Self-Rewarded Multimodal Coherent Reasoning Across Diverse Visual Domains

Jusheng Zhang[1] **Ningyuan Liu**[1] **Kaitong Cai**[1] **Sidi Liu**[2]
Ziliang Chen[3] **Wenhao Wang**[4*] **Keze Wang**[1*]

[1]Sun Yat-sen University
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]Southern University of Science and Technology
[4]Vast Intelligence Lab

Abstract

Multimodal LLMs often produce fluent yet unreliable reasoning, exhibiting weak step-to-step coherence and insufficient visual grounding, largely because existing alignment approaches supervise only the final answer while ignoring the reliability of the intermediate reasoning process. We introduce SR-MCR, a lightweight and label-free framework that aligns reasoning by exploiting intrinsic process signals derived directly from model outputs. Five self-referential cues – semantic alignment, lexical fidelity, non-redundancy, visual grounding, and step consistency – are integrated into a normalized, reliability-weighted reward that provides fine-grained process-level guidance. A critic-free GRPO objective, enhanced with a confidence-aware cooling mechanism, further stabilizes training and suppresses trivial or overly confident generations. Built on Qwen2.5-VL, SR-MCR improves both answer accuracy and reasoning coherence across a broad set of visual benchmarks; among open-source models of comparable size, SR-MCR-7B achieves state-of-the-art performance with an average accuracy of 81.4%. Ablation studies confirm the independent contributions of each reward term and the cooling module.

1 Introduction

Multimodal reasoning must deliver not only correct answers but also explanations that are coherent and visually grounded. Yet recent MLLMs (e.g., LLaVA [27], Qwen-VL [2]) often drift in their intermediate steps—contradicting themselves, hallucinating evidence [24, 66, 38], or repeating trivial content. Existing alignment pipelines, including instruction tuning [45] and preference finetuning (DPO [33], RLHF [31, 39]), rely on costly human-labeled rewards or external evaluators [64]. While effective, they remain brittle under domain shift and largely *outcome-centric*, supervising answers rather than the reasoning process [25, 5].

This outcome-centric focus leaves two issues unresolved: (i) *Lack of intrinsic process reward*: supervising only final answers leaves step coherence and visual grounding underconstrained; (ii) *Cross-domain instability*: single reward proxies (e.g., lexical overlap) miscalibrate signals across heterogeneous tasks, causing over-alignment to spurious patterns.

At the same time, a single forward pass of an MLLM already exposes multiple measurable signals that correlate with "good" reasoning—semantic alignment, lexical fidelity, non-redundancy, grounding to visual evidence, and local step consistency [55]. These signals are inexpensive to compute, task-agnostic, and complementary in nature. Rather than relying solely on external preference labels, we ask: *can we turn these process-level signals into a practical, intrinsic self-reward for multimodal*

*Co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

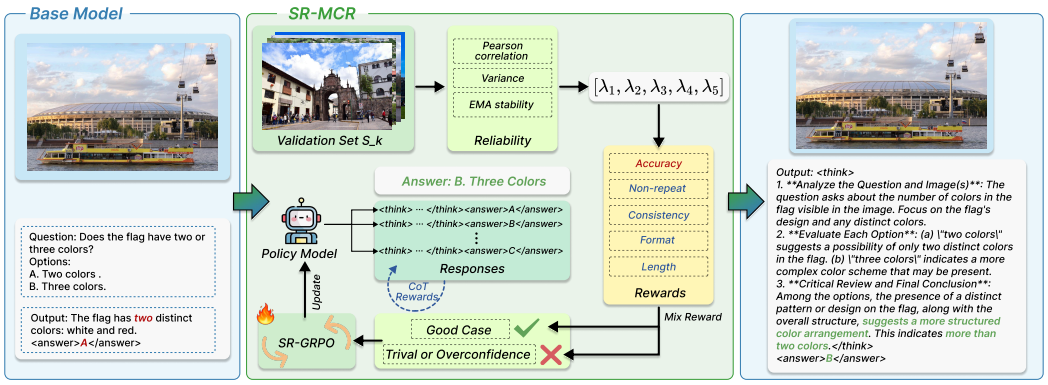


Figure 1: Overview of **SR-MCR**. Given an image–text input (I, x), the policy π_θ generates multiple responses, each scored by five process-level self-reward terms and an adaptive reliability estimator to form a mixed reward. **SR-GRPO** then favors high-reward, reliable outputs while suppressing trivial ones, updating π_θ on top of the frozen base policy π_0.

reasoning? We posit that, once normalized and reliability-weighted, such signals can serve as a unified *self-reward* for both training and diagnosis [52].

Based on this observation, we develop **Self-Rewarded Multimodal Coherent Reasoning (SR-MCR)**, a unified framework that instantiates process-aware, label-free alignment for MLLMs. Given an image I, textual input x, and model outputs $(\hat{y}_a, \hat{y}_t)$ (final answer and reasoning trace), SR-MCR computes a normalized self-reward $\mathcal{R}(I, x, \hat{y}_a, \hat{y}_t) = \sum_{k\in\{\text{sem,lex,nr,vis,step}\}} \lambda_k \, \tilde{s}_k, \qquad \lambda_k = \frac{\exp(\alpha \, \text{Relia}_k)}{\sum_j \exp(\alpha \, \text{Relia}_j)}$.Here $\tilde{s}_k \in [0, 1]$ are min–max normalized scores for semantic similarity, lexical overlap, non-redundancy, visual grounding, and step-wise coherence. The term Relia_k estimates each signal's reliability (e.g., held-out correlation or inverse variance), producing adaptive weights λ_k that attenuate noisy or unstable signals and amplify consistent ones.

Optimization. Instead of constructing human preference pairs, we directly treat this self-reward as the optimization target and fine-tune the model with a *self-rewarded GRPO* [37] objective. The policy is updated to increase the likelihood of high-reward generations while constraining drift from the base model via a KL term [36, 35] (Sec. 2, Eq. 9). To further stabilize training under noisy self-rewards, we introduce a dynamic *cooling weight* (Eq. 8) that down-weights trivial or overconfident samples based on their normalized negative log-likelihood [47, 23]. Together, these components yield a simple, label-free, and process-aware alignment recipe that can be applied across diverse visual domains. **Contributions. SR-MCR** is a lightweight GRPO-based framework that learns a reliability-weighted reward from five self-signals for coherent and visually grounded multimodal reasoning. **Process-aware optimization** with a simple cooling scheme stabilizes reward-driven training and improves reasoning consistency. **Empirical gains** show consistent improvements over Qwen2.5-VL baselines, with ablations validating the contribution of each component.

2 Method

Following the motivation in Sec. 1, we introduce **Self-Rewarded Multimodal Coherent Reasoning (SR-MCR)**, a training framework that enables multimodal large language models (MLLMs) to *self-evaluate* and *self-improve* their reasoning process without human preference labels [1, 6]. SR-MCR is built on three components: (i) a unified self-reward that aggregates multiple process-level signals, (ii) an adaptive reliability weighting scheme, and (iii) a critic-free GRPO objective with cooling regularization. Together, these components turn multimodal alignment from discrete preference matching into continuous, self-guided optimization over process quality.

2.1 Problem Setup

Given an image I, text input x, a frozen base policy π_0, and a trainable policy π_θ, the model produces a final answer $\hat{y}_a$ and a reasoning trace $\hat{y}_t$ [46, 18], forming the full response $\hat{y} = (\hat{y}_a, \hat{y}_t)$. Our goal

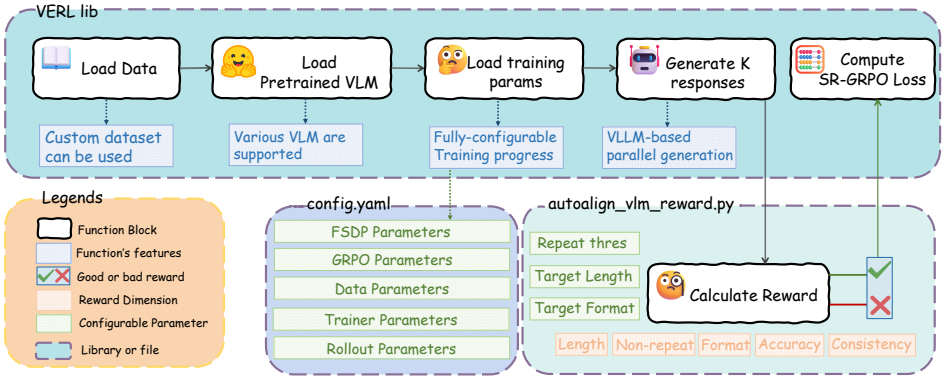


Figure 2: Training pipeline of SR-MCR. We load data and a pretrained VLM, sample K responses via vLLM [20], compute component-wise self-rewards, and train a LoRA adapter with the SR-GRPO loss. Rewards and hyperparameters are specified in a single YAML configuration.

is to align π_θ such that $\hat{y} \sim \pi_\theta(\cdot \mid I, x)$ is (i) semantically correct, (ii) visually grounded, and (iii) step-wise coherent—*without* human preferences or ground-truth labels. Training relies solely on intrinsic, multi-signal self-rewards derived from $(I, x, \hat{y})$.

2.2 Unified Self-Reward

Each forward pass of an MLLM yields several low-cost signals that correlate with reasoning quality. We combine five such signals into a unified, normalized self-reward:

$$R(I, x, \hat{y}_a, \hat{y}_t) = \sum_{k \in \{\text{sem,lex,nr,vis,step}\}} \lambda_k \, \tilde{s}_k \;\; \in [0, 1], \tag{1}$$

where each raw score s_k is first winsorized (clipped at the 1st and 99th percentiles) and then min–max normalized to $\tilde{s}_k \in [0, 1]$. The five components are: **Semantic consistency** ($\tilde{s}_{\text{sem}}$): sentence-level semantic alignment between reasoning and answer; **Lexical fidelity** ($\tilde{s}_{\text{lex}}$): surface-level overlap between reasoning and answer; **Non-redundancy** ($\tilde{s}_{\text{nr}}$): penalty on repeated n-grams in the reasoning trace; **Visual grounding** ($\tilde{s}_{\text{vis}}$): alignment between textual mentions and visual regions; **Step-wise coherence** ($\tilde{s}_{\text{step}}$): local logical consistency between adjacent steps.

All signals are computed purely from the model's own outputs and pretrained, frozen tools (e.g., Sentence-BERT [34, 63], CLIP, NER, detection, NLI), and thus do not require ground-truth labels in the training loop. Crucially, although these signals are intrinsic, they correlate strongly with human preferences. Empirical analysis confirms strong alignment with human judgments, yielding Pearson correlation coefficients of $r = 0.65$ for visual grounding and an average of $r = 0.51$ across all components.

Semantic and Lexical Signals. We first measure how well the reasoning trace supports the final answer. Let the reasoning trace be split into N sentences $\{\hat{y}_t^{(i)}\}_{i=1}^{N}$. We define

$$\tilde{s}_{\text{sem}} = \frac{1}{N} \sum_{i=1}^{N} \cos\Big(f_{\text{SBERT}}(\hat{y}_t^{(i)}), f_{\text{SBERT}}(\hat{y}_a)\Big), \tag{2}$$

where $f_{\text{SBERT}}(\cdot)$ is a Sentence-BERT encoder and $\cos(\cdot, \cdot)$ is cosine similarity. Lexical consistency is quantified via ROUGE-L [26]: $\tilde{s}_{\text{lex}} = \text{ROUGE-L}(\hat{y}_t, \hat{y}_a)$. These signals can optionally be compared against ground-truth answers on a held-out validation set for calibration, but ground truth is *not* required in the training loop.

Non-Redundancy. To discourage degenerate repetition [14], we compute

$$\tilde{s}_{\text{nr}} = 1 - \frac{|\text{repeated } n\text{-grams}(\hat{y}_t)|}{|\text{all } n\text{-grams}(\hat{y}_t)|}, \tag{3}$$

where we use $n = 2$ in our experiments. Higher values correspond to more informative and concise reasoning traces.

Algorithm 1 SR-MCR: Self-Rewarded GRPO Training Loop

Input: base policy π_0, trainable policy π_θ
Hyperparameters: learning rate η, KL weight β_{KL}
for each training iteration **do**
  Candidate sampling: sample K responses $\mathcal{C} \leftarrow \{\hat{y}^{(i)} \sim \pi_\theta(\cdot \mid I, x)\}_{i=1}^{K}$
  Reward computation: compute $\{s_k\}$ and fuse into R via Eq. (1)
  Reliability update: update Relia_k (e.g., EMA or variance), then λ_k via Eq. (6)
  Policy update: compute $R_{\text{adj}}(\hat{y})$ via Eq. (8) and $\mathcal{L}_{\text{SR-GRPO}}$ via Eq. (9)
    update parameters $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{SR-GRPO}}$
  Monitoring: track component rewards and coherence violations; early-stop when a moving average of R plateaus
end for
Return: optimized policy π_θ

Visual Grounding. We align textual mentions in $\hat{y}_t$ with visual regions in I. Textual mentions $\{m_j\}_{j=1}^{M}$ are extracted using a pretrained NER model (e.g., spaCy). Visual regions $\{r_i\}$ are obtained from a pretrained detector (e.g., DETR [3]) as bounding boxes. We then compute CLIP embeddings and define

$$\tilde{s}_{\text{vis}} = \frac{1}{M} \sum_{j=1}^{M} \max_i \cos\Big(f_{\text{CLIP}}(r_i), f_{\text{CLIP}}(m_j) \Big), \tag{4}$$

where $f_{\text{CLIP}}(\cdot)$ denotes CLIP encoders for image crops and text spans. AlignScore [54] can be used as a drop-in alternative, but we adopt CLIPScore [13] for efficiency.

Step-wise Coherence. Finally, we assess local logical consistency within the reasoning trace. Let the trace be decomposed into steps $\{\text{step}_1, \ldots, \text{step}_L\}$. A pretrained NLI model (e.g., DeBERTa-v3-large [12]) provides entailment probabilities e_i and contradiction probabilities c_i between adjacent steps $(\text{step}_i, \text{step}_{i+1})$. We define

$$\tilde{s}_{\text{step}} = \frac{1}{L-1} \sum_{i=1}^{L-1} \min\big(e_i,\, 1 - c_i\big), \tag{5}$$

which encourages both strong entailment and low contradiction between consecutive steps.

Adaptive Reliability Weighting. Different signals may have varying reliability across domains (e.g., lexical overlap is more informative in QA than in math reasoning). We therefore assign adaptive weights λ_k based on a reliability score Relia_k:

$$\lambda_k = \frac{\exp\big(\alpha\, \mathrm{Relia}_k\big)}{\sum_j \exp\big(\alpha\, \mathrm{Relia}_j\big)}, \quad \sum_k \lambda_k = 1, \tag{6}$$

where $\alpha > 0$ is a temperature that controls sharpness. The reliability score Relia_k can be instantiated as

$$\mathrm{Relia}_k \in \{\rho_k,\ 1/\mathrm{Var}[s_k],\ \text{EMA-stability}(s_k)\}\,. \tag{7}$$

Here ρ_k denotes the Pearson correlation between the raw score s_k and task accuracy on a small, labeled validation set (optional). This calibration stage is the *only* place that may use ground-truth labels; the main optimization in Sec. 2.3 is fully self-rewarded. If no labeled validation set exists, we use GT-free proxies—inverse variance, EMA stability, or uniform weights $\lambda_k = 1/5$.

2.3 Self-Rewarded GRPO with Cooling

To optimize continuous self-rewards without constructing preference pairs, we adopt the **Generalized Reversed Policy Optimization (GRPO)** objective. Directly using the normalized self-reward R in Eq. (1) can be unstable, as low-reward trivial samples and overconfident degenerate ones introduce noisy gradients. We therefore define an *adjusted reward* that applies reward- and confidence-aware cooling:

$$R_{\text{adj}}(\hat{y}) = \tilde{R}(\hat{y})^{\gamma} \cdot \sigma\big(\kappa\,(\bar{\ell}_\theta(\hat{y}) - c)\big), \tag{8}$$

where $\tilde{R}(\hat{y}) \in [0,1]$ is the (optionally re-normalized) self-reward, $\bar{\ell}_\theta(\hat{y})$ is the average token negative log-likelihood (NLL) under π_θ, $\sigma(\cdot)$ is the sigmoid function, and γ, κ, c are hyperparameters

Table 1: **Multimodal benchmark results.** SR-MCR (3B/7B) improves its Qwen2.5-VL bases and is competitive with prior models. Best open-source results are shown in **bold**. * denotes models with a separate "Thinking" mode.

Model	MMMU val	MMB v1.1	MME	ChartQA test	AI2D	HallBench	Avg.
Closed-source models							
Gemini-2.0-Pro	72.6	83.0	86.1	91.2	84.8	49.8	77.9
GPT-4o-latest	70.7	84.3	84.2	91.5	86.3	57.0	79.0
Open-source baselines							
InternVL3-2B	47.1	84.3	77.4	80.4	78.7	41.4	68.2
Qwen2.5-VL-3B	48.1	82.4	77.5	87.0	80.7	48.3	70.7
WeThink-7B	50.9	87.8	82.9	90.8	84.5	55.1	75.3
InternVL3-8B	57.3	87.7	85.2	89.6	85.2	53.7	76.5
Qwen2.5-VL-7B	50.3	86.7	82.2	89.5	84.0	56.0	74.8
VL-Rethinker-7B	54.8	88.2	82.9	91.5	83.6	55.1	76.0
VLAA-Thinker-7B	51.9	86.9	83.3	89.5	78.9	51.5	73.7
Keye-VL-8B-Thinking*	63.4	81.7	83.5	88.0	86.4	**62.7**	77.6
Kimi-VL-A3B-Thinking*	60.4	89.7	87.0	92.1	83.1	58.3	78.4
SAIL-VL2-8B-Thinking*	66.1	90.4	86.0	93.6	87.4	61.5	80.8
Our models							
SR-MCR-3B	52.8	86.9	80.8	90.9	84.3	51.9	74.6
SR-MCR-7B	**67.6**	**91.2**	**87.3**	**94.5**	**88.2**	59.4	**81.4**

controlling reward scaling, cooling sharpness, and the NLL baseline (e.g., $\gamma = 1.0$, $\kappa = 5.0$, $c = 0.1$).

This design downweights gradients from two types of uninformative samples: (i) *trivial low-reward* samples, where $\tilde{R}(\hat{y}) \to 0$ drives $R_{\text{adj}}(\hat{y}) \to 0$, and (ii) *overconfident* samples, where $\bar{\ell}_\theta(\hat{y}) < c$ makes the sigmoid term close to zero. By contrast, high-reward yet moderately uncertain samples obtain the strongest gradient signal.

We finally optimize the model with the SR-GRPO loss. Unlike standard GRPO [37] which relies on ratio clipping for stability, our confidence-aware cooling (Eq. 8) acts as a sample-wise soft constraint. Combined with the explicit KL penalty, this effectively filters destabilizing gradients from overconfident errors without requiring artificial truncation, allowing more efficient learning from high-reward exploration.

$$\begin{aligned}\mathcal{L}_{\text{SR-GRPO}} = &- \mathbb{E}_{(I,x),\, \hat{y}\sim\pi_\theta}\Big[R_{\text{adj}}(\hat{y})\, \log \frac{\pi_\theta(\hat{y} \mid I, x)}{\pi_0(\hat{y} \mid I, x)}\Big] \\ &+ \beta_{\text{KL}}\, \text{KL}\big(\pi_\theta \,\|\, \pi_0\big).\end{aligned} \tag{9}$$

Interpretation and Pipeline. The first term steers π_θ toward responses it deems coherent and grounded, while the KL term limits drift from π_0. Unlike pairwise objectives (e.g., DPO, RLHF), SR-GRPO optimizes continuous self-rewards, yielding smoother and more stable label-free training. We train a LoRA adapter [15] on the frozen backbone and re-rank M sampled candidates using R at inference. Without reference answers, SR-MCR falls back to $(\text{vis}, \text{nr}, \text{step})$, which still enforces visual grounding and step coherence.

3 Experiments

3.1 Experimental Setup

Backbone and Training. We adopt Qwen2.5-VL [43] as the base multimodal backbone. We then apply our proposed **Self-Rewarded Multimodal Coherent Reasoning (SR-MCR)** training procedure, which leverages GRPO with step-wise reasoning coherence rewards. All training is performed efficiently using LoRA on top of the original checkpoints. Crucially, our method does

Table 2: **Evaluation on V* Bench.** We compare SR-MCR with Qwen2.5-VL and other VLMs. Best open-source results are shown in **bold**. * denotes models with a separate "Thinking" mode.

Model	Attribute	Spatial	Overall
Closed-source models			
GPT-4o	62.03	72.00	66.00
Open-source models			
InternVL3-2B [67]	59.13	63.16	60.73
Qwen2.5-VL-3B	78.26	60.53	71.20
VLAA-Thinker-7B [4]	56.52	59.21	57.59
WeThink-7B [50]	82.61	75.00	79.58
VL-Rethinker-7B [42]	64.35	71.05	67.02
InternVL3-8B	70.43	72.37	71.20
Keye-VL-8B-Thinking* [41]	66.96	75.00	70.16
Qwen2.5-VL-7B	80.87	75.00	78.53
Kimi-VL-A3B-Thinking* [40]	51.30	60.53	54.97
SAIL-VL2-8B-Thinking* [51]	51.30	65.79	57.07
Our models			
SR-MCR-3B (Ours)	81.16	61.60	73.12
SR-MCR-7B (Ours)	**83.48**	**76.32**	**80.63**

not require training any external reward model and produces a **single, unified model**, obviating the need for separate "Instruct" vs. "Thinking" variants common in prior work. Crucially, reward computation is strictly a training-time operation; SR-MCR incurs **zero additional latency or memory overhead** during inference. **Benchmarks.** We evaluate our models on a comprehensive suite of benchmarks. For general multimodal understanding (evaluated using **ACC**, results in Table 1), we use: **MMMU**$_{\text{val}}$ [53] — general-domain knowledge and reasoning across diverse subjects; **MMBench**$_{\text{v1.1}}$ [28] — instruction-following, world knowledge, and general perception; **MME** [10] — fine-grained vision–language alignment and grounding; **ChartQA**$_{\text{test}}$ [29] — table and chart understanding with numerical reasoning; **AI2D** [17] — diagram parsing and diagrammatic reasoning; **HallBench** [11] — robustness against visual and factual hallucination. Additionally, we specifically evaluate comprehensive visual reasoning using **V*Bench** [49] (results in Table 2).

3.2 Main Results on General Benchmarks

Key Finding. As presented in Table 1, **SR-MCR-7B** establishes a new state-of-the-art among open-source Vision-Language Models (VLMs) with an impressive average score of **81.4**. Crucially, SR-MCR surpasses specialized "Thinking" variants (such as SAIL-VL2-8B-Thinking and Kimi-VL-A3B-Thinking) without requiring the architectural overhead of a dual-checkpoint (Instruct/Thinking) paradigm. By optimizing intrinsic, process-level rewards, our unified framework inherently generates logical and coherent reasoning traces within a single forward policy. It not only leads in the overall average but also dominates demanding individual benchmarks requiring deep domain knowledge, achieving 67.6 on MMMU and 94.5 on ChartQA.

Scalability and Efficacy. The performance improvements introduced by SR-MCR are highly scalable across different parameter sizes. It delivers substantial and consistent gains over its strong Qwen2.5-VL bases at both the 3B scale (**74.6** vs. 70.7, a +3.9 absolute improvement) and the 7B scale (**81.4** vs. 74.8, a +6.6 absolute improvement). This non-linear growth indicates that our *reasoning-coherence alignment* effectively unlocks the latent reasoning capacities of larger base models rather than merely relying on raw parameter scaling. Furthermore, the robust SR-MCR-3B model outpaces several traditional 7B and 8B baselines (e.g., VLAA-Thinker-7B at 73.7), validating the parameter efficiency of our method.

3.3 Results on V* Bench

Fine-Grained Visual Understanding. The V* Bench is specifically designed to rigorously test a model's capacity for complex spatial and attribute reasoning—areas where conventional VLMs often suffer from hallucination and contextual drift. As detailed in Table 2, our SR-MCR models demonstrate exceptional robustness in these challenging scenarios. The **SR-MCR-7B** model achieves the state-of-the-art **Overall** score of **80.63** among all open-source models, outperforming strong baselines like WeThink-7B (79.58) and greatly exceeding the closed-source GPT-4o (66.00).

Attribute and Spatial Reasoning Capabilities. A detailed breakdown reveals that SR-MCR-7B secures the highest open-source scores in both the **Attribute** (**83.48**) and **Spatial** (**76.32**) sub-benchmarks. This superior performance directly reflects the benefits of our continuous self-reward mechanism. By explicitly incorporating visual grounding and step-wise coherence into the optimization objective, SR-MCR strictly enforces alignment between textual mentions and actual visual regions. Consequently, the model successfully avoids intermediate hallucinatory drift, enabling it to track intricate visual elements and spatial relationships across multi-step analysis without losing fidelity. Even at the 3B scale, SR-MCR yields an overall score of 73.12, further validating that our training procedure effectively instills fine-grained perception abilities that transfer across model sizes.

Table 3: **Math reasoning benchmarks.**

Model	MathVista	MathVision
Reasoning Baselines		
VLAA-Thinker-7B [4]	68.0	26.4
WeThink-7B [50]	71.6	26.7
VL-Rethinker-7B [42]	74.9	32.3
Base vs. Ours		
Qwen2.5-VL-7B	68.2	25.07
+ SR-MCR-7B	**72.0** (+3.8)	**31.17** (+6.1)
Qwen3-VL-8B	74.8	53.90
+ SR-MCR-8B	**78.3** (+3.5)	**60.10** (+6.2)

We further evaluate SR-MCR on mathematical reasoning benchmarks, including MathVista (testmini) and MathVision. As shown in Table 3, SR-MCR consistently improves both Qwen2.5-VL-7B and Qwen3-VL-8B backbones. Specifically, SR-MCR-7B improves Qwen2.5-VL-7B from 68.2/25.07 to **72.0/31.17**, while SR-MCR-8B improves Qwen3-VL-8B from 74.8/53.90 to **78.3/60.10**. These gains demonstrate that SR-MCR effectively enhances mathematical reasoning across different model families.

3.4 Generalization to Mathematical Reasoning

Beyond Perception-Heavy Tasks. To verify that our process-level rewards generalize beyond perception-heavy tasks, we evaluate SR-MCR on MathVista and MathVision across two generations of Qwen architectures (Table 3). SR-MCR delivers consistent absolute gains ranging from +3.5 to +6.2 points over the base models. **Enhancing Logical Trajectories.** While pure mathematical computation benefits slightly less from visual grounding ($\tilde{s}_{vis}$), our unified reward mechanism remains highly effective. The step-wise coherence ($\tilde{s}_{step}$) and semantic consistency ($\tilde{s}_{sem}$) signals impose strict constraints on intermediate reasoning steps, significantly enhancing logical trajectories by preventing logical drift and ensuring complex derivations remain structurally sound. By explicitly penalizing contradictions between adjacent steps, SR-MCR effectively mitigates the compounding errors commonly seen in lengthy Chain-of-Thought generations. Furthermore, it encourages the model to produce intermediate mathematical expressions that strictly align with both the initial premises and the ultimate objective. **Cross-Architecture Robustness.** Crucially, SR-MCR yields significant improvements on the state-of-the-art **Qwen3-VL-8B** model (+3.5 on MathVista, +6.2 on MathVision) without modifying any hyperparameters. This demonstrates the strong cross-architecture robustness of our label-free alignment strategy, validating it as a highly transferable method for enhancing complex reasoning across cutting-edge multimodal foundation models.

3.5 Ablation Studies

Ablation on Reward Components Experimental Setup. We study the unified five-component self-reward $\mathcal{R}$ (Section 2.2) by removing each term individually on both model scales. We also test a two-term variant using only $\tilde{s}_{vis}$ (visual grounding) and $\tilde{s}_{step}$ (step-wise coherence). **Results and Analysis.** Figure 3 shows that: **(1) All components matter.** Removing any reward term consistently

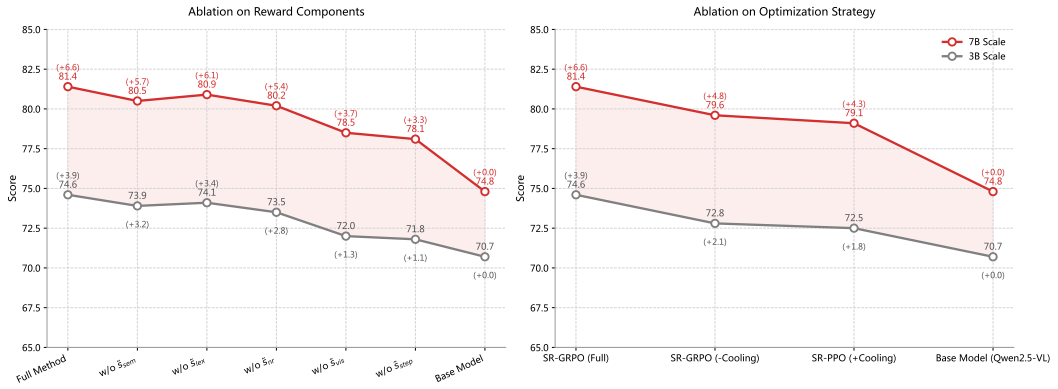


Figure 3: **Ablation studies on reward components and optimization strategies. Left:** Performance impact of individually removing each fine-grained reward component ($\tilde{s}_{sem}$, $\tilde{s}_{lex}$, $\tilde{s}_{nr}$, $\tilde{s}_{vis}$, and $\tilde{s}_{step}$) from the full method. **Right:** Comparison of optimization strategies, illustrating the effectiveness of SR-GRPO over SR-PPO, as well as the positive contribution of the cooling mechanism. Both charts evaluate the 3B and 7B scales, with numbers in parentheses indicating the absolute performance gains relative to the Base Model (Qwen2.5-VL).

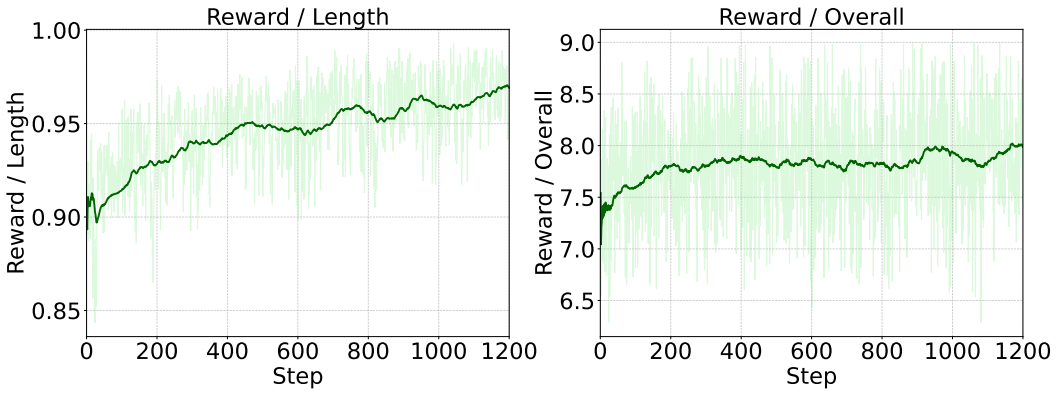


Figure 4: Reward curves for SR-MCR-7B. Both Length and Overall Rewards increase steadily, indicating improved reasoning.

reduces performance on both 3B and 7B models. **(2) Grounding and coherence are critical.** The largest drops occur when ablating $\tilde{s}_{vis}$ or $\tilde{s}_{step}$, confirming their central role in visual-semantic alignment and reasoning stability. **(3) Unified reward is optimal.** While $\tilde{s}_{vis}$+$\tilde{s}_{step}$ performs well (80.7 on 7B), the full five-term reward yields the best results (81.4), showing that auxiliary terms ($\tilde{s}_{sem}$, $\tilde{s}_{lex}$, $\tilde{s}_{nr}$) provide complementary constraints. **Ablation on Optimization Strategy Experimental Setup.** We compare our full method with two variants: (a) SR-GRPO without cooling ($R_{adj} = \tilde{R}$), and (b) SR-PPO, which replaces GRPO with PPO and trains a separate critic. **Results and Analysis.** Figure 3 shows two main findings. **(1) Cooling is essential.** Removing the cooling module leads to consistent drops (3B: 74.6→72.8; 7B: 81.4→79.6), indicating that R_{adj} stabilizes training by down-weighting trivial or overly confident samples. **(2) GRPO outperforms PPO.** SR-GRPO (81.4 on 7B) surpasses SR-PPO (79.1), suggesting that the critic-free GRPO formulation better fits continuous self-reward signals and avoids instability from value estimation.

3.6 Analysis of Training Progress

To examine how reasoning ability evolves during training, we analyze the reward curve of SR-MCR-7B in Fig. 4. The Length Reward rises sharply, indicating that SR-MCR encourages longer and more articulated reasoning chains. Meanwhile, the overall reward increases simultaneously, showing that these extended trajectories correspond to genuine improvements in reasoning quality rather than simple verbosity. This synchronization ensures that the expansion of the reasoning trace is driven by logical necessity rather than stochastic repetition. Furthermore, the steady growth of the

overall reward confirms that our cooling-stabilized objective effectively guides the model toward more reliable reasoning states, successfully suppressing the noise typically associated with longer, more complex outputs. Overall, this pattern reflects that SR-MCR gradually strengthens the model's ability to generate complete, coherent reasoning processes, leading to steady gains on complex tasks.

4 Related Work

Vision-Language Models. Large Vision-Language Models (VLMs) couple visual encoders (e.g., CLIP [32], ViT [9]) with LLMs for multimodal understanding. CLIP benefits from large-scale contrastive pretraining that yields strong zero-shot transfer and alignment between images and text, but it can struggle with compositional queries, and inherits biases present in its training data. **LLaVA** demonstrated that visual instruction-following can be achieved with a lightweight projector and a two-stage pipeline. Subsequent models such as **Qwen-VL** [2] improved perception and multilinguality via larger data and refined architectures, while **GPT-4V** [30] advanced visual reasoning. **InstructBLIP** [8] further enhances cross-modal fusion through Q-Former [21]. These works establish the foundation of modern multimodal reasoning. Complementary efforts improve visual interfaces through hybrid token compression [59] and compact video encoders built on frozen image representations [61]. These methods address visual representation and efficiency, while SR-MCR targets the reliability of generated reasoning.

Alignment Methods for VLMs. Early alignment of vision–language models typically begins with **Supervised Fine-Tuning (SFT)** on curated multimodal datasets [44, 22]. While SFT imparts task formats and basic instruction following, it struggles to model nuanced human preferences [7] and often overfits to annotator distributions. CF-VLM [57] uses targeted counterfactual samples to improve compositional reasoning and factual consistency. **RLHF** extends SFT by training on preference data with a learned reward model [65], improving subjective alignment but incurring substantial annotation cost and introducing potential bias from the reward model itself. In contrast, **RLVR** (Reinforcement Learning with Verifiable Rewards) replaces learned rewards with rule-based or verifier-derived correctness signals, enabling more transparent and scalable optimization for tasks with objectively checkable outcomes (e.g., VQA or grounded reasoning), and avoiding the failure modes associated with preference-model drift.

Reinforcement Learning Algorithms. PPO (Proximal Policy Optimization) [36] is widely used for RL-based alignment; it uses an actor–critic setup with advantage estimation, which is often via GAE, and a clipped surrogate objective to stabilize updates, but requires an additional value network and thus incurs significant compute and memory overhead. To mitigate this computational cost, critic-free approaches such as **REINFORCE** [48] and its variance-reduced variants, **Reinforce++** [16] and **RLOO** (Reinforce Leave-One-Out) [19], have gained traction. These methods estimate policy gradients using multiple samples per prompt, employing baselines derived from peer samples to stabilize training without training a separate value model. Building on this direction, **GRPO** (Group Relative Policy Optimization) [37] constructs relative advantages by comparing rewards within sampled groups of completions. It effectively combines the efficiency of critic-free methods with the stability of group-based normalization, substantially reducing computational and memory cost while retaining strong alignment performance in preference-based generation tasks. At the training-pipeline level, MAT-Agent [56] dynamically coordinates augmentation, optimizer selection, learning-rate schedules, and loss functions for multi-label image classification.

Multi-Agent Reasoning and Expert Coordination. GAM-Agent [58] combines game-theoretic visual reasoning with uncertainty-aware verification. KABB [60] uses knowledge-aware Bayesian bandits for expert selection, while Markets of Minds [62] allocates expert assistance under shared resource constraints. These approaches coordinate multiple agents or experts; SR-MCR improves a single policy through process-level self-rewards.

5 Conclusion

We introduced **SR-MCR**, which fuses five intrinsic self-signals into a reliability-weighted reward and aligns multimodal reasoning through cooling-stabilized GRPO. By jointly modeling answer correctness, visual grounding, logical consistency, semantic coherence, and uncertainty-aware reliability, SR-MCR provides a label-free and process-oriented optimization signal for improving

multimodal reasoning. It improves accuracy, grounding, and coherence over Qwen2.5-VL, demonstrating the effectiveness of self-rewarded alignment in reducing hallucinations and encouraging more faithful reasoning trajectories. Looking ahead, we plan to extend this framework to video reasoning and embodied agents, where temporal consistency, long-horizon planning, and physical grounding pose greater challenges. Furthermore, investigating how self-reward mechanisms scale with larger foundation models remains a promising direction for achieving fully autonomous machine intelligence

Limitations. Despite these promising results, SR-MCR still has several limitations. First, our evaluation mainly focuses on static visual question-answering benchmarks and moderate-scale open-source models, leaving its effectiveness on larger foundation models, dynamic video reasoning, and embodied interactive settings to be further verified. Second, the proposed reward formulation relies on off-the-shelf proxy evaluators, such as visual grounding and consistency models, whose biases or failure modes may introduce noise into the self-rewarding process. Finally, although cooling-stabilized GRPO helps suppress unreliable generations, it may also reduce the diversity of uncertain but potentially valid reasoning paths. Future work should further examine the scalability, robustness, and fairness of self-rewarded multimodal reasoning before real-world deployment.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283 and Grant 62632003, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006 and 23hytd006-2, and in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

[4] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025.

[5] Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning convertsweak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[7] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc.

[8] 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025.

[11] 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)*, pages 14375–14385, June 2024.

[12] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020.

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[14] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019.

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[16] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025.

[17] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth A dozen images. *CoRR*, abs/1603.07396, 2016.

[18] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[19] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! In *DeepRLStructPred@ICLR*, 2019.

[20] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[22] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.

[23] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. pages 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics.

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, December 2023. Association for Computational Linguistics.

[25] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[26] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[28] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[29] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[30] OpenAI. GPT-4V(ision). `https://openai.com/contributions/gpt-4v/`, 2023.

[31] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019.

[35] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015.

[36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[37] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[38] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore, December 2023. Association for Computational Linguistics.

[39] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[40] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-VL technical report, 2025.

[41] Kwai Keye Team. Kwai keye-vl technical report, 2025.

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

[43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[44] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

[45] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[47] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *CoRR*, abs/1908.04319, 2019.

[48] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992.

[49] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

[50] Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. Wethink: Toward general-purpose vision-language reasoning via reinforcement learning, 2025.

[51] Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, Wenzhuo Liu, Xiao Liang, Shuicheng Yan, and Chao Feng. Sail-vl2 technical report, 2025.

[52] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024.

[53] 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. In *Proceedings of CVPR*, 2024.

[54] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics.

[55] Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. pages 7113–7139, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[56] Jusheng Zhang, Kaitong Cai, Yijia Fan, Ningyuan Liu, and Keze Wang. MAT-Agent: Adaptive multi-agent training optimization. In *Advances in Neural Information Processing Systems*, volume 38, pages 77165–77208. Curran Associates, Inc., 2025.

[57] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. CF-VLM:counterfactual vision-language fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pages 58056–58094. Curran Associates, Inc., 2025.

[58] Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. GAM-Agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pages 143569–143629. Curran Associates, Inc., 2025.

[59] Jusheng Zhang, Xiaoyang Guo, Kaitong Cai, Qinhan Lv, Yijia Fan, Wenhao Chai, Jian Wang, and Keze Wang. Hybrid token compression for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31889–31899, June 2026.

[60] Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. KABB: Knowledge-aware Bayesian bandits for dynamic expert coordination in multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 74966–74996. PMLR, 2025.

[61] Jusheng Zhang, Wenhao Wang, Longqi Cai, Liangzhe Yuan, Yuxiao Wang, and Ming-Hsuan Yang. Video encoders built on image representations. *arXiv preprint arXiv:2610.06616*, 2026.

[62] Jusheng Zhang, Xuekun Wu, Wenhao Wang, Mark Wang, and Ming-Hsuan Yang. Markets of minds: A 51-billion-state economy of intelligence. *arXiv preprint arXiv:2610.32190*, 2026.

[63] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.

[64] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[65] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[66] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[67] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

Table 4: **Necessity of Online Optimization.** We compare the base model, rejection sampling fine-tuning (RFT), and SR-MCR at the 7B scale. RFT uses the same 5-term reward to select best-of-N samples ($N = 4$) for supervised training. SR-MCR substantially outperforms RFT, demonstrating the value of online exploration.

Method	Optimization	MMMU	MM Bench	MME	V* Bench	Avg.	Δ
Qwen2.5-VL-7B Base	–	50.3	86.7	89.5	78.5	74.8	–
RFT-7B Best-of-N	Offline SFT	56.4	88.1	91.2	79.1	78.7	+3.9
SR-MCR-7B Ours	Online GRPO	**67.6**	**91.2**	**94.5**	**80.6**	**81.4**	+6.6

A Dataset Details

ScienceQA. ScienceQA is a multimodal multiple-choice dataset containing approximately 20k Question-Answer (QA) pairs. We derive the *ScienceQA-IMG* subset by filtering for instances that include image contexts. In this subset, each entry consists of an image, a question, and multiple candidate options, requiring Vision-Language Models (VLMs) to synergize visual perception with textual reasoning to derive the correct answer. During training, we utilize the full ScienceQA-IMG set but explicitly exclude the hint data provided in the original dataset. This exclusion compels the model to reason from scratch based solely on visual and textual inputs, thereby enhancing its independent problem-solving capabilities.

Visual7W. Visual7W is a dataset comprising 70k QA pairs with multiple-choice options, designed to evaluate object-level semantic understanding beyond simple local region recognition. We employ the complete Visual7W dataset for training to bolster the model's semantic perception and fine-grained visual grounding abilities.

B In-depth Analysis on Optimization and Efficiency

To rigorously validate the necessity of our reinforcement learning framework and assess its computational feasibility, we conducted two additional sets of experiments: (1) a comparison with Rejection Sampling Fine-Tuning (RFT) to verify the need for online optimization, and (2) a comprehensive training efficiency analysis.

B.1 Necessity of Online Optimization: SR-MCR vs. RFT

A critical question regarding the SR-MCR framework is whether the complex GRPO optimization is strictly necessary, or if simply filtering high-quality data using our proposed reward signals would suffice. To investigate this, we implemented a **Rejection Sampling Fine-Tuning (RFT)** baseline. Specifically, for each prompt in the training set, we sampled $N = 4$ responses using the base model π_0, scored them using our unified self-reward R, selected the highest-scoring response, and fine-tuned the base model using standard Supervised Fine-Tuning (SFT).

Analysis. As shown in Table 4, while RFT yields a notable improvement over the base model (+3.9% on average), **SR-MCR significantly outperforms RFT** (+6.6% on average). This performance gap highlights a fundamental limitation of offline methods like RFT: they are strictly bound by the exploration capability of the *initial* policy. If the base model rarely generates a perfectly coherent reasoning chain for a complex query (e.g., in MMMU), RFT has no positive signal to harvest.

In contrast, SR-MCR employs **Online Policy Improvement**. Through iterative GRPO updates, the model's distribution shifts dynamically during training, allowing it to explore reasoning paths that were initially low-probability but high-reward. This "exploration-exploitation" loop enables SR-MCR to solve complex tasks where the base model's naive "best" answer is still suboptimal. Thus, the RL component is not over-engineering; it is essential for breaking the performance ceiling of the base model.

Table 5: **Training Efficiency Analysis.** We compare training time, peak GPU memory usage, and performance gain on 7B models. While SR-MCR introduces overhead compared to SFT, it is substantially more efficient than PPO and achieves higher performance than DPO.

Method	CriticModel?	Reward Type	Time/Step(s)	Peak Mem(GB)	Avg. Acc.(%)
SFT	✗	None	**2.5**	**42**	76.2
SR-PPO	✓	Neural	8.4	78	79.1
DPO	✗	Ref. model	3.8	65	78.8
SR-MCR (Ours)	✗	Neural tools	5.2	56	**81.4**

Table 6: **Results on InternVL3-8B.** InternVL-Base denotes the vanilla InternVL3-8B model without our alignment interventions.

Model	MMMU	MMBench	MME	V*Bench	Average
InternVL-Base	57.3	87.7	89.6	71.2	76.5
+ SR-MCR (Ours)	**61.8**	**89.4**	**92.1**	**74.5**	**79.5**
Gain	+4.5	+1.7	+2.5	+3.3	**+3.0**

B.2 Training Efficiency and Computational Cost

One potential concern with SR-MCR is the overhead introduced by computing the five-component self-reward (involving DETR, DeBERTa, and CLIP) during training. We analyze the training efficiency in Table 5, comparing SR-MCR against standard SFT, PPO, and DPO. Experiments were conducted on 8×H800 GPUs with a global batch size of 128.

B.3 Inference Latency and Memory Overhead

Training Overhead: As illustrated in Table 5, SR-MCR peaks at 56 GB GPU memory usage, a notable 28% reduction compared to the 78 GB required by SR-PPO. This efficiency gain stems from Group Relative Policy Optimization (GRPO), which eliminates the need to maintain a separate critic model. **Inference Latency:** Crucially, the external reward tools and cooling mechanism are utilized exclusively during training, meaning the deployed model is architecturally identical to the base Qwen2.5-VL. Consequently, there is zero structural latency or parameter overhead. However, as SR-MCR encourages Chain-of-Thought (CoT), the generative latency may increase linearly with the length of reasoning traces, a necessary trade-off for improved interpretability.

C Generalization and Reward Mechanism Analysis

To further demonstrate the robustness of SR-MCR, we extend our evaluation along two critical dimensions: (1) cross-architecture generalization on different base models (InternVL2 and LLaVA-NeXT), and (2) a comparative analysis of our tool-based reward system versus LLM-based self-evaluation.

C.1 Cross-Architecture Generalization

To verify the universality of our approach beyond the Qwen architecture, we applied the SR-MCR framework to the InternVL series without distinguishing hyperparameters. As presented in Table 6, our method demonstrates robust generalization capabilities, consistently improving the reasoning and grounding performance of the InternVL baseline. Specifically, SR-MCR achieves a significant gain on V*Bench, indicating that the proposed visual grounding rewards effectively transfer across different visual encoders.

C.2 Ablation of Reasoning Process Quality

To evaluate whether SR-MCR produces superior reasoning trajectories compared to existing visual CoT methods, we conducted a pairwise preference evaluation anchored on the base model. We

Table 7: **Reasoning quality preference evaluation on V-Star Bench.** Values denote the win rate of each reasoning model against the base model Qwen2.5-VL-7B, judged by GPT-4o based on logic, coherence, and completeness.

Category	WeThink 7B	VL-Rethinker 7B	SR-MCR-7B (Ours)
Overall	78.5%	84.2%	**90.6%**
Attribute	75.2%	81.4%	**86.1%**
Spatial	83.5%	88.7%	**97.4%**

Table 8: **Reward Source Comparison.** We compare our external tool-based reward with an internal self-reflection reward, where the LLM acts as a judge. External tools provide objective signals that reduce hallucination on HallBench and improve grounding on V*Bench, whereas self-reflection suffers from mode collapse.

Reward Mechanism	Signal Source	MMMU Reasoning	HallBench Faithfulness	V*Bench Grounding
Base Model	–	50.3	48.3	78.5
Self-Reflection	Internal prompt	55.2 (+4.9)	46.1 (−2.2)	77.8 (−0.7)
SR-MCR (Ours)	**External tools**	**67.6** (+17.3)	**59.4** (+11.1)	**80.6** (+2.1)

extracted full reasoning chains from Qwen2.5-VL-7B (Base), WeThink-7B, VL-Rethinker-7B, and our SR-MCR-7B across the comprehensive V-Star benchmark. Using GPT-4o as an impartial judge, we performed blind A/B comparisons between each reasoning-enhanced model and the base model. The judge assessed which trajectory was superior based on logical soundness, structural coherence, and explanatory completeness.

The results, summarized in Table 7, report the win rate of each model when compared head-to-head against the Qwen2.5-VL-7B baseline. While established reasoners such as WeThink-7B and VL-Rethinker-7B demonstrate strong performance (achieving win rates of 78.5% and 84.2%, respectively), SR-MCR-7B achieves an even higher overall win rate of 90.6%. These results suggest that our method generates more rigorous and reliable reasoning processes, particularly in demanding spatial and attribute-related tasks.

C.3 Ablation: External Tools vs. LLM-as-a-Judge

Our method relies on external tools (DETR, CLIP, NLI) to compute rewards. A natural alternative is to rely on the model's own reflective capabilities ("LLM-as-a-Judge") to self-evaluate its reasoning quality. We conducted an ablation where the unified reward R is replaced by a scalar score generated by the policy model itself via a self-evaluation prompt (e.g., "Rate the quality of your previous reasoning from 0 to 1").

D Empirical Robustness to Reward Noise

To empirically validate the theoretical noise-suppression properties of our cooling mechanism (Eq. 12-14), we conducted an empirical noise injection experiment on the Qwen2.5-VL-7B backbone. We artificially corrupted the visual grounding scores ($\tilde{s}_{vis}$) by adding Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ during training.

As shown in Table 9, without the confidence-aware cooling mechanism, performance degrades sharply (-7.2 points at $\sigma = 0.3$). However, with cooling enabled, the degradation is remarkably graceful (-2.3 points). This confirms that $w_{cool} = \sigma(\kappa(c - l_\theta))$ effectively suppresses gradient updates from high-NLL samples (which are most likely noise-corrupted anomalies), ensuring robust policy alignment even when external reward tools behave unreliably.

Deep Analysis: Why External Tools Matter? Table 8 reveals a critical insight: while Self-Reflection improves general reasoning (MMMU +4.9%), it **degrades** performance on hallucination

Table 9: **Empirical Noise Robustness.** We report average accuracy under varying levels of injected Gaussian noise σ on the visual grounding reward $\tilde{s}_{\text{vis}}$.

Noise Level σ	**0 (Clean)**	**0.1**	**0.2**	**0.3**
SR-MCR (no cooling)	81.4	79.8	77.5	74.2 (-7.2)
SR-MCR (with cooling)	**81.4**	**81.0**	**80.3**	**79.1** (-2.3)

and grounding benchmarks (HallBench -2.2%, V*Bench -0.7%). We attribute this divergence to the *"Self-Delusion"* phenomenon in RL fine-tuning:

- **Breaking the Hallucination Loop:** If a base model hallucinates an object (e.g., seeing a "cat" where there is none), its self-reflection module—sharing the same parameters—is highly likely to validate this hallucination as "correct." This creates a positive feedback loop that reinforces errors. Our external **Object Detection Reward (DETR)** breaks this loop by providing an objective, model-agnostic ground truth surrogate. If DETR does not see a cat, the model is penalized, regardless of its confidence.
- **Objective vs. Subjective Signals:** LLM-as-a-Judge is subjective and prone to sycophancy. In contrast, our **CLIP-based Visual Grounding** and **NLI-based Consistency** signals offer stable, metric-driven rewards. The stark difference in HallBench performance (59.4 vs. 46.1) confirms that relying on external "eyes" is essential for grounding multimodal reasoning in reality, turning the "dependency on external tools" into a vital feature for reliability.

Analysis. The results in Table 5 reveal three key findings regarding the trade-off between cost and performance:

- **Speed vs. PPO:** While SR-MCR is slower than simple SFT (5.2s vs. 2.5s per step), it is **significantly faster than PPO (8.4s)**. This is because PPO requires maintaining and updating a separate Value (Critic) network, which effectively doubles the forward/backward passes. SR-MCR's reward computation, although involving external models, is performed only in inference mode (no gradient computation) and is efficiently batched.
- **Memory Efficiency:** SR-MCR reduces Peak GPU Memory usage by $\sim$**28% compared to SR-PPO** (56GB vs. 78GB). By eliminating the Critic model, GRPO allows us to allocate more memory to larger batch sizes or longer context windows, which is crucial for multimodal reasoning tasks.
- **Inference Cost:** It is important to note that the overhead from DETR, CLIP, and DeBERTa exists **only during training**. The resulting SR-MCR model is architecturally identical to the base model, incurring **zero additional latency** during inference.

In summary, SR-MCR strikes an optimal balance: it accepts a manageable training overhead to leverage sophisticated, process-aware rewards that lightweight methods (like DPO) cannot easily incorporate, while avoiding the prohibitive costs of full Actor-Critic RL.

E Prompts

To elicit structured and reflective reasoning, we design task-specific prompt templates tailored to the distinct formats of each benchmark:

- **MME**: Tailored for binary (Yes/No) questions, this template enforces a strict separation between the internal thought process (encapsulated within `<think>` tags) and the final verdict, ensuring a reflective reasoning stage before the conclusion.
- **V-Star**: Designed for complex multiple-choice questions, this prompt guides the model through a systematic three-step workflow: (1) analyzing the question and visual content, (2) evaluating each option against the evidence, and (3) conducting a critical review before concluding.
- **GPT-4o as a Judge**: Employed to quantitatively assess reasoning quality, this prompt instructs the evaluator to compare two Chain-of-Thought (CoT) responses, prioritizing the clarity, logic, and validity of the reasoning process over the mere correctness of the final answer.

MME Prompt Template

Your task is to answer the following yes/no question based on the provided image(s).
Inside the <think> tag, you must demonstrate a careful and reflective thought process.
After your thinking process, provide the final answer inside the <answer> tag. The answer must be **Yes** or **No**.
Question: {Question}
<think> [Your thought process here, clearly showing the three steps above.] </think>
<answer> {Answer} </answer>

V-Star Prompt Template

Your task is to answer the following multiple-choice question based on the provided image(s).
Inside the <think> tag, you must demonstrate a careful and reflective thought process following these steps: 1. **Analyze the Question and Image(s)**: Break down the question, identify key information in the image(s), and determine what is being asked. 2. **Evaluate Each Option**: Systematically review each option. Provide reasoning for why an option is plausible or incorrect, referencing specific visual evidence and your knowledge. 3. **Critical Review and Final Conclusion**: Compare the options based on your analysis. State your final choice and provide a confident justification for why it is the best answer.
After your thinking process, provide the final answer inside the <answer> tag. The answer must be **only the letter** of the correct option (e.g., A, B, C, D).
Question: {Question}
<think> [Your thought process here, clearly showing the three steps above.] </think>
<answer> {Answer} </answer>

GPT-4o Prompt

You are a professional and impartial AI evaluation assistant. Your task is to carefully read the following question and two alternative 'Chain of Thought' (CoT) answers. You need to determine which Chain of Thought (Choice A or Choice B) is of higher quality, which is clearer, more logical, and more reasonable in its reasoning. Do not choose an answer solely because its final conclusion is correct; focus on evaluating the *reasoning process* itself. In your response, first analyze the strengths and weaknesses of both A and B, and then clearly state your final choice.

F MME Accuracy on Various Steps

As shown in Table 10, MME ACC+ steadily increases with training steps across the four task types, indicating a consistent strengthening of multimodal reasoning. Code reasoning and numerical calculation exhibit the fastest early-stage gains, suggesting that the model quickly absorbs structural and arithmetic patterns. OCR begins with a higher baseline and therefore improves only slightly. Commonsense reasoning grows more slowly, reflecting its relatively greater complexity.

After around 1k steps, most tasks begin to plateau. OCR even shows a minor dip, which may indicate mild overfitting or a trade-off introduced during optimization. Overall, the trends suggest that most improvements occur early, with later steps offering diminishing returns.

G MME Performance Details

Figure 5 illustrates the performance of SR-MCR-7B across the MME benchmark. The model exhibits robust visual understanding, particularly in fine-grained tasks such as landmark, poster, and celebrity recognition. ACC measures the proportion of correctly answered yes-or-no questions, while ACC+ denotes the stricter metric where both questions for a given image must be answered correctly. The consistently strong results across diverse categories highlight SR-MCR-7B's ability to maintain high reasoning quality and prediction accuracy.

Table 10: **MME ACC+ performance across training steps.** We report performance across four task types. Step 0 denotes the baseline Qwen2.5-VL-7B performance.

Task Type	Step 0	Step 500	Step 800	Step 1000	Step 1200
Commonsense Reasoning	58.5	65.1	66.1	**70.0**	68.9
Code Reasoning	65.0	79.2	80.0	85.0	**86.0**
OCR	80.0	82.5	86.5	**90.0**	88.0
Numerical Calculation	55.0	75.0	81.5	85.0	**86.0**

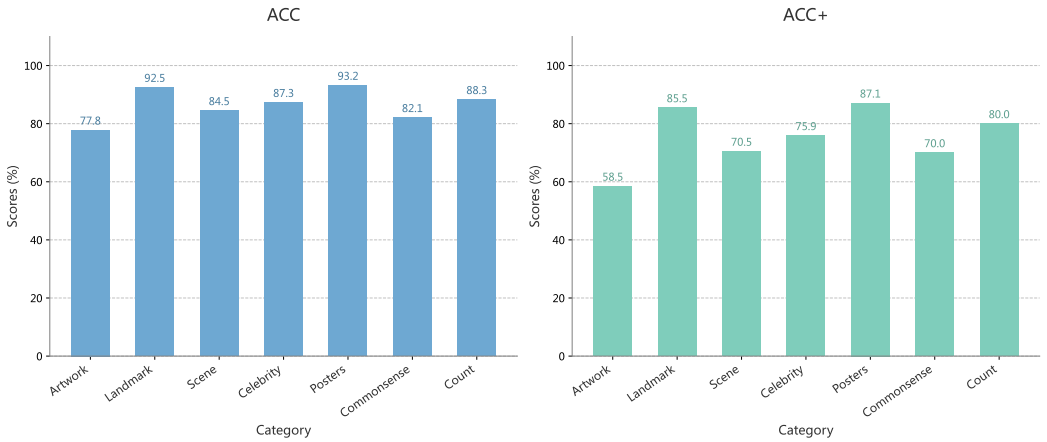


Figure 5: Detailed performance breakdown on the MME benchmark for SR-MCR-7B. The left chart illustrates the accuracy (ACC) across seven categories, while the right chart displays the accumulated accuracy (ACC+).

H Theoretical Analysis: Gradient Properties and Noise Robustness

To theoretically substantiate the superiority of SR-MCR over baseline paradigms (Supervised Fine-Tuning and standard RL), we analyze the asymptotic properties of its gradient estimator. We demonstrate that SR-MCR provides a denser supervision signal than SFT via group-relative ranking and possesses an intrinsic noise-suppression property lacking in standard PPO due to our confidence-aware cooling mechanism.

H.1 Dense Supervision via Relative Advantage

Let the latent true reward distribution be $r(y|x, I)$. **Supervised Fine-Tuning (SFT)** or Rejection Sampling (RFT) effectively optimizes the log-likelihood of the single best-of-N sample, denoted as $y^* = \text{argmax}_{y_i \sim \pi} r(y_i)$. The gradient approximation is:

$$\nabla_\theta \mathcal{J}_{\text{SFT}} \approx \mathbb{E}_{y^* \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(y^*|x, I)] \tag{10}$$

Analysis of Gradient Sparsity: Theoretically, SFT approximates the reward landscape as a Dirac delta function centered at y^*, i.e., $r(y) \approx \delta(y - y^*)$. This introduces two critical limitations: (1) **Information Loss**: It treats y^* as a ground-truth label (probability 1), discarding the valuable relative signals contained in suboptimal samples; (2) **High Bias**: It reinforces y^* even if y^* is merely a "lucky guess" or a local optimum, leading to potential overfitting.

In contrast, **SR-MCR (via GRPO)** optimizes the relative advantage within a sampled group $\{y_1, ..., y_G\}$. The gradient estimator is defined as:

$$\nabla_\theta \mathcal{J}_{\text{SR-MCR}} \approx \frac{1}{G} \sum_{i=1}^{G} \underbrace{(R_{adj}(y_i) - \bar{R})}_{\text{Relative Advantage } \hat{A}_i} \nabla_\theta \log \pi_\theta(y_i|x, I) \tag{11}$$

Proof of Superiority: Unlike SFT, SR-MCR utilizes the *full distribution* of the generated group. Even a sample with mediocre quality contributes to the optimization: if a sample y_i is better than

Table 11: **Theoretical Comparison of Alignment Paradigms.** Comparison of gradient properties and resource requirements. SR-MCR occupies a unique sweet spot: it offers dense, process-level supervision similar to PPO but with the low memory footprint and intrinsic noise robustness lacking in standard RL methods.

Paradigm	Gradient Estimator	Supervision Signal	Noise Robustness	Critic Overhead	Granularity
SFT / RFT	$\nabla \log \pi(y^*)$	Sparse (Best-of-N)	Low (Overfits bias)	None	Outcome
PPO (RLHF)	$\hat{A}_t \nabla \log \pi(y_t)$	Dense (All samples)	Low (Prone to Hacking)	High	Outcome/Process
DPO	$\nabla \log \frac{\pi_\theta}{\pi_{ref}}$	Pairwise Relative	Medium (Implicit)	None	Outcome
SR-MCR (Ours)	$(R_{adj} - \bar{R}) \nabla \log \pi$	**Dense (Group Relative)**	**High (via Cooling)**	**None**	**Process**

the group average $\bar{R}$ (yet worse than y^*), it yields a positive signal ($\hat{A}_i > 0$); conversely, inferior samples provide negative gradients ($\hat{A}_i < 0$). This **fine-grained relative ranking** prevents the "mode collapse" often observed in SFT and drives the policy to continuously explore the upper bounds of the reward landscape.

H.2 Noise Robustness via Cooling Mechanism

A fundamental challenge in "Self-Reward" implies that the proxy reward $R(y)$ inherently contains noise ϵ (e.g., false positives from DETR or hallucinations in self-reflection). Standard RL algorithms (like PPO) blindly optimize this noisy signal, creating a risk of "reward hacking." We theoretically prove that our Cooling mechanism functions as a **Dynamic Variance Reducer**.

Recall our adjusted reward formulation: $R_{adj} = R(y) \cdot w_{\text{cool}}$, where $w_{\text{cool}} = \sigma(\kappa(c - l_\theta(y)))$ is the confidence-gated weight derived from the negative log-likelihood (NLL) $l_\theta(y)$.

Proposition 1 (Gradient Suppression under Uncertainty). Assume the reward estimation contains noise such that $R_{proxy} = R_{true} + \epsilon$. When the model hallucinates or exploits the reward model (reward hacking), the generation typically manifests as a statistical anomaly with high perplexity (high NLL $l_\theta \gg c$). As $l_\theta(y_i) \to \infty$, the cooling weight follows $w_{\text{cool}} \to 0$. Consequently, the norm of the gradient contribution is bounded:

$$\|\nabla_\theta \mathcal{J}^{(i)}_{\text{Cooling}}\| \leq w_{\text{cool}} \cdot \|\hat{A}_i \nabla_\theta \log \pi_\theta\| \xrightarrow{l_\theta \gg c} 0 \tag{12}$$

Interpretation: The cooling term acts as a differentiable "soft gate":

- **Case A (High Reward, High Uncertainty):** If $R(y_i)$ is high but the model is uncertain (indicating a potential hallucination or out-of-distribution reward hack), $w_{\text{cool}} \to 0$ effectively suppresses this gradient, preventing the policy from learning the noise.
- **Case B (High Reward, High Confidence):** If $R(y_i)$ is high and the model is confident ($l_\theta < c$), $w_{\text{cool}} \approx 1$, allowing full optimization strength.

This mechanism provides a theoretical guarantee that SR-MCR is robust against reward noise, applying a **soft trust-region constraint** based on generation confidence.

H.3 Theoretical Bound on Reward Hacking

We formally define the **Reward Hacking Region** $\mathcal{H}$ as the subspace of generations where the proxy reward is deceptively high ($R(y) > \tau$) but the generation is statistically anomalous with respect to the truthful language manifold ($l_\theta(y) \gg c$).

Proposition 2 (Exponential Damping of Hacking Gradients). In SR-MCR, the gradient contribution from the hacking region $\mathcal{H}$ is exponentially suppressed compared to standard Policy Gradient (PG).

Proof. Let $g_{\text{PG}}(y)$ be the standard gradient vector. The magnitude of the update in SR-MCR is scaled by the sigmoid cooling factor:

$$\|g_{\text{SR}}(y)\| = \sigma(\kappa(c - l_\theta(y))) \cdot \|g_{\text{PG}}(y)\| = \frac{1}{1 + e^{-\kappa(c - l_\theta)}} \|g_{\text{PG}}(y)\| \tag{13}$$

For a hacking sample $y \in \mathcal{H}$ where $l_\theta(y) \to \infty$, the scaling factor decays exponentially:

$$\lim_{l_\theta \to \infty} \frac{\|g_{\text{SR}}(y)\|}{\|g_{\text{PG}}(y)\|} \approx e^{-\kappa l_\theta} \to 0 \tag{14}$$

This ensures that the policy update is strictly confined to the reliable region of the language model, theoretically preventing the catastrophic forgetting or mode collapse often triggered by reward hacking.

Discussion: The Bias-Variance Trade-off. It is worth acknowledging that the cooling mechanism introduces a bias by down-weighting low-confidence samples, even if they are factually correct. However, in the context of *Self-Reward*, where the reward signal stems from imperfect proxies (external tools), the *variance* of the gradient (driven by reward hacking and hallucinations) is the dominant factor destabilizing training. SR-MCR explicitly trades off a small degree of bias (ignoring potentially correct but uncertain samples) for a significant reduction in variance. This ensures a monotonic improvement trajectory, preventing the model from collapsing into high-reward, low-quality modes.

I Case Study

Table 12 and Table 13 illustrate success scenarios where SR-MCR-7B outperforms baselines. In Table 12, while Qwen2.5-VL struggles with color recognition and others fail to detect the object entirely, SR-MCR-7B accurately identifies the suitcase, demonstrating robust grounding. Similarly, Table 13 highlights our model's precision in localizing a small-scale yellow car under challenging lighting, correctly inferring spatial relationships where competing methods falter.

However, limitations persist in specific challenging scenarios. As shown in Table 14 and Table 15, all evaluated models, including SR-MCR-7B, fail to provide correct answers. Table 14 reveals a susceptibility to context-induced hallucination, where models mistake a red toothbrush for blue due to the dominant clinical environment colors. Table 15 highlights difficulties in spatial reasoning under ambiguous perspectives, leading to consistent directional errors across all models.

J Hyperparameters

In all experiments, we set the maximum prompt length and maximum response length to 4096. For GRPO optimization, KL regularization is enabled with a coefficient of 1.0×10^{-2}, and each prompt is sampled with $n = 4$ generations. The rollout batch size and global batch size are both set to 4. We train the target model in BF16 precision using AdamW with a learning rate of 1.0×10^{-6}. For the reward function, we use `all-MiniLM-L6-v2` as the sentence transformer and set the target response length to 1536.

Table 12: Qualitative comparison on V*Bench (Attribute Recognition). The question asks for the color of a suitcase. **Correct Answer: D (brown)**.

User Prompt

Question: What is the color of the suitcase?



Options: (A) red (B) gray (C) black (D) brown

Model	Reasoning & Output
Qwen2.5-VL-7B-Instruct	The suitcase is located near the bottom right... The color is not clearly distinguishable... appears to be dark... either black or brown. ✗ **Answer:** C
SAIL-VL2-8B-Thinking	*Got it, let's try to figure out the color... The options are red, gray, black, brown... Wait, maybe in the image, the suitcase is gray...* ✗ **Answer:** B
VL-Rethinker-7B	*There are people standing near the entrance, but no suitcases are clearly visible in the image. ...* ✗ **Answer:** (No Answer)
SR-MCR-7B (Ours)	*1. Analyze: Focus on the suitcase's appearance... 2. Evaluate: (d) "brown" indicates a specific shade that suggests a structured look. ... 3. Conclusion: A consistent tone suggests a deliberate design choice.* ✓ **Answer:** D

Table 13: Qualitative comparison on V*Bench (Spatial Relationship)). The question asks for the position of a yellow car relative to a pool. **Correct Answer: A (left)**.

User Prompt

Question: Is the yellow car on the left or right side of the pool?



Options: (A) left (B) right

Model	Reasoning & Output
VL-Rethinker-7B	*The yellow car appears to be situated on the right side of the pool... Thus, the spatial arrangement indicates the right side. ...* ✗ **Answer:** B
WeThink-7B	*The yellow car is visible on the right side of the pool. ... Therefore, the yellow car is on the right side.* ✗ **Answer:** B
Kimi-VL-A3B-Thinking	*... On the right side, near the edge of the pool, there's a yellow vehicle. ... So the answer should be the right side...* ✗ **Answer:** B
SR-MCR-7B (Ours)	*1. Analyze: Identifying key objects... 2. Evaluate: The car is clearly positioned on the left side of the pool... 3. Conclusion: Option (A) is the correct choice.* ✓ **Answer:** A

Table 14: Failure Case on V*Bench (Attribute Recognition). The question asks for the color of a toothbrush. **Correct Answer: D (red)**.

User Prompt

Question: What is the color of the toothbrush?



Options: (A) green (B) blue (C) yellow (D) red

Model	Reasoning & Output
Qwen2.5-VL-7B-Instruct	The dental office features light green cabinets... The toothbrush inside the holder appears to have a handle that matches the teal/blue decor... It is likely blue. ✗ **Answer:** B
SAIL-VL2-8B-Thinking	*Scanning the room... The environment is sterile with mint tones. The object in the cup holder looks blueish under this lighting... Not seeing any bright red objects.* ✗ **Answer:** B
VL-Rethinker-7B	*Focusing on the dental instruments... The toothbrush seems to be green or blue, blending with the tray. No red toothbrush is clearly visible.* ✗ **Answer:** A
SR-MCR-7B (Ours)	*1. Analyze: Focus on the dental setting... 2. Evaluate: ... (b) "blue" suggests a possibility of a specific color associated with the setup. ... (d) "red" is less indicated... 3. Conclusion: ...consistent color scheme... indicates a specific color for dental tools.* ✗ **Answer:** B

Table 15: Failure Case on V*Bench (Spatial Relationship). The question asks for the position of a soccer ball relative to a bench. **Correct Answer: B (left)**.

User Prompt

Question: Is the soccer ball on the left or right side of the long bench?



Options: (A) right (B) left

Model	Reasoning & Output
Qwen2.5-VL-7B-Instruct	The bench is located along the path. From this angle, the soccer ball appears to be positioned on the right side of the bench structure... ✗ **Answer:** A
SAIL-VL2-8B-Thinking	*Locating the ball and the bench... The bench stretches along the right. The ball sits on the ground to the right of the bench's end... roughly aligned with the trees.* ✗ **Answer:** A
VL-Rethinker-7B	*Analyzing spatial layout... The soccer ball is clearly visible on the right side relative to the bench... The perspective makes it appear further right.* ✗ **Answer:** A
SR-MCR-7B (Ours)	*1. Analyze: Focus on the bench and surrounding elements. 2. Evaluate: (a) "right" does not indicate... (b) "left" suggests... 3. Conclusion: ...alignment with the bench's edge... indicates the ball's position relative to the bench.* ✗ **Answer:** A

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims about SR-MCR improving multimodal coherent reasoning, which are supported by the empirical results in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We demonstrate limitations in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper proposes an empirical reinforcement learning framework and does not include theoretical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides experimental setups, formulas for self-rewards, and implementation details necessary for reproducibility in Section 3 and 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper currently does not provide a repository URL or open-source data and code in this version.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the base model, benchmarks used, and evaluation setups in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports average accuracy metrics across benchmarks without multiple runs for error bars due to computational overhead.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Table 5 and the efficiency analysis in the Appendix specify the use of 8x H800 GPUs and provide the peak memory and time per step.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work builds upon standard visual instruction tuning and conforms to general ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The broader impacts discussion in the impact statement section has been commented out.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new high-risk datasets or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Datasets and base models are properly cited, but their licenses and terms of use are not explicitly mentioned in the text.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce novel datasets or assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research relies on automated metrics and does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.