SELF-REWARDING VISION-LANGUAGE MODEL VIA REASONING DECOMPOSITION AND MULTI-REWARD POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) often suffer from visual hallucinations – generating things that are not consistent with visual inputs – and language shortcuts, where they skip the visual part and just rely on text priors. These issues arise because most post-training methods for VLMs rely on simple verifiable answer matching and supervise only final outputs, leaving intermediate visual reasoning without explicit guidance. As a result, VLMs receive sparse visual signals and often learn to prioritize language-based reasoning over visual perception. To mitigate this, some existing methods add visual supervision using human annotations or distilled labels from external large models. However, human annotations are labor-intensive and costly, and because external signals cannot adapt to the evolving policy, they cause distributional shifts that can lead to reward hacking.

In this paper, we introduce Vision-SR1, a self-rewarding reinforcement learning method that improves visual reasoning without relying on external visual supervisions. Vision-SR1 decomposes VLM reasoning into two stages: *visual perception* and *language reasoning*. The model is first prompted to produce self-contained visual perceptions that are sufficient to answer the question without referring back the input image. To validate this self-containment, the same VLM model is then re-prompted to perform language reasoning using only the generated perception as input to compute reward. This self-reward is combined with supervision on the final outputs through multi-reward policy optimization, providing a balanced training signal that strengthens both visual perception and language reasoning. In addition, Multi-Reward Policy Optimization separately computes advantages and log probabilities for both the visual reasoning reward and the answer accuracy reward. The method then calculates KL divergence regularization and actor loss using the combined sum of these two reward components. Our experiments show that Vision-SR1 improves visual reasoning, mitigates visual hallucinations, and reduces reliance on language shortcuts across diverse vision-language tasks.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) have progressed by integrating pre-trained language models and vision encoders with instruction tuning (Liu et al., 2023b; et al, 2024; Chen et al., 2024; Bai et al., 2025; Li et al., 2025d). Despite these successes, a critical limitation remains in their reasoning capabilities: VLMs often produce visual hallucinations – descriptions of content that is not actually present in the image (Guan et al., 2024; Liu et al., 2024; Li et al., 2025e; Liu et al., 2023a) – or rely on language shortcuts, where the model bypasses visual understanding and instead depends solely on text priors (Si et al., 2022; Bleeker et al., 2024). Very recently, R1-style reinforcement learning (RL) methods have been shown to improve the reasoning abilities of VLMs across diverse tasks (Huang et al., 2025b; Shen et al., 2025; Xia et al., 2025; Zhang et al., 2025). However, these methods often encourage "thinking over seeing" that lean heavily on language reasoning while underutilize visual perception (Liu et al., 2025; Yao et al., 2025). This imbalance makes VLMs susceptible to reward hacking (Fu et al., 2025) and spurious effects (Shao et al., 2025) observed in RL training. Although VLMs trained with RL often have apparent improvements, they can largely re-

flect probability shifts toward the style of training and test data, leading to language shortcut answers from prior knowledge and overlooking hallucination risks (Li et al., 2025b).

In essence, most existing post-training methods for VLMs rely on simple verifiable answer matching and thus lack explicit supervision of visual information. As a result, VLMs receive sparse visual signals and often learn to prioritize language-based reasoning over visual perception. To mitigate this, some methods introduce intermediate visual supervision through human annotations (Thawakar et al., 2025) or distilled labels (e.g., pre-extracted key steps) from external models (Xu et al., 2024; Zhang et al., 2025; Xiao et al., 2025; Xia et al., 2025; Lu et al., 2025). However, these solutions face significant limitations. Human annotations are labor-intensive, costly, and difficult to scale across multimodal tasks, while distilled signals inherit biases from source models and often fail to generalize across diverse domains. Moreover, distributional shifts between fixed intermediate signals and the continually updated policy can lead to reward hacking (Gao et al., 2023). Most importantly, both approaches remain limited by their reliance on external supervision and their use of single advantage and log probability computations across multiple intermediate rewards, , restricting their scalability and applicability.

In this paper, we introduce **Vision-SR1**, a reinforcement learning framework that encourages VLM to produce *self-contained* visual reasoning that can be verified by the VLM itself, without external supervision. Vision-SR1 decomposes the reasoning process into two stages: *visual perception* and *language reasoning*. The visual perception is required to capture all details relevant to answering the query, so that the reasoning stage can proceed without re-accessing the original image. We explicitly compute advantages and rollouts separately for each stage, then calculate individual Actor policy losses and KL divergence terms for the visual perception and language reasoning stages before combining them into a unified training objective.

The training has two **rollout passes** and one **training objective optimization** of the same VLM:

– First pass (standard rollout): $(\text{Image}, \text{Query}) \rightarrow (\text{Visual Perception}, \text{CoT Reasoning}, \text{Answer})$

- The model generates a structured output that explicitly separates visual perception, chain-of-thought (CoT) reasoning, and the final answer.
- An **accuracy reward** is computed by comparing the final answer with the ground truth.

– Second pass (self-reward rollout): $(\text{Query}, \text{Visual Perception}) \rightarrow (\text{CoT Reasoning}, \text{Answer})$

- The model is re-prompted to reason using only the generated perception (without re-accessing the original image). If the correct answer is derived, the perception is considered **faithful**, and a **self-visual reward** is assigned.

– Multi-Reward Policy Optimization (objective optimization):

- The multi-reward policy optimization enables the policy model to receive distinct feedback for visual reasoning quality and answer accuracy through separate advantage computations and rollout-specific loss terms.

These separate reward signals are then combined through our multi-policy loss objective to provide balanced training that strengthens both visual perception and language reasoning without the entangled learning signals of traditional reward summation. Experiments show that Vision-SR1 improves visual reasoning, mitigates hallucinations, and reduces language shortcuts across diverse vision-language tasks.

2 METHOD

We build upon advancement of Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for improving VLM reasoning. We first review the key concepts then introduce our method.

2.1 PRELIMINARY: REINFORCEMENT LEARNING FOR VLM WITH VERIFIABLE REWARD

We denote a pre-trained VLM as a policy model π to be optimized in reinforcement learning. Given a multimodal question (Q) consists of an image i and a text question q, where $Q = \{i, q\}$, the policy model π generates a reasoning response s. We use GRPO to optimize the response s for the policy model. For each multimodal question $Q = \{i, q\}$ we sample a *group* of K candidate responses

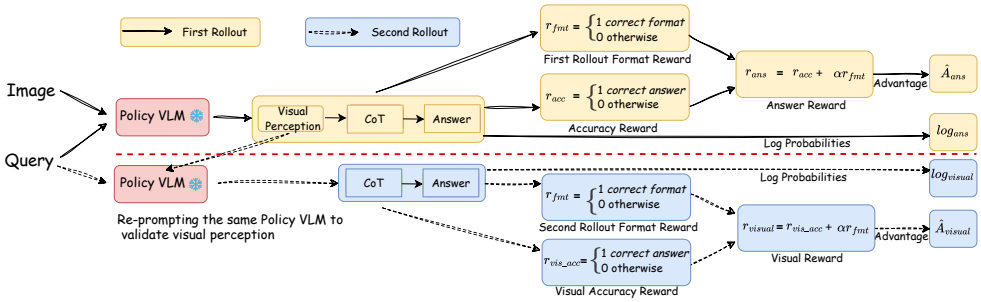

Figure 1: Overall framework of Vision-SR1. During RL training, the VLM performs two rollouts. In the first pass, the model takes an image–query pair and generates a structured output (visual perception, CoT reasoning, and answer), with answer reward computed against the ground truth. In the second pass, the model is re-prompted to answer using only query and its generated visual perception. If the correct answer is derived, a self-visual reward is assigned. We compute the advantages and log probabilities for each rollout for Multi-Reward Policy Optimization.

$\mathcal{S}_Q = \{s_1, \ldots, s_K\},\ s_k \sim \pi_\theta(\cdot|Q)$. Each response is scored by a scalar reward $r(Q, s_k)$ (defined in Sec. 2.2), and we compute a *group-relative* advantage

$$\hat{A}^{\text{grp}}(Q, s_k) \;=\; r(Q, s_k) \;-\; \frac{1}{K}\sum_{j=1}^{K} r(Q, s_j), \tag{1}$$

which centres rewards within the group, removing question-level biases while retaining pairwise preferences. We update the policy by maximizing

$$\mathcal{L}_{\text{GRPO}}(\theta) \;=\; \mathbb{E}_{Q\sim\mathcal{D}}\Big[\sum_{k=1}^{K} \hat{A}^{\text{grp}}(Q, s_k)\ \log \pi_\theta(s_k \,|\, Q) \;-\; \beta\ \text{KL}\big(\pi_\theta(\cdot|Q)\ \|\ \pi_{\theta_0}(\cdot|Q)\big)\Big], \tag{2}$$

where π_{θ_0} is the frozen, pre-trained reference model and β controls the strength of the KL penalty that keeps the updated policy close to its original behavior.

The group-centred baseline in equation 1 guarantees $\sum_k \hat{A}^{\text{grp}}(Q, s_k) = 0$, thereby reducing the variance of policy-gradient estimates without requiring an external value critic.

2.2 Steps 1: Self-Rewarding VLM via Reasoning Decomposition

As we discussed, incorporating intermediate visual supervision can strengthen the reasoning ability of VLMs. However, existing methods suffer from key limitations: methods based on human annotations are labor-intensive and costly (Thawakar et al., 2025), while approaches that distill supervision from external models provide static signals that cannot adapt as the policy model itself evolves during training (Zhang et al., 2025; Xiao et al., 2025; Xia et al., 2025). To overcome these issues, we introduce a self-rewarding framework that enables the VLM to reward its own visual perception. The key idea is to decompose the visual reasoning process into structured components, i.e., the VLM first produces a self-contained visual perception and then assesses whether this perception is sufficient for produce the final answer. This decomposition reduces reliance on external supervision and allows the reward signal to adapt dynamically as the model improves.

Decomposed VLM Reasoning. To encourage the VLM to perform self-contained visual reasoning, we require every response to adhere to a ***See-Think*** generation format (Jia et al., 2024; Xia et al., 2025) format. Specifically, for a vision-language task, $Q = \{i, q\}$ where i is the input image and q is the textual query, the model produces the following structured output:

$$\langle\text{visual\_reasoning}\rangle\, c\, \langle/\text{visual\_reasoning}\rangle\ \|\ \langle\text{think}\rangle\, t\, \langle/\text{think}\rangle\ \|\ \langle\text{answer}\rangle\, a\, \langle/\text{answer}\rangle$$

where c is a *self-contained* visual perception that captures all visual information necessary to solve the task, so that the following language reasoning can proceed without re-accessing the original input image. Besides, t is the language reasoning trace, and a denotes the final answer.

Self-Reward for Visual Reasoning. A challenge is judging whether the visual perception c is *self-contained* – i.e. whether it encodes *all* the visual information needed to answer the question $Q = \{i, q\}$ correctly. To address this, our idea is to treat the visual perception as a *text-only proxy* for the image and validate it by re-prompting the VLM itself to perform language reasoning using only the generated perception as input. If the model can derive the correct answer from (c, q) alone, we consider c to be visually faithful and assign the corresponding visual reward.

$$\hat{a} \;=\; f_\theta\big(c, q\big), \quad r_{\text{visual}}(Q, c) \;=\; \mathbb{I}\big[\hat{a} = a^*\big], \tag{3}$$

where a^* is the ground-truth answer. Instead of using an external reward model, we leverage the policy model's own reasoning ability for self-evaluation. The model itself determines the reward by answering the question using only its generated visual reasoning (Figure 1).

Reward Composition. The reward composition comprises three *aligned* components, each conditioned on the question $Q = \{i, q\}$:

• **Format reward** $r_{\text{fmt}}(s)$: measures whether the response strictly follows the required layout. This reward is being applied to both Visual reward and Accuracy reward with respect to their format requirements.

• **Answer reward** $r_{\text{ans}}(Q, a)$: measures the correctness of the final answer (r_{acc}) *plus* the corresponding format reward. Because a is generated after the reasoning trace t, the term implicitly rewards CoT reasoning. This is computed at first rollout with hyper-parameters $(0 \leq \alpha \leq 1)$:

$$r_{ans}(Q, a) = \; r_{\text{acc}}(Q, a), + \alpha \, r_{\text{fmt}}(s) \tag{4}$$

• **Visual reward** $r_{\text{visual}}(Q, c)$: measures whether the visual reasoning output is self-contained, i.e., sufficient to answer the question without image (r_{vis_acc}) *plus* corresponding format reward. A reward of 1 is assigned if, given only the question and the visual reasoning, the VLM can give the correct answer. This is computed at second rollout:

$$r_{visual}(Q, a) = \; r_{\text{vis\_acc}}(Q, a), + \alpha \, r_{\text{fmt}}(s) \tag{5}$$

2.3 Step 2: Multi-Reward Optimization with Multi-Advantage Loss Computation.

Simply *summing* the visual reasoning reward and the final-answer accuracy reward could produce a *sparse and entangled* learning signal: the policy has little to tell which rollout was responsible for which part of the scalar return. To disentangle visual reasoning and answer accuracy assignment, we keep the two rollouts—answer generation and visual reasoning—*separate* throughout the update. Each rollout receives its own log-probabilities, advantage, and KL term, and the gradients are combined only at the very end. This turns the single *multi-reward* problem into two single-reward sub-problems that share parameters with individually optimized feedback.

Reward-Specific Log-Probability Tracking. During sampling we cache the behavioral log probabilities for every token in each rollout:

$$\log \pi^{(i)}_{\text{old}}\big(a_{\text{ans},t}\big), \; \log \pi^{(i)}_{\text{old}}\big(a_{\text{visual},t}\big),$$

where $a_{\text{ans},t}$ is the action at step t of the first rollout, and $a_{\text{visual},t}$ is the action (token) at step t of the second rollout. At update time we compute the corresponding $\log \pi_\theta$ under the current parameters to compute the policy and KL losses.

Group-wise Z-Score Advantage. For each reward we follow GRPO to compute the advantage:

$$A^{(i)}_{\text{ans}} \;=\; \frac{r^{(i)}_{\text{ans}} - \mu_{\text{ans}}}{\sigma_{\text{ans}} + \varepsilon}, \qquad A^{(i)}_{\text{visual}} \;=\; \frac{r^{(i)}_{\text{visual}} - \mu_{\text{visual}}}{\sigma_{\text{visual}} + \varepsilon}, \tag{6}$$

with means and standard deviations $\mu_{\text{ans}} = \frac{1}{B}\sum_i r^{(i)}_{\text{ans}}$, $\sigma^2_{\text{ans}} = \frac{1}{B}\sum_i \big(r^{(i)}_{\text{ans}} - \mu_{\text{ans}}\big)^2$, where B is the rollout batch size (and analogously for the visual group). Broadcasting A_{ans} to all caption tokens and A_{visual} to all answer tokens gives two advantage masks that weight the corresponding log-probabilities during backpropagation for each sub-task.

Actor Loss (Policy Gradient Loss). The actor loss computes weighted policy gradients for the two reward signals (answer and visual), with separate coefficients λ_{ans} and λ_{visual} indicating their contributions.[1]

$$\mathcal{L}_{\text{actor}} = -\frac{\lambda_{\text{ans}}}{B} \sum_{i,t} A^{(i)}_{\text{ans},t} \log \pi_\theta(a^{(i)}_{\text{ans},t}) - \frac{\lambda_{\text{visual}}}{B} \sum_{i,t} A^{(i)}_{\text{visual},t} \log \pi_\theta(a^{(i)}_{\text{visual},t}) \tag{7}$$

KL Divergence Regularization Loss. The KL regularization applies separate penalty coefficients β_{cap} and β_{ans} to prevent excessive policy deviation for each reward component.

$$\mathcal{L}_{\text{KL}} = \frac{\beta_{\text{ans}}}{B} \sum_{i=1}^{B} \sum_{t} \left[\log \pi_{\text{old}}(a^{(i)}_{\text{ans},t}) - \log \pi_\theta(a^{(i)}_{\text{ans},t})\right] + \frac{\beta_{\text{visual}}}{B} \sum_{i=1}^{B} \sum_{t} \left[\log \pi_{\text{old}}(a^{(i)}_{\text{visual},t}) - \log \pi_\theta(a^{(i)}_{\text{visual},t})\right] \tag{8}$$

Multi-Reward Loss Objective. The total loss combines multi-reward policy gradients with component-specific regularization to optimize the model across both reward signals.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{actor}} + \mathcal{L}_{\text{KL}} \tag{9}$$

2.4 Theoretical Analysis

We analyze why Multi-Reward Policy Optimization with separate advantage computation could improve VLM RL training compared to using only answer rewards. In standard RL training, the objective depends solely on final answer correctness:

$$\nabla_\theta \, \mathbb{E}_{s\sim\pi_\theta}[r_{\text{ans}}(a, a^*)] \tag{10}$$

where $s = (t, a)$ contains visual reasoning and language reasoning trace t and final answer a. Since r_{ans} only measures whether a matches ground truth a^*, the intermediate visual reasoning t receives no direct supervision signal. For VLMs, the stronger LLM backbone dominates generation of t, and continued RL training leads to potential reward hacking where the model exploits language priors to achieve correct answers without visual grounding Pantazopoulos & Özyiğit (2025).

Multi-Reward Loss Decomposition. We decompose the loss computation itself into separate components as shown in Equation 9, where the actor loss handles visual and answer components separately (with the KL regularization term following similar component-wise structure):

$$\mathcal{L}_{\text{actor}} = -\lambda_{\text{ans}}\mathbb{E}[A_{\text{ans}} \log \pi_\theta(a_{\text{ans}})] - \lambda_{\text{visual}}\mathbb{E}[A_{\text{visual}} \log \pi_\theta(a_{\text{visual}})] \tag{11}$$

Since each advantage is computed from different reward components (visual and answer) shown in Equation 6, this approach creates clear gradient paths from each reward to its corresponding components, enabling independent optimization of visual reasoning and language reasoning capabilities (Zhu et al., 2025; Lyu et al., 2025).

From an information-theoretic perspective, let I denote the visual input, Q the question, C the visual reasoning representation, and A the final answer. Mutual information $I(U; V)$ measures how much knowing U reduces uncertainty about V (Shannon, 1948). If training relies only on r_{ans}, the model primarily maximizes $I(A; Q)$—making answers strongly dependent on the question—while neglecting $I(A; I)$, the dependence of answers on the visual input. This permits shortcut solutions that bypass perception. By additionally optimizing r_{visual}, which enforces high $I(C; I)$, the model strengthens the path from I to A, thereby increasing $I(A; I)$ and ensuring that answers remain grounded in visual reasoning rather than language-only correlations.

2.5 Data Preparation

Vision-SR1-47K. Our RL dataset consists of approximately 47K examples collected from 24 open-source VLM benchmarks. It spans three key reasoning domains (Figure 1): mathematical reasoning (30.5%), which strengthens quantitative and logical abilities; commonsense knowledge (30%); and general visual understanding (39.5%), which grounds the model in visual question answering.

3 Experiments

[1] We use 0.5 for λ_{ans} and λ_{visual}.

To implement our Vision-SR1, we use **Qwen2.5-VL-3B** and **7B**, Mimo-7B-VL Team et al. (2025) as base models. Additionally, we use Lora finetuning to train Qwen2.5-VL-72B to evaluate the method's generalization to larger model sizes and Lora finetuning.[2] We train the base model with GRPO. The RL phase is trained for 1 epoch on the Vision-SR1-47K dataset. During training, the policy model first generates visual reasoning from the input image, then produces language reasoning and final answer. We then compute a self-reward for visual reasoning by re-prompting the frozen policy model to answer the question using only its generated visual reasoning, without access to the original image i. Finally, we compute advantages and log probabilities separately for each reward component and combine them in the final loss (Figure 1).[3]

Table 1: Vision-SR1-47K data comprises three domains—Math, Knowledge, and General Visual Reasoning—providing diverse supervision for VLM generalization and adaptation.

Category	Included Datasets	Size	(%)
Math	CLEVR-Math, GeoQA+, UniGeo, GEOS, Geometry3K, Super-CLEVR	14K	30.5%
Science Knowledge	TQA, ScienceQA, AI2D, PMC-VQA, VQA-RAD, EXAMS-V-train	14K	30%
General Visual Reasoning	ChartQA, DVQA, PlotQA, FigureQA, MapQA, TabMWP, A-OKVQA, IconQA, visual7w, OpenSpaces, SpaceIlava	18K	39.5%

3.1 Baseline Methods

Vision-R1 (Huang et al., 2025b): The first R1-style reinforcement learning approach, which relies solely on answer rewards as the training signal. However, since the original Vision-R1 was trained only on math-domain data and performs poorly on general-domain reasoning, we reproduce it using our 47K dataset to ensure a fair comparison.

Perception-R1 (Xiao et al., 2025): Similar in training style to Vision-R1, but incorporates pre-extracted visual annotations as an additional reward signal. These visual annotations are derived from a state-of-the-art proprietary multimodal LLM (not specified in the paper).

Visionary-R1 (Xia et al., 2025): Trained to produce a caption–reason–answer output format during RL, where the supervision signal comes from an external text-only LLM (not specified in the paper).

For fair comparisons, we only re-train Vision-R1 on our 47K dataset, since both Perception-R1 and Visionary-R1 require access to external annotations or supervision signals, which are undisclosed.

3.2 Benchmarks and Metrics

Our evaluation covers three areas to evaluate VLMs abilities. Specifically, the domains include (1) general visual understanding, (2) multimodal math reasoning (3) visual hallucination detection.

General Visual Understanding. We evaluate general visual understanding across five diverse benchmarks. **MMMU** (Yue et al., 2024) tests cross-modal reasoning and subject knowledge with 11.5K college-level, four-choice questions spanning six disciplines. **MMMU-Pro** (Yue et al., 2025) increases the difficulty with ten choices per question and adds a challenging *vision-only* setting, where all text is embedded within the image to necessitate robust visual parsing. **RealWorldQA** (xAI, 2024) features ∼700 real-world images from vehicle captures, paired with spatially grounded questions that require verifiable answers. **VisNumBench** (Weng et al., 2025) specifically targets visual number sense through ∼1.9K questions covering seven numerical attributes and four estimation tasks.

Multimodal Mathematical Reasoning. We assess mathematical reasoning using two specialized benchmarks. **MathVerse** (Zhang et al., 2024a) consists of 2.6K diagram-centric problems (e.g., geometry, functions), each rendered in six visual-text variants to disentangle true visual understanding from linguistic shortcuts. Evaluation is based on step-by-step Chain-of-Thought (CoT) correctness. **MATH-Vision** (Wang et al., 2024) presents ∼3K competition-grade problems across 16 disciplines and five difficulty levels, stressing advanced multimodal reasoning.

[2]For Qwen2.5-VL-72B, we used ms-swift Zhao et al. (2024) to perform Lora finetuning on 80 steps, batch size 8 for resource and time limitation. For 72B finetuning, the Lora adapter is frozen during both rollouts.

[3]The policy model remains frozen during both rollouts.

Table 2: Vision-SR1 vs. baselines. For Vision-R1, as noted in Section 3.1, the original model checkpoint was trained only on math-domain data. So we also reproduce it using our 47K dataset.

Methods	General Visual Understanding				Visual Math & Hallucination			Avg.
	MMMU -Pro	MMMU	RealWorld QA	VisNum Bench	Math Verse	MATH -Vision	Hallusion Bench	
Visionary-R1 (3B) by Xia et al. (2025)	27.4	30.6	56.9	10.0	45.0	40.4	26.7	33.9
Percention-R1 (7B) by Xiao et al. (2025)	36.8	40.9	69.4	15.9	52.1	35.7	65.4	45.2
Vision-R1 (7B) by Huang et al. (2025b)	34.9	42.8	60.1	33.0	57.3	51.2	32.2	44.5
Backbone model: Qwen2.5-VL-3B								
Zero-shot Inference (before RL)	30.5	25.5	65.4	15.7	44.3	40.4	27.1	35.5
Vision-R1 47K data (fair comparison)	40.3	49.5	63.0	36.7	42.8	29.9	67.4	47.1
Vision-SR1 (ours)	40.8	49.6	66.1	41.9	45.8	29.3	68.3	48.8
Backbone model: Qwen2.5-VL-7B								
Zero-shot Inference (before RL)	34.2	33.5	68.5	21.4	49.2	31.9	51.7	41.5
Vision-R1 47K data (fair comparison)	39.8	51.8	66.6	43	53.2	33.8	66.6	50.7
Vision-SR1 (ours)	**40.7**	**52.2**	**69.2**	**43.5**	**54.5**	**36.2**	**68.9**	**52.2**
Backbone model: Mimo-VL-7B								
Zero-shot Inference (before RL)	38.0	45.6	68.2	30.2	35.5	21.6	71.9	44.4
Vision-R1 47K data (fair comparison)	38.7	47.3	67.1	33.5	35.3	25.7	74.3	46.0
Vision-SR1 (ours)	**39.3**	**49.5**	**68.1**	**44.6**	**40.0**	**29.6**	**75.6**	**49.5**
Backbone model: Qwen2.5-VL-72B								
Zero-shot Inference (before Lora RL)	40.6	45.0	69.5	26.1	51.3	33.5	68.7	47.8
Vision-R1 47K data (fair comparison)	43.8	45.3	72.1	47.1	50.5	34.6	73.2	52.4
Vision-SR1 (ours)	**47.6**	**52.8**	**75.1**	**47.9**	**53.6**	**34.5**	**74.4**	**55.1**

Hallucination Diagnosis. To diagnose model failures, we use **HallusionBench** (Guan et al., 2024), a benchmark designed to pinpoint specific errors: (i) language-side hallucination, where visual context is ignored, and (ii) visual-illusion errors, where the image is misinterpreted. The benchmark's binary yes/no format enables precise error analysis.

For our evaluations, we all use Gemini-2.5-flash (Comanici et al., 2025) to judge response correctness on non-multiple choice format question, serving as a proxy for human judgment.

3.3 Experimental Results

3.3.1 Vision-SR1 v.s. Baseline Methods

Table 2 presents a comprehensive comparison of Vision-SR1 with several baseline methods across diverse vision-language benchmarks. For example, with the Qwen2.5VL-72B backbone, Vision-SR1 reaches 47.6 on MMMU-Pro and 52.8 on MMMU, outperforming Vision-R1 fair comparison runs (43.8 and 45.3, respectively). When averaged across all benchmarks, Vision-SR1 establishes a clear margin of improvement. With the 72B backbone, it achieves an average score of 55.1, compared to 52.4 for Vision-R1. Even with the smaller 3B and 7B backbone, Vision-SR1 achieve 48.8 and 52.2 average, outperforming all comparable baselines. For results on Mimi-VL-7B, a model outside the Qwen-VL family, we observe a similar trend: the average accuracy improves from 44.4 to 49.5. This shows that our method generalizes beyond the Qwen-VL. These results demonstrate that Vision-SR1 outperforms prior baseline models across both general-purpose and math-specific visual reasoning tasks, validating the effectiveness of our approach.

Table 3: Our method also can improve VLMs' abilities on spatial reasoning and language shortcut (LS) robustness.

Methods	ViLP (LS)	MMSI -Bench	Omni Spatial	Avg.
Backbone: Mimo-VL-7B				
before RL	56.4	28.2	40.3	41.6
Vision-R1	58.2	27.7	40.4	42.1
Vision-SR1	59.3	28.0	42.7	43.3
Backbone: Qwen2.5-VL-7B				
before RL	45.1	24.0	27.3	32.1
Vision-R1	51.3	21.9	31.1	34.8
Vision-SR1	52.6	27.7	44.2	41.5
Backbone: Qwen2.5-VL-72B				
Zero-shot	55.9	33.4	36.5	41.9
Vision-R1 (Lora)	55.4	35.4	36.4	42.4
Vision-SR1 (Lora)	61.8	35.3	38.6	45.2

3.3.2 Ablation study on Spatial Reasoning and Language Shortcut Datasets

In addition to evaluating Vision-SR1 on standard visual-reasoning benchmarks, we further evaluate its effectiveness on additional datasets to probe two complementary challenges: spatial reasoning and language-shortcut (LS) robustness. MMSI-Bench Yang et al. (2025) and OmniSpatial Jia et al. (2025) target multi-image spatial understanding, requiring models to integrate spatial relationships across multiple images. In contrast, ViLP Luo et al. (2025) evaluates language shortcuts by pairing

each question with images that can be answered either through textual priors alone or only through pure visual reasoning. Table 3 shows that Vision-SR1 generalizes well to spatial reasoning benchmarks and substantially improves robustness to visual–language shortcuts. In particular, explicitly generating visual descriptions helps the model avoid shortcut behavior and rely more on the actual visual content. Next we propose a systematic way to evaluate VLMs' language shortcut frequencies on standard VLM benchmarks.

3.3.3 Analysis on Language Shortcut

Table 4: Language Shortcut Rate (LSR) across different benchmarks. Lower values indicate better performance, as a reduced LSR reflects fewer language shortcuts during reasoning. Adding additional reward supervision can reduce the change of visual reasoning reward hacking.

Methods	General Visual Understanding				Visual Math & Hallucination			Avg.
	MMMU -Pro	MMMU	RealWorld QA	VisNum Bench	Math Verse	MATH -Vision	Hallusion Bench	
Vision-SR1 (3B)	7.5	6.3	10.8	5.4	10.3	8.3	10.1	9.4
⊢ w/o self-reward	9.0	9.6	11.9	4.2	11.4	9.2	8.5	10.4
Vision-SR1 (7B)	8.0	6.5	13.4	4.2	11.5	10.7	6.8	9.8
⊢ w/o self-reward	8.7	5.3	10.8	3.9	12.7	10.7	9.1	10.1

We also introduce the Language Shortcut Rate (LSR), a metric designed to quantify how often a model produces the correct answer with an incorrect visual perception. A high LSR suggests the model is leveraging language knowledge prior rather than genuine visual understanding.

Our evaluation, follows a two-step process and uses Gemini-2.5-flash as a judge: (1) Visual Perception Extraction: for each model output, we extracted the generated visual reasoning, denoted as $\hat{C}$. (2) Self-Containment Check: we then provide the $\hat{C}$ and the original question Q to Gemini-2.5-Flash evaluator. If the evaluator can reproduce the correct ground-truth answer using *only* this information, $\hat{C}$ is deemed self-contained. Based on this process, we define the metrics: The **Language Shortcut Rate (LSR)** is defined as the percentage of instances where the model produces an *incorrect (not self-contained) visual reasoning* but still gives the *correct final answer*:

$$\text{LSR} = \frac{\#\{\text{incorrect visual reasoning \& correct answer}\}}{\#\{\text{total samples}\}}$$

A higher LSR indicates that the model is answering correctly while bypassing visual perception, suggesting reliance on language prior shortcuts. An LSR of 0 indicates no shortcutting, i.e., every correct answer is supported by a correct, self-contained visual reasoning.

We compute the LSR for 7B model w/ and w/o self rewards on seven selected benchmarks for demo example in Table 4. An important finding is that the visual shortcut is the *highest in multimodal mathematical reasoning*, which raises important question to previous work R1-VL (Zhang et al., 2025), VLM-R1 (Shen et al., 2025), Vision-R1 (Huang et al., 2025b): is multimodal RL training truly improving VLMs' abilities to perform visual reasoning, or simply awakes the models' language reasoning ability to guess without actually looking at visual information?

4 Related Work

4.1 Post-Training Vision-Language Models

Recent vision-language models have increasingly leveraged post-training alignment techniques, including instruction tuning and reinforcement learning, to enhance general-purpose multimodal performance (Liu et al., 2023b; Bai et al., 2025; Chen et al., 2024; et al, 2024; Huang et al., 2025b). For example, LLaVA (Liu et al., 2023b) is tuned on GPT-4 generated (image, question, answer) pairs, coupling a CLIP encoder with Vicuna to produce a visual chat assistant that imitates some GPT-4 vision capabilities. InstructBLIP (Dai et al., 2023) introduces an instruction-aware query transformer tuned on 26 datasets, which yields a model that substantially outperforms even larger models on zero-shot benchmarks. Beyond standard instruction-tuning methods like LLaVA and InstructBLIP, recent work increasingly uses reinforcement learning (RL) to align vision-language models

for better reasoning (Huang et al., 2025b; Xia et al., 2025; Xiao et al., 2025). Many of these methods, inspired by techniques from DeepSeek-R1 (DeepSeek-AI et al., 2025), focus on sophisticated reward engineering. Strategies include providing step-wise rewards to supervise the intermediate reasoning (Zhang et al., 2025), adding explicit visual annotations to ground truth for calculating visual rewards (Xiao et al., 2025), and applying RL in a two-stage curriculum that first strengthens text-only reasoning (Peng et al., 2025b). As a complementary approach, RL from AI Feedback for VLMs demonstrates that preference-based alignment is also a powerful signal, showing it can substantially reduce object hallucination by learning from AI-generated feedback (Yu et al., 2024).

4.2 Self-Rewarding Reinforcement Learning

The existing reinforcement learning with verifiable rewards (RLVR) methods heavily rely on high-quality reward models or human feedback, creating a major bottleneck for scalability (Peng et al., 2025a; Dai et al., 2025; Li et al., 2025c; Luu et al., 2025). To overcome this, recent work explores self-rewarding approaches, where the model itself provides intrinsic reward signals during RL post-training, an idea first pioneered by Yuan et al. (2025). Building on self-rewarding language models, methods replace external reward models with the model's own confidence and uncertainty (logit-based self-certainty) or self-verification of its solutions, and even elicit a latent *endogenous* reward already present inside base LLMs (Zhao et al., 2025; Li et al., 2025a; Simonds et al., 2025; Zheng et al., 2025; van Niekerk et al., 2025; Huang et al., 2025a; Zhou et al., 2025). For example, RLIF leverages self-certainty as a reward, achieving comparable performance to GRPO while improving out-of-distribution generalization (Zhao et al., 2025). Similarly, RLSC optimizes a self-confidence reward to secure large accuracy gains with only a few training samples (Li et al., 2025a).

Although self-generated reward signals have thrived in text-only LLMs, only a few works extend this idea to VLMs (Zhou et al., 2024; Lee et al., 2025; Holmes & Chi, 2025), largely due to the complexity of the visual modality and the difficulty of properly defining and evaluating reward signals that capture visual perception. Recent progress includes Calibrated Self-Rewarding, which iteratively generates candidates, self-scores them with step-wise, visually constrained rewards, and fine-tunes via direct preference optimization (DPO) (Zhou et al., 2024). Similarly, RG-VLM uses a VLM to directly label rewards for offline trajectories in long-horizon visual tasks, serving as an auxiliary signal that boosts generalization (Lee et al., 2025). Beyond judgment-based signals, ARES derives dense shaped rewards from attention weights to accelerate learning under sparse or delayed feedback (Holmes & Chi, 2025). These works show that internal visual signals can provide rich reward feedback for VLM alignment without costly supervision, yet the reward is not integrated end-to-end, where the policy receives both visual perception and answer rewards during training.

5 Conclusion and Future Work

In this paper, we introduce Vision-SR1, a self-rewarded reinforcement learning framework that decomposes vision-language understanding into visual reasoning and language reasoning components. Our approach uses the VLM itself to generate explicit rewards for visual understanding, then applies Multi-Reward Policy Optimization to provide clear gradient attribution and backpropagation pathways for each reward component. Vision-SR1 strengthens visual perception and reduces language shortcuts, thereby improving VLM performance across several domains of vision-language tasks. Our proposed metric LSR further shows how perception reward lowers the tendency of models to answer via language shortcut rather than genuine visual reasoning.

This work opens up several future research directions. First, future work can focus on improving the efficiency of the *visual reasoning then think* generation format by treating the visual reasoning component as latent thinking, thereby reducing the number of decoded tokens while still enabling reward attribution to latent visual processes during the RL phase. It is also important to recognize that some of the observed mathematical gains from RL training in VLMs may come from spurious effects – for instance, recalibrating the LLM backbone's output distribution can boost multimodal math performance without true visual grounding (Shao et al., 2025). This suggests that improvements in accuracy may sometimes reflect better exploitation of language shortcuts rather than genuine perception gains. Therefore, future work can also explore more analysis to disentangle visual grounding from shortcut learning.

Reproducibility Statement

To ensure the reproducibility of our research, we provide information regarding our prompt templates and experimental setup in the main paper and Appendix. All datasets and code will be released upon conference decision release.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Maurits Bleeker, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. Demonstrating and reducing shortcuts in vision-language representation learning, 2024. URL https://arxiv.org/abs/2402.17510.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Gheorghe Comanici, Eric Bieber, and et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Runpeng Dai, Tong Zheng, Run Yang, and Hongtu Zhu. R1-re: Cross-domain relationship extraction with rlvr. *arXiv preprint arXiv:2507.04642*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu,

Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

OpenAI et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf, 2025. URL https://arxiv.org/abs/2502.18770.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

Ian Holmes and Min Chi. Attention-based reward shaping for sparse and delayed rewards, 2025. URL https://arxiv.org/abs/2505.10802.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. 2025a. URL https://arxiv.org/abs/2508.05004.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025b. URL https://arxiv.org/abs/2503.06749.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models, 2025. URL https://arxiv.org/abs/2506.03135.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training, 2024. URL https://arxiv.org/abs/2404.14604.

Younghwan Lee, Tung M. Luu, Donghoon Lee, and Chang D. Yoo. Reward generation via large vision-language model in offline reinforcement learning, 2025. URL https://arxiv.org/abs/2504.08772.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. Confidence is all you need: Few-shot rl fine-tuning of language models, 2025a. URL https://arxiv.org/abs/2506.06395.

Zhimin Li, Haichao Miao, Xinyuan Yan, Valerio Pascucci, Matthew Berger, and Shusen Liu. See or recall: A sanity check for the role of vision in solving visualization question answer tasks with multimodal llms, 2025b. URL https://arxiv.org/abs/2504.09809.

Zongxia Li, Yapei Chang, Yuhang Zhou, Xiyang Wu, Zichao Liang, Yoo Yeon Sung, and Jordan Lee Boyd-Graber. Semantically-aware rewards for open-ended r1 training in free-form generation, 2025c. URL https://arxiv.org/abs/2506.15068.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pp. 1587–1606, June 2025d.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding, 2025e. URL https://arxiv.org/abs/2505.01481.

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

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. URL https://arxiv.org/abs/2304.08485.

Yangxiao Lu, Ruosen Li, Liqiang Jing, Jikai Wang, Xinya Du, Yunhui Guo, Nicholas Ruozzi, and Yu Xiang. Multimodal reference visual grounding. *arXiv preprint arXiv:2504.02876*, 2025.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/luo25b.html.

Tung Minh Luu, Younghwan Lee, Donghoon Lee, Sunho Kim, Min Jun Kim, and Chang D. Yoo. Enhancing rating-based reinforcement learning to effectively leverage feedback from large vision-language models, 2025. URL https://arxiv.org/abs/2506.12822.

Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, Weihan Cao, Jiangning Liu, Hongwei Liu, Junnan Liu, Songyang Zhang, Dahua Lin, and Kai Chen. Exploring the limit of outcome reward for learning mathematical reasoning, 2025. URL https://arxiv.org/abs/2502.06781.

Georgios Pantazopoulos and Eda B. Özyiğit. Towards understanding visual grounding in visual language models, 2025. URL https://arxiv.org/abs/2509.10345.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems, 2025a. URL https://arxiv.org/abs/2502.19328.

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

Neale Ratzlaff, Man Luo, Xin Su, Vasudev Lal, and Phillip Howard. Training-free mitigation of language reasoning degradation after multimodal instruction tuning. In *Proceedings of the AAAI Symposium Series*, volume 5, pp. 384–388, 2025.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr, 2025. URL https://arxiv.org/abs/2506.10947.

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

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

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3698–3712, 2022.

Toby Simonds, Kevin Lopez, Akira Yoshiyama, and Dominique Garmier. Rlsr: Reinforcement learning from self reward, 2025. URL https://arxiv.org/abs/2505.08827.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

Carel van Niekerk, Renato Vukovic, Benjamin Matthias Ruppik, Hsien chin Lin, and Milica Gašić. Post-training large language models via reinforcement learning from self-feedback, 2025. URL https://arxiv.org/abs/2507.21931.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. URL https://arxiv.org/abs/2402.14804.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. Visnumbench: Evaluating number sense of multimodal large language models, 2025. URL https://arxiv.org/abs/2503.14939.

xAI. Realworldqa: Real-world spatial understanding benchmark. https://x.ai/blog/grok-1.5v-and-realworldqa, 2024. CC BY-ND 4.0 license. Benchmark dataset released with Grok-1.5 Vision.

Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14677.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Advancing multimodal reasoning capabilities of multimodal large language models via visual perception reward, 2025. URL https://arxiv.org/abs/2506.07218.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence, 2025. URL https://arxiv.org/abs/2505.23764.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination?, 2025. URL https://arxiv.org/abs/2505.23646.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness, 2024. URL https://arxiv.org/abs/2405.17220.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2025. URL https://arxiv.org/abs/2401.10020.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

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, 2025. URL https://arxiv.org/abs/2409.02813.

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

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?, 2024a. URL https://arxiv.org/abs/2403.14624.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Wings: Learning multimodal llms without text-only forgetting. *Advances in Neural Information Processing Systems*, 37:31828–31853, 2024b.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards, 2025. URL https://arxiv.org/abs/2505.19590.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Tong Zheng, Lichang Chen, Simeng Han, R Thomas McCoy, and Heng Huang. Learning to reason via mixture-of-thought for logical reasoning. *arXiv preprint arXiv:2505.15817*, 2025.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models, 2024. URL https://arxiv.org/abs/2405.14622.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*, 2025.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, Qizheng Zhang, Lin Chen, Fanghao Shao, Bo Xue, Yunchong Song, Zhenjie Yang, Ganqu Cui, Ning Ding, Jianfeng Gao, Xiaodong Liu, Bowen Zhou, Hongyuan Mei, and Zhouhan Lin. Flowrl: Matching reward distributions for llm reasoning, 2025. URL https://arxiv.org/abs/2509.15207.

A Appendix

A.1 The Use of Large Language Models (LLMs)

We acknowledge the use of large language models (LLMs) as assistive tools in this research. Our use of LLMs was limited to refine grammar and improve language clarity. All outputs from these models were meticulously reviewed, revised, and verified by the authors, who retain full responsibility for all content presented in this paper.

B Experiment Details

B.1 Prompt Templates

This section presents the prompt templates used for constructing the cold start training data and Model Training prompt. The ***See-Think*** prompt is used for generating SFT ***See-Think*** data and model training. The Caption-Reasoner prompt is used to generate text-only caption reasoner SFT data and self-reward during training.

See-Think Prompt Template

{*Question*}
You are tasked with analyzing an image/video to generate a detailed description to help you answer the question. First analyze the image/video and produce a self-contained description—detailed enough that can lead to the correct answer. Wrap the entire description in $< description >< /description >$ tags.

Next, engage in an internal dialogue and include self-reflection or verification in your reasoning process. Provide your detailed, step-by-step reasoning based on the image/video description information and image/video, and enclose this part within $< think >< /think >$ tags.

Finally, provide a single word or phrase answer to the question in \boxed{}.

The output format should be: $< description >$ image/video description here $< /description >$ $< think >$reasoning process here $< /think >$ \boxed{FINAL ANSWER here}.

Note: {`Question`} *is a placeholder for the actual question.*

Caption-Reasoner (Self-Reward) Prompt Template

Text description: {Description}

Question: {Question}

You are provided a text description of a problem and a question. Determine the answer to the question based on the text description. First provide an internal step-by-step reasoning within $< think >< /think >$ tags, then provide a single word or phrase answer in \boxed{}.

Note: {`Description`} *is a placeholder for the actual text caption.* {`Question`} *is a placeholder for the actual question.*

Table 5: Through self-reward, the model is implicitly rewarded for text-only reasoning, leading to improved performance in general reasoning and reduced degradation in math reasoning benchmarks.

Model	MMLU-Pro	SuperGPQA	GSM8K	MATH-500
Backbone model: Qwen2.5-VL-3B				
Before RL	34.3	15.1	78.5	65.2
Vision-R1	47.7	23.1	82.2	66.0
Vision-SR1	48.1	23.2	83.2	68.6
Backbone model: Qwen2.5-VL-7B				
Before RL	33.4	17.1	86.0	73.4
Vision-R1	53.4	26.7	85.5	68.2
Vision-SR1	56.1	26.3	87.6	70.8

Vision Reasoner (CoT) Prompt Template

Question: {Question}

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{}.

Note: {`Question`} *is a placeholder for the actual question.*

B.2 LLM-AS-A-JUDGE PROMPT

We use Gemini-2.5-flash as our LLM-as-a-Judge to evaluate

LLM-as-a-Judge Prompt Template

- **Model**: Gemini-2.5-flash

Prompt Message:
Question: {Question}

Reference: {Reference}

Candidate: {Candidate}

You are provided a question, a gold answer, and a candidate answer. Your task is to judge the correctness of the candidate answer. Return your judgment enclosed with $< judgment >< /judgment >$.

Note: {`Question`} *is a placeholder for the actual question;* {`Reference`} *is a placeholder for the gold answer;* {`Candidate`} *is a placeholder for the model response.*

B.2.1 ANALYSIS ON TEXT-ONLY REASONING

An interesting question is how different training strategies affect the text-only reasoning capabilities of VLMs. In particular, by decoupling visual perception and language reasoning with two separate rewards, we ask whether these abilities can mutually reinforce one another. To examine this, we evaluated the text-only performance of VLMs after RL fine-tuning on multimodal data.

Specifically, we tested on four text-only datasets: MMLU-Pro and SuperGPQA (multi-disciplinary, general-domain benchmarks), and MATH-500 and GSM8K (mathematical reasoning tasks). Our results (Table 5) compare Vision-R1, our method, and pre-RL training baselines.

First, we observe that on GSM8K and MATH-500, multimodal RL training, including both Vision-R1 and our method, degrades text-only reasoning performance. This observation aligns with recent findings on "text-only forgetting" in VLMs Zhang et al. (2024b); Ratzlaff et al. (2025), which show

Table 6: Results of ablation study: Vision-SR1 v.s. Vision-SR1 w/o visual perception self-reward.

Methods	General Visual Understanding				Visual Math & Hallucination			Avg.
	MMMU -Pro	MMMU	RealWorld QA	VisNum Bench	Math Verse	MATH -Vision	Hallusion Bench	
Vision-SR1 (3B)	40.8	49.6	66.1	41.9	45.8	29.3	68.3	48.8
⊢ w/o self-reward	40.0	48.0	62.6	41.6	45.1	30.2	65.8	47.6
Vision-SR1 (7B)	**40.7**	**52.2**	**69.2**	**43.5**	**54.5**	**36.2**	**68.9**	**52.2**
⊢ w/o self-reward	42.8	51.8	67.3	35.7	52.6	34.4	67.8	50.3

that visual instruction tuning can impair language reasoning (particularly in mathematics) depending on the underlying LLM. Second, compared to Vision-R1, our method proved more effective at mitigating performance degradation on text-only mathematical benchmarks (MATH-500, GMS8K) and yielded larger gains on general knowledge tasks (MMLU-Pro, SuperGPQA). This indicates that separating the optimization signals for visual perception and language reasoning helps preserve text-only competencies, while still enabling improvements from multimodal training.

B.2.2 Ablation Study on Self-Reward

We train a control version of our model without the visual reasoning self-reward and Multi-Reward Policy Optimization (Vision-SR1 w/o self-reward). This ablated model still follows a structured output (visual perception, CoT reasoning, and answer) but is optimized only with answer and format rewards. The self-visual reward for self-evaluating visual reasoning and Multi-Reward Policy Optimization are removed. We note the only difference between Vision-SR1 w/o self-reward and Vision-R1 (Huang et al., 2025b) lies in the output structure, i.e., using different system prompts, while all supervision signals (answer reward and format rewards) remain the same. Interestingly, our system prompt yields slightly better performance (+1.0 on average). Table 6 reports the ablation results. We find that not including visual reasoning reward and Multi-Reward Policy Optimization could lead to overall worse VLM task performance compared to including them in the training process.