, prior work suggests perception as a key bottleneck for VLM post-training.

Mitigation approaches for perception. A growing body of work attempts to mitigate the perception bottleneck. On the SFT side, various work tries to augment CoT differently (Shao et al., 2024a; Wang et al., 2025b; Man et al., 2025; Xu et al., 2025; Li et al., 2025b), but their customized chain-of-thought designs limit their application to generic post-training. On the RL side, mitigations largely focus on perception-aware reward shaping (Xia et al., 2025; Xiao et al., 2025; Li et al., 2025c; Wang et al., 2025d; Chen et al., 2025). In §3.3, we study representative perception-reward designs under our dianostic framework.

Controlled analysis of post-training. Our methodology follows on a line of work that uses controlled synthetic tasks to dissect model training. Such settings make it possible to remove confounders present in real-world benchmarks and to access ground-truth properties of the data-generating process. For example, (Chu et al.) uses synthetic settings to compare SFT memorization with RL generalization; (Li et al., 2025a) probes compositional generalization; and Li et al. (2026) localizes the internal changes induced by RL. Our framework continues this line of work by introducing two synthetic tasks to disentangle reasoning and perception optimized in post-training.

5 Conclusions

This work investigates asymmetric optimization in VLM post-training, where reasoning improves more reliably than perception, causing perception to become a bottleneck for end-to-end visual reasoning. Using a controlled framework with disentangled perception and reasoning, we show that both SFT and RL exhibit a consistent perception-reasoning asymmetry, but through different mechanisms. In SFT, the asymmetry is primarily caused by token imbalance, and loss reweighting—either manually with a fixed perception weight, or dynamically via NGDiff—mitigates this imbalance. In RL, the asymmetry is instead mainly driven by reward coupling, and augmenting training with perception-aware rewards—either ground-truth perception rewards or reliable surrogate rewards—helps reduce the asymmetry. Overall, our findings reveal a pronounced asymmetry induced by default post-training objectives, suggesting that VLM post-training should not rely on these objectives alone. Instead, paradigm-specific interventions are needed to mitigate perception-reasoning asymmetry, strengthen perception learning, and ultimately improve end-to-end visual reasoning.

Limitations

This work studies whether, and why, reasoning and perception improve differently during VLM post-training, using a suite of two synthetic visual reasoning tasks. One possible concern is that the synthetic setting abstracts away from the full diversity of real-world vision-language data. However, this abstraction is central to the purpose of the study rather than a limitation of its validity. Our goal is not to benchmark end-task performance, but to isolate the training dynamics that cause reasoning and perception to be optimized asymmetrically. For this purpose, synthetic tasks provide a uniquely clean experimental setting: they allow full control over the input distribution, canonical perception annotations, and reliable programmatic evaluation. Without such control, errors in perception and reasoning are often entangled, making it difficult to measure each component independently or to identify the mechanisms responsible for their divergence.

As discussed in §4, this style of controlled task design is widely used in prior work to study model behavior while minimizing confounding factors. More importantly, our main claims do not depend on the surface form of the synthetic tasks. The mechanisms we identify, token imbalance in SFT and reward coupling in RL, arise from general properties of post-training objectives rather than from task-specific artifacts. Accordingly, the proposed mitigations, including loss reweighting, dynamic objective balancing, and perception-aware rewards, are not tied to the synthetic benchmark itself. They are practical interventions that can be incorporated into real-world VLM post-training pipelines whenever perception supervision, proxy perception signals, or surrogate rewards are available.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. Preprint, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. Preprint, arXiv:2502.13923.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930.

Yan Chen, Long Li, Teng Xi, Long Zeng, and Jingdong Wang. 2025. Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models. arXiv preprint arXiv:2509.13031.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In Forty-second International Conference on Machine Learning.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Yilun Du, Jiayuan Mao, and Joshua B Tenenbaum. 2024. Learning iterative reasoning through energy diffusion. In International Conference on Machine Learning, pages 11764–11776. PMLR.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Alex Heyman and Joel Zylberberg. 2025. Evaluating the systematic reasoning abilities of large language models through graph coloring. arXiv preprint arXiv:2502.07087.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720.

Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. 2025. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11278–11294.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient

memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles.

Yann LeCun, Corinna Cortes, and CJ Burges. 2010. Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2.

Tianle Li, Jihai Zhang, Yongming Rao, and Yu Cheng. 2025a. Unveiling the compositional ability gap in vision-language reasoning model. arXiv preprint arXiv:2505.19406.

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

Yifan Li, Zhenghao Chen, Ziheng Wu, Kun Zhou, Ruipu Luo, Can Zhang, Zhentao He, Yufei Zhan, Wayne Xin Zhao, and Minghui Qiu. 2025b. Unleashing perception-time scaling to multimodal reasoning models. CoRR, abs/2510.08964.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 conference on empirical methods in natural language processing, pages 292–305.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. 2025c. Self-rewarding vision-language model via reasoning decomposition. Preprint, arXiv:2508.19652.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In International Conference on Learning Representations, volume 2024, pages 39578–39601.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. arXiv preprint arXiv:2505.21523.

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

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.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 14268–14280.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems, 37:8612–8642.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv:2409.19256.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. 2025. More thought, less accuracy? on the dual nature of reasoning in vision-language models. arXiv preprint arXiv:2509.25848.

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

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2026. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. Advances in Neural Information Processing Systems, 38:30865–30891.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and 1 others. 2025b. Vgr: Visual grounded reasoning. arXiv preprint arXiv:2506.11991.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025c. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and 1 others. 2025d. Perception-aware policy optimization for multimodal reasoning. arXiv preprint arXiv:2507.06448.

Weiyun1025. 2025. verl-internvl. https://github.com/Weiyun1025/verl-internvl. GitHub repository. Accessed: 2026-05-25.

Juncheng Wu, Hardy Chen, Haoqin Tu, Xianfeng Tang, Freda Shi, Hui Liu, Hanqing Lu, Cihang Xie, and Yuyin Zhou. 2026. From seeing to thinking: Decoupling perception and reasoning improves post-training of vision-language models. arXiv preprint arXiv:2605.20177.

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

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2025. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. Preprint, arXiv:2506.07218.

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

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2376–2385.

Guanyu Yao, Qiucheng Wu, Yang Zhang, Zhaowen Wang, Handong Zhao, and Shiyu Chang. 2025. Rethinking the text-vision reasoning imbalance in mllms through the lens of training recipes. arXiv preprint arXiv:2510.22836.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2026. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. Advances in Neural Information Processing Systems, 38:29918–29952.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15134–15186.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. 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. Association for Computational Linguistics.

A Responsible Research Checklist

A.1 Used Scientific Artifacts

1. `Qwen/Qwen3-VL-2B-Instruct` (Bai et al., 2025a) for post-training, released under Apache-2.0 license;
2. `OpenGVLab/InternVL3_5-2B` (Wang et al., 2025c) for post-training, released under Apache-2.0 license;
3. `LlamaFactory` (Zheng et al., 2024) for SFT training, released under Apache-2.0 license;
4. `verl` (Sheng et al., 2024) for training the `Qwen3-VL` model with RL, released under Apache-2.0 license;
5. `verl-internvl` (Weiyun1025, 2025) for training the `InternVL3.5` model with RL, released under Apache-2.0 license;
6. `vllm` (Kwon et al., 2023) for inference and evaluation, released under Apache-2.0 license;
7. The graph coloring dataset (Heyman and Zylberberg, 2025) released under MIT license is adapted for building our graph coloring task;
8. The Sudoku dataset (Du et al., 2024) released under MIT license is adapted for building our Sudoku task;

9. MNIST images (LeCun et al., 2010) released under MIT license for rendering Sudoku images.

We have reviewed the licenses and intended-use conditions of all existing artifacts we used. The models and training frameworks we adopt are released under Apache-2.0 licenses, and the datasets we adapt are released under MIT licenses. Our use is fully within a research context: the models, frameworks, and datasets are used only for research-oriented post-training, supervised fine-tuning, reinforcement learning, and benchmark/-task construction.

A.2 Created Scientific Artifacts

Code. The code artifacts include: (1) training code for our targeted interventions, built on top of `LlamaFactory` and `verl`; and (2) detailed scripts for data processing, supervised fine-tuning, reinforcement learning, evaluation, and result reproduction.

Datasets. The dataset artifacts cover both GC and Sudoku. Each task includes an SFT training dataset, an RL training dataset, and a test set. All examples are in English. Since the tasks are synthetic or puzzle-based, they do not represent natural-language demographic groups, human subjects, or personally identifiable information. Therefore, demographic coverage is not applicable. Statistics are reported in Table 5.

Task	# Test	# SFT Train	# RL Train
GC	500	2,000	16,000
Sudoku	500	2,000	16,000

Table 5: Statistics for the created dataset artifacts. The tasks are synthetic reasoning tasks in English and do not involve human subjects, demographic groups, or personally identifiable information.

For GC, we curate random graphs with 7, 8, or 9 nodes and render them as images using `networkx`. For Sudoku, we adapt the problem generation code from (Du et al., 2024) and render images by combining MNIST digit images (LeCun et al., 2010). Detailed dataset synthesis scripts will also be released under the MIT license.

Release plan. We will release all artifacts upon publication. The code will be released under the Apache-2.0 license, and the data will be released under the MIT license. For components adapted from existing artifacts, we will preserve the required copyright, license, and attribution notices.

A.3 AI Assistants In Research Or Writing

I used AI assistants solely for stylistic improvements in writing, such as improving clarity, grammar, and phrasing. I did not use AI assistants for coding, brainstorming, research design, data analysis, interpretation of results, or any other critical intellectual contribution.

B Experimental Details and Additional Results

B.1 Setups for Main Experiments

SFT training. We perform SFT using the `LlamaFactory` package. The hyperparameter settings are reported in Table 6. Each experiment is trained on 4 $\times$ 80GB A100 GPUs with `per_device_train_batch_size` set to 2 and `gradient_accumulation_steps` set to 1, giving an effective batch size of 8. Each run takes approximately 3 hours.

Hyperparameter	Value
Effective batch size	8
`per_device_train_batch_size`	2
`gradient_accumulation_steps`	1
`learning_rate`	`1.0e-6`
`lr_scheduler_type`	`cosine`
`warmup_ratio`	0.1
`bf16`	`true`
`max_steps`	500

Table 6: Hyperparameter settings for SFT.

RL training. For the `QwenVL` model, we perform RL training using the `verl` package. The hyperparameter settings are reported in Table 7. Each experiment is trained on 4 $\times$ 80GB A100 GPUs, with `ppo_micro_batch_size_per_gpu` set to 10 and `log_prob_micro_batch_size_per_gpu` set to 80. Each GC run takes approximately 30 hours, while each Sudoku run takes approximately 50 hours, mainly because Sudoku examples have longer output sequences.

For the `InternVL` model, due to compatibility issues, we use the `verl-internvl` implementation from `verl-internvl` (Weiyun1025, 2025) and largely follow its default hyperparameter settings to ensure stable performance.

Hyperparameter	Value
train_batch_size	128
lr	1e-6
ppo_mini_batch_size	128
use_kl_loss	true
kl_loss_coef	0.01
total_training_steps	100

Table 7: Hyperparameter settings for QwenVL RL.

The hyperparameter settings are reported in Table 8. We also train on 4 × 80GB A100 GPUs, with use_dynamic_bsz set to true, ppo_max_token_len_per_gpu set to 20480, and log_prob_micro_batch_size_per_gpu set to 16. Training is faster than QwenVL RL: each GC run takes approximately 18 hours, while each Sudoku run takes approximately 30 hours.

Hyperparameter	Value
lr	1e-6
ppo_mini_batch_size	128
ppo_micro_batch_size_per_gpu	16
use_kl_loss	false
loss_mode	gspo
rollout.n	5
total_training_steps	100

Table 8: Hyperparameter settings for InternVL RL.

Inference and evaluation. For all experiments, we run inference using the vllm package with max_tokens=10000, temperature=0.7, top_p=1.0, top_k=-1. Standard inference and counterfactual inference for measuring reasoning accuracy each take approximately ~ 30 min on a single A100 GPU. Evaluation is performed using rule-based scripts based on ground-truth perception annotations. For all experiments, we report a single run due to the high computational cost. However, since each task uses a large controlled test set ($N = 500$), we expect the reported results to provide reliable estimates of model performance.

B.2 Full Results

In the main paper, we present representative results for selected model-task settings. Here, we provide the corresponding results across all four settings: QwenVL-GC, InternVL-GC, QwenVL-Sudoku, and InternVL-Sudoku. Specifically, Figure 12 extends the SFT optimization curves in Figure 3, Figure 13 extends the SFT loss-reweighting trade-off in Figure 4, Figure 14 extends the stronger-perception-initialization results in Figure 6, Figure 15 extends the RL optimization curves in Figure 9, and Figure 16 extends the RL reward-augmentation trade-off in Figure 10.

B.3 Token Reweighting for RL

In the main text, we show that token imbalance explains the perception-reasoning asymmetry in SFT, but not in RL. Here, we describe the RL token-reweighting intervention in more detail and provide the full optimization trajectory.

Recall that standard GRPO optimizes the token-averaged objective

$$\mathcal{J}_{\text{GRPO}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{y}^{(i)}|}\sum_{t=1}^{|\mathbf{y}^{(i)}|}\rho_{i,t}(\theta)\hat{A}_{i,t}\right],$$

where each response is decomposed into perception tokens $\mathbf{p}^{(i)}$ and reasoning tokens $\mathbf{r}^{(i)}$, i.e., $\mathbf{y}^{(i)} = [\mathbf{p}^{(i)}, \mathbf{r}^{(i)}]$. To mirror the SFT loss-reweighting intervention, we first decompose the GRPO objective into perception and reasoning components:

$$\mathcal{J}_p = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{p}^{(i)}|}\sum_{y_t\in\mathbf{p}^{(i)}}\rho_{i,t}(\theta)\hat{A}_{i,t}\right],$$

$$\mathcal{J}_r = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{r}^{(i)}|}\sum_{y_t\in\mathbf{r}^{(i)}}\rho_{i,t}(\theta)\hat{A}_{i,t}\right].$$

We then define the reweighted GRPO objective as

$$\mathcal{J}_{\text{GRPO},\lambda} = \lambda\mathcal{J}_p + (1-\lambda)\mathcal{J}_r,$$

where $\lambda \in [0, 1]$ controls the relative emphasis on perception-token updates.

In practice, we implement this objective by keeping the original token-averaged GRPO form and rescaling the token-level advantages. Specifically, we define

$$\hat{A}_{i,t}^{(\lambda)} = \begin{cases} \lambda \cdot \dfrac{|\mathbf{y}^{(i)}|}{|\mathbf{p}^{(i)}|}\hat{A}_{i,t}, & y_t \in \mathbf{p}^{(i)}, \\ (1-\lambda)\cdot\dfrac{|\mathbf{y}^{(i)}|}{|\mathbf{r}^{(i)}|}\hat{A}_{i,t}, & y_t \in \mathbf{r}^{(i)}. \end{cases}$$

Substituting $\hat{A}_{i,t}^{(\lambda)}$ into the standard token-averaged objective,

$$\mathcal{J}_{\text{GRPO},\lambda} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{y}^{(i)}|}\sum_{t=1}^{|\mathbf{y}^{(i)}|}\rho_{i,t}(\theta)\hat{A}_{i,t}^{(\lambda)}\right],$$

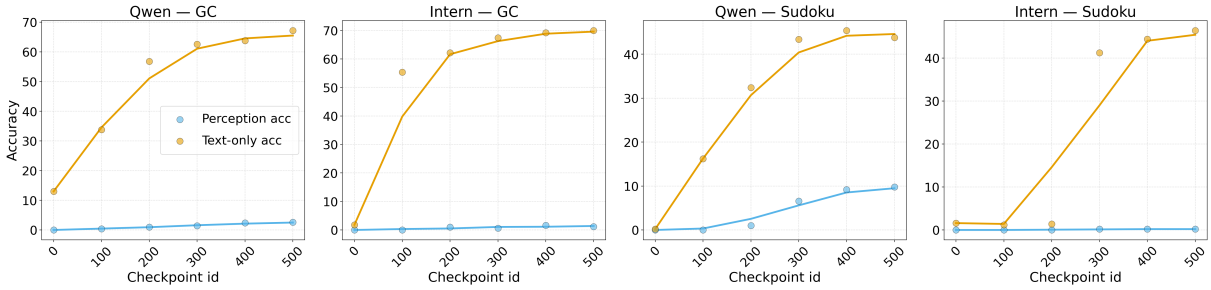

Figure 12: Full SFT optimization curves across all four model-task settings. This figure extends Figure 3 in the main text.

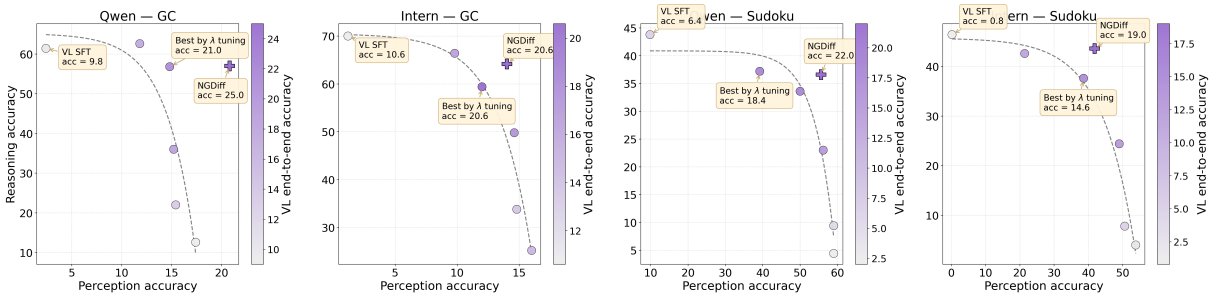

Figure 13: Full perception-reasoning trade-off results for SFT loss reweighting across all four model-task settings. This figure extends Figure 4 in the main text.

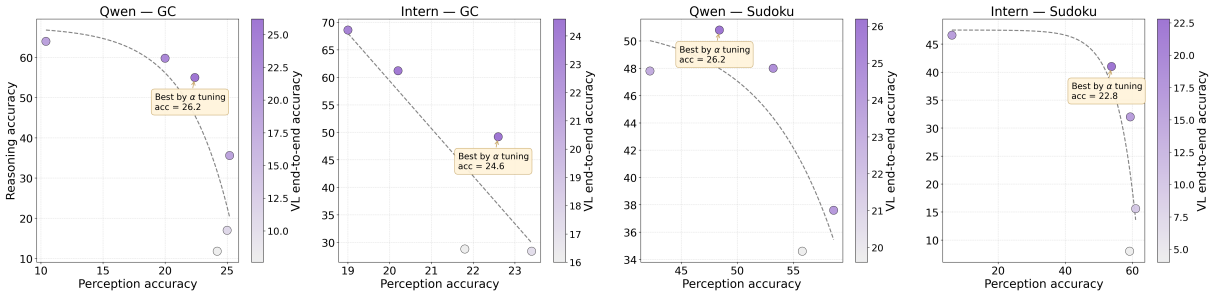

Figure 14: Full perception-reasoning trade-off results after stronger perception initialization across all four model-task settings. This figure extends Figure 6 in the main text.

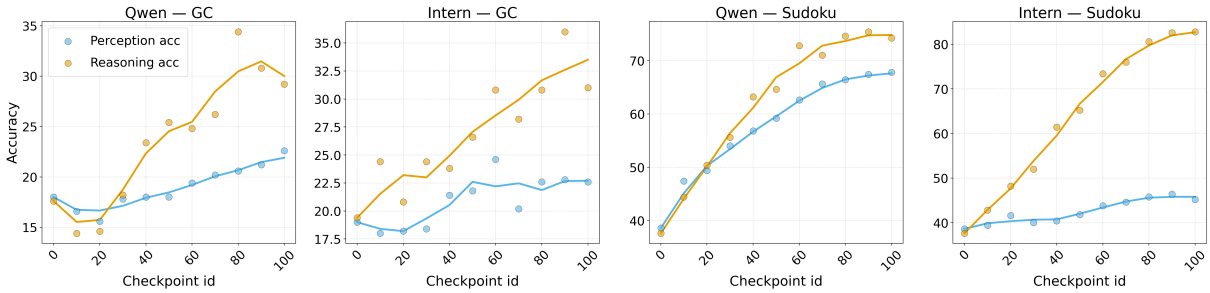

Figure 15: Full RL optimization curves across all four model-task settings. This figure extends Figure 9 in the main text.

recovers the decomposed reweighted objective above. Thus, although the intervention is conceptually an objective-level reweighting between perception and reasoning terms, our implementation applies it through advantage rescaling.

Figure 17 shows the optimization trajectories under different values of λ. Unlike in SFT, token reweighting does not mitigate the perception-reasoning asymmetry in RL. Instead, increasing the perception-token weight destabilizes training

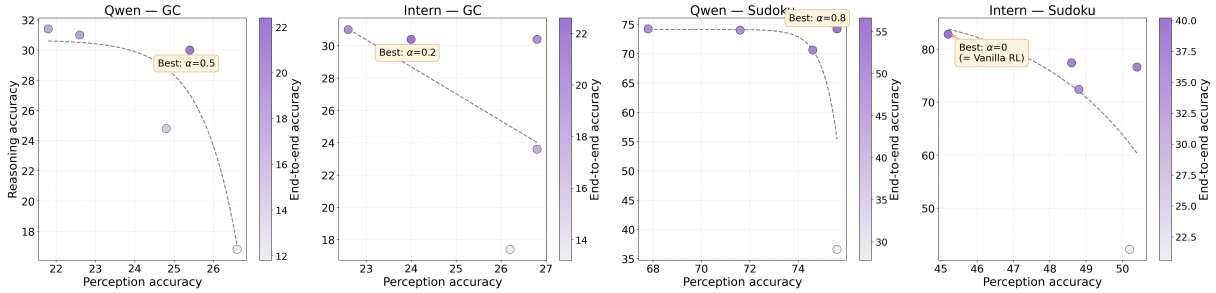

Figure 16: Full perception-reasoning trade-off results for RL reward augmentation across all four model-task settings. This figure extends Figure 10 in the main text.

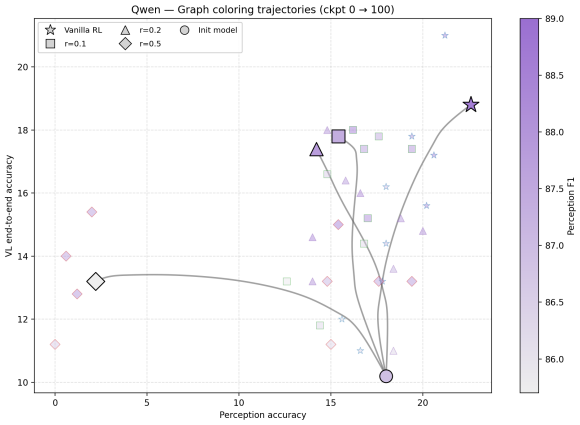

Figure 17: Optimization trajectories of RL token reweighting. Unlike SFT loss reweighting, directly increasing the weight on perception-token policy-gradient terms does not improve perception or end-to-end accuracy, and instead degrades training. This indicates that RL asymmetry is not primarily driven by token imbalance.

and degrades both perception and end-to-end performance. This suggests that the RL asymmetry is not primarily caused by the number of perception tokens, even though perception remains a small fraction of the output sequence. Rather, the dominant issue is reward coupling: the outcome reward provides a weaker and noisier learning signal for perception than for reasoning.

B.4 Surrogate Perception Rewards

CLIP-based perception reward. For the CLIP-based surrogate, following Chen et al. (2025), we score the generated perception by its image-text similarity to the input image. Specifically, we encode the image and the extracted perception segment with CLIP and use their cosine similarity as the surrogate perception reward. This implementation is lightweight and does not require additional rollouts or teacher queries. However, because CLIP similarity is coarse-grained, it may fail to distinguish fine-grained structural errors in canonical perception representations, such as missing graph edges or incorrect Sudoku entries.

Self-rewarding perception reward. Following Li et al. (2025c), we construct a self-reward by testing whether the model's generated perception is sufficient for downstream reasoning. Specifically, for each generated CoT, we extract the perception segment and run a secondary rollout in which the model reasons only from this textual perception representation, without access to the original image. We then use the final-answer accuracy of this secondary rollout as the surrogate perception reward. The original self-rewarding setup uses a single rollout ($N = 1$). We additionally evaluate a higher-budget variant with $N = 4$ secondary rollouts per perception and use the average final-answer accuracy as the reward. This provides a more reliable surrogate signal, at the cost of additional inference budget. For training, we build on the official Vision-SR1 codebase.

Teacher-based perception reward. We also implement a teacher-based surrogate reward inspired by Xiao et al. (2025). Since our tasks use simplified canonical perception representations, we do not require a teacher-as-judge formulation. Instead, we directly query the teacher model, `gemini-3-flash-preview`, to generate a canonical perception representation from the image, and compute the reward by comparing the model's predicted perception against this teacher-generated representation. This yields a perception reward in the same format as our ground-truth perception reward, allowing us to use the standard RL training codebase without additional judge-model infrastructure.