Enhancing the Outcome Reward-based RL Training of MLLMs with Self-Consistency Sampling

Jiahao Wang[1,3*†], **Weiye Xu**[2,3*†], **Aijun Yang**[1], **Wengang Zhou**[2],
Lewei Lu[4], **Houqiang Li**[2], **Xiaohua Wang**[1✉], **Jinguo Zhu**[3✉]
[1]Xi'an Jiaotong University, [2]University of Science and Technology of China,
[3]Shanghai Artifcial Intelligence Laboratory, [4]SenseTime Research
wjhwdscience@stu.xjtu.edu.cn, ustcxwy0271@mail.ustc.edu.cn
xhw@mail.xjtu.edu.cn, lechatelia@gmail.com

Abstract

Outcome-reward reinforcement learning (RL) is a common—and increasingly significant—way to refine the step-by-step reasoning of multimodal large language models (MLLMs). In the multiple-choice setting—a dominant format for multimodal reasoning benchmarks—the paradigm faces a significant yet often overlooked obstacle: unfaithful trajectories that guess the correct option after a faulty chain of thought receive the same reward as genuine reasoning, which is a flaw that cannot be ignored. We propose Self-Consistency Sampling (SCS) to correct this issue. For each question, SCS (i) introduces small visual perturbations and (ii) performs repeated truncation-and-resampling of an initial trajectory; agreement among the resulting trajectories yields a differentiable consistency score that down-weights unreliable traces during policy updates. Based on Qwen2.5-VL-7B-Instruct, plugging SCS into RLOO, GRPO, and REINFORCE++ series improves accuracy by up to 7.7 percentage points on six multimodal benchmarks with negligible extra computation. SCS also yields notable gains on both Qwen2.5-VL-3B-Instruct and InternVL3-8B, offering a simple, general remedy for outcome-reward RL in MLLMs. Our code is available at https://github.com/GenuineWWD/SCS.

1 Introduction

With the remarkable success of state-of-the-art Large Language Models (LLMs) such as OpenAI-o3 and DeepSeek-R1[10], enhancing models' reasoning abilities through Reinforcement Learning (RL) method have been seen as a mainstream route toward Artificial General Intelligence (AGI) [13, 47]. A series of studies has reported breakthrough performance when training LLMs with RL, highlighting the considerable potential of RL-based training paradigms [39, 10, 1, 14]. Driven by the rapid development of LLMs, considerable progress [54, 29, 26, 6] has been made in improving multimodal capabilities of Multimodal Large Language Models (MLLMs). In particular, RL-driven approaches have markedly improved MLLMs' capabilities in mathematics, video comprehension, and visual logical reasoning [45, 42, 49, 26].

However, applying reinforcement learning to multimodal *multiple-choice problems* exposes a critical weakness. Existing outcome-based approaches compute the outcome reward solely by checking whether the chosen option matches the ground truth, completely ignoring the faithfulness of the intermediate reasoning trajectory. This makes it easy for the model to reach the correct answer via spurious or hallucinatory chains of thought. We illustrate the issue with an exploratory experiment on Geometry3K [22]. As shown in Figure 2(a), training in the multiple-choice setting improves

[*]equal contribution; [†] interns at OpenGVLab, Shanghai AI Laboratory; [✉] corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-4395

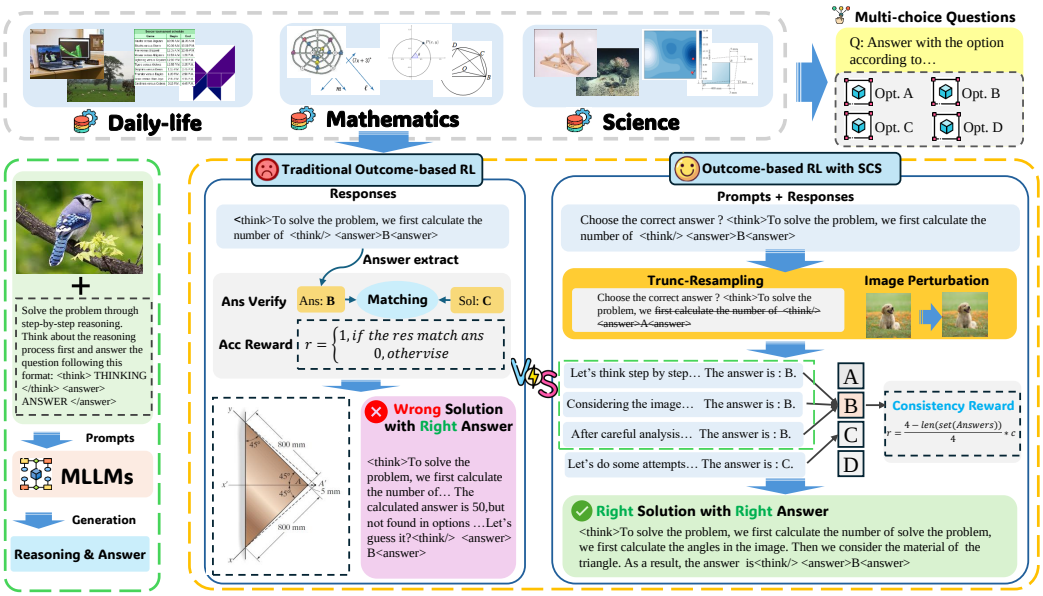


Figure 1: **Overview of our work.** When applied to multiple-choice problems, traditional outcome reward-based RL methods that rely solely on accuracy-based rewards often lead to situations where the selected option is correct, but the reasoning process is flawed. Our method introduces an additional consistency reward, which significantly reduces the occurrence of such cases.

accuracy by only 5.6%, which is 6.4 percentage points lower than the 12.0% gain obtained under the open-ended QA setting. Qualitative analyses in Figures 2(b) and (d) confirm that multiple-choice training encourages unfaithful reasoning: the model often generates incorrect rationales yet still selects the correct option by chance. We further probe this phenomenon on additional benchmarks (ScienceQA [35], MMMU [56], M^{3}CoT [7] and MathVision [43]). For each question we first generate a complete rationale, then truncate the rationale at several positions and resume generating from each truncated prefix. Figure 2(c) shows that divergent answer options frequently emerge from identical prefixes, underscoring the pervasiveness of unfaithful trajectories in the sampling procedure under multiple-choice regime.

To solve this challenge, we introduce Self-Consistency Sampling (SCS), a framework that improves the faithfulness of reasoning trajectory. As shown in Figure 1, building on recent advances in consistency-based reasoning [46, 30, 41], SCS first generates an initial trajectory and then explores its neighboring trajectories through two mechanisms. In the *truncation–resampling* mechanism, the initial trajectory is truncated and generation is resumed to produce several continuations; in the *visual-perturbation* mechanism, the input image is added by subtle gaussian noise before resampling, encouraging the policy to reason over perturbed visual evidence. The resulting set of trajectories provides multiple answer candidates for the same question, from which SCS derives a consistency reward that quantifies the agreement among their responses. This reward, reflecting the reliability of intermediate reasoning, is used to update the policy, steering it toward trajectories that remain self-consistent under perturbations and thereby enhancing overall reasoning faithfulness.

We assess the effectiveness of self-consistency sampling (SCS) when combined with several outcome-reward reinforcement-learning algorithms—namely RLOO [19], REINFORCE++ series [14], and GRPO [39]. Based on Qwen2.5-VL-7B-Instruct [3], on six benchmarks covering four task categories, integrating SCS with RLOO yields an average absolute improvement of 7.7 percentage points over the RLOO baseline. When SCS is incorporated into REINFORCE++-baseline, REINFORCE++, and GRPO, the respective gains are 1.7, 2.0, and 0.9 percentage points relative to their no-SCS counterparts. SCS generalizes across models of different scales and architectures, achieving consistent gains—for example, 3.2 and 1.6 more percentage points on Qwen2.5-VL-3B-Instruct and InternVL3-8B when applied with RLOO. Trajectory analysis reveal that models trained with SCS generate substantially more faithful reasoning paths (Figure 2). Ablation studies further indicate that the two components of SCS are essential: Truncation-Resampling (TR) and Visual Perturbation (VP)

https://doi.org/10.52202/085713-4395

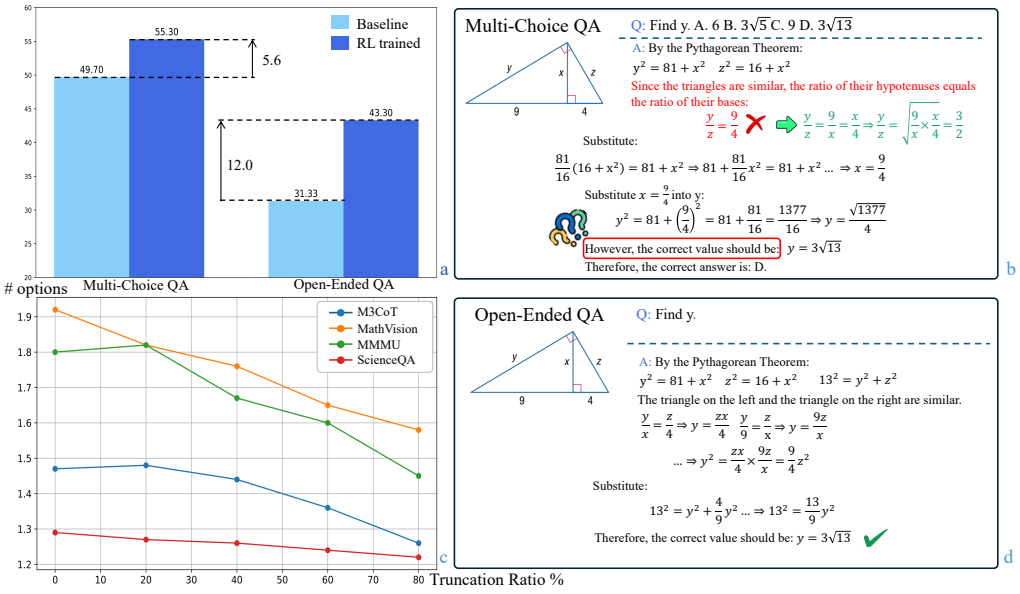


Figure 2: **Illustration of unfaithful reasoning phenomenon.** (a) Compared with open-ended questions, training in the multi-choice format yields smaller performance gains. (b) Examples of unfaithful reasoning generated by model on Multi-Choice QA problems. (c) The curve of the relationship between the average number of final options for each question and the trajectory truncation ratio(%). (d) Correct reasoning trajectories generated by models with Open-Ended QA form.

contribute 5.2 and 5.0 percentage-points improvements on their own, and together recover the full 7.7 points gain. Our main contributions are as follows:

- We empirically show that outcome-reward training encourages unfaithful reasoning in multimodal multiple-choice tasks, where models often arrive at correct answers through incorrect or inconsistent reasoning processes.
- We propose a novel Self-Consistency Sampling, which resamples truncated rationales and perturbed visuals to compute a consistency reward, thereby aligning reasoning trajectories with their answers.
- Integrated with RLOO, GRPO, and REINFORCE++ variants, Self-Consistency Sampling yields up to +7.7 pp accuracy with Qwen2.5-VL-7B-Instruct [3] and exhibits stable performance in extensive ablations.

2 Related Work

RL methods for LLMs and MLLMs. Reinforcement Learning (RL) [37, 39, 1, 14] has demonstrated significant success across a wide range of domains, particularly in game playing [27, 40, 17, 16], robotic control [51, 12, 18], and autonomous driving [5, 36, 23]. With the rapid advancement of Multi-modal Large Language Models (MLLMs), RL techniques have increasingly been adopted to enhance model performance on tasks such as VQA [8, 21], image captioning [28, 11], visual reasoning [52, 50], and referring expression comprehension [55, 25]. Early approaches, including Reinforcement Learning from Human Feedback (RLHF) [4], often employed methods like PPO [37] and DPO [32] to align model outputs with human preferences data. Some works introduce more efficient alternatives like RLOO [1], REINFORCE++ [14], ReMax [20], Group-wise Relative Policy Optimization (GRPO) [39], eliminates critic model from RL training, significantly reducing the computational overhead.

Reward Models in RL training. Reward models play an important role in training LLMs and MLLMs with reinforcement learning. Common types of reward models include generative rewards [57, 60, 24], implicit rewards [34, 48, 59], process rewards [53, 44, 9], and outcome rewards [14, 20, 39]. Generative reward models leverage auxiliary generative models to evaluate or

https://doi.org/10.52202/085713-4395

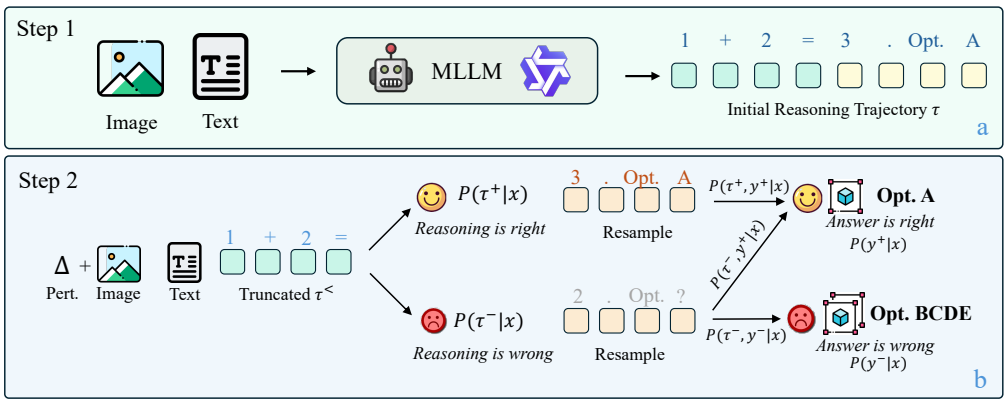


Figure 3: **Pipeline of our method.** (a) illustrates the initial reasoning trajectory generated by the MLLM; (b) shows the sampling and probability propagation across reasoning steps.

guide the behavior of the primary model, often providing flexible supervision in an unsupervised or weakly-supervised manner. Unlike explicit reward models, implicit reward methods, such as DPO [32] and SLiC-HF [59], aligned variables or loss functions to convey optimization signals with reduced resource overhead. Process reward models [53, 44, 9] assign feedback based on the reasoning trajectory rather than the final output, aiming to encourage interpretable and robust problem-solving. In contrast, outcome reward models [14, 20, 39] focus solely on the correctness of the final answer, ignoring the generation path. While efficient, this may lead to issues like reasoning hallucinations or flawed logic. Process reward models can mitigate such problems but are typically resource-intensive. To strike a balance, self-consistency sampling adopt outcome-based reward formulations while incorporating consistency checks during reasoning.

3 Method

3.1 Preliminary

Outcome Reward. Outcome reward-based Reinforcement Learning is a training paradigm that exploits objectively checkable task outcomes to train large language models (LLMs) and Multi-modal large language models (MLLMs). Unlike RL from Human Feedback (RLHF), where reward signals are noisy, subjective, and expensive to collect, outcome reward-based RL relies on a deterministic verification protocol that returns a scalar reward $r \in \{0, 1\}$ indicating whether the model's output satisfies a well-defined correctness criterion. Typical application domains include mathematics, programming, structured reasoning questions with a single ground-truth solution and fine-grained tasks such as image detection and OCR. Formally, let π_θ be an autoregressive policy that generates a trajectory $\tau = (y_1, \ldots, y_T)$ token-by-token given a problem instance x. Once the final answer token y_T is produced, an external verifier $\mathcal{V}$ checks τ and returns.

$$r = \mathcal{V}(x, \tau) = \begin{cases} 1 & \text{if } \tau \text{ solves } x, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Outcome Reward-based RL Algorithm. The learning objective is the standard expected-reward maximization, optimized through any policy-gradient estimator (e.g., GRPO [39], RLOO [19] and REINFORCE++ [14]). GRPO refines the PPO [38] framework by eliminating the separate value network—thereby lowering computational overhead—and employs a group-relative advantage estimator together with an explicit KL-divergence penalty to keep policy updates both efficient and stable. GRPO adapts PPO to the outcome reward-based RL setting while discarding the value-network critic. In GRPO, a group-relative advantage is then computed for every response,

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}, \tag{2}$$

where mean and std are taken over the G rewards.

https://doi.org/10.52202/085713-4395

Algorithm 1 Outcome Reward-based RL Training with Self-Consistency Sampling (SCS)

Require: Dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, pretrained model parameters θ, truncation ratio $k\,(0 < k < 1)$, number of resample m, consistency weight c, learning rate α, reample number N.

Initialize optimizer $\mathcal{O}$ with θ
for each minibatch $\{x\}$ in $\mathcal{D}$ **do**
  Sample initial answer & reasoning trajectory $a, \tau \sim \pi_\theta(\cdot \mid x)$;
  $r \leftarrow r_{acc}$;
  $\mathcal{A} \leftarrow \emptyset$;
  for $t = 1$ to m **do**
    $\tau^{<} \leftarrow$ Truncate(τ, k);
    Add Noise to Image in $x^* \leftarrow x + \mathcal{N}(\mathbf{0}, \sigma_t^2)$;
    Sample new answer a_t after $\tau^{<}, x^*$;
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_t\}$;
  end for
  $r_{\text{con}} \leftarrow (N - |A|)$;
  $r \leftarrow r_{\text{for}} + r_{\text{acc}} + r_{\text{con}}$;
  Compute baseline of reward b;
  Compute policy gradient $g \leftarrow \nabla_\theta \log \pi_\theta(a_0 \mid x)(r - b)$;
  $\theta \leftarrow \theta + \alpha\mathcal{O}(g)$;
end for
return Updated parameters θ

Reinforcement Learning with Leave-One-Out (RLOO) generates K independent roll-outs $\{\tau_i\}_{i=1}^{K}$ for the same input. Since the baseline reuses the rewards of peer roll-outs, RLOO enjoys lower variance than REINFORCE while avoiding an explicit value network.

The core idea of REINFORCE++ is to fold a suite of PPO-style optimizations into the classic REINFORCE algorithm to boost both performance and stability. Specifically, REINFORCE++ augments plain REINFORCE with token-level KL penalties, mini-batch updates, reward normalization and clipping, and advantage normalization.

3.2 Self-Consistency Sampling

This section presents the theoretical foundation of our method, including the underlying assumptions, modeling process, and derivation steps. Figure 3 illustrates the modeling of our SCS method.

We see the entire reasoning process as a tree structure and make three key assumptions:

1. **Uniqueness of the correct trajectory.** There is exactly one correct reasoning trajectory in the tree. **Justification:** This assumption is reasonable, as most problems admit a unique solution.
2. **Leaf/Choice Alignment.** Every trajectory of the reasoning tree ends at one of the answer choices. **Justification:** In the majority of cases, models ultimately outputs a single option as its solution, whether it is correct or not.
3. **Relationship between correct/incorrect reasoning trajectories and answer options.** If the model follows the correct trajectory, it must arrive at the correct option; if it follows an incorrect trajectory, it may still pick either the correct option or a wrong one. **Justification:** With correct reasoning the model just retrieve the correct answer, whereas an incorrect reasoning chain can lead to multiple outcomes (e.g., guessing when the correct answer is not reachable). For the quantitative verification, refer to Appendix B.1.

When the model follows an incorrect reasoning trajectory, it may still select the correct answer y^+ with probability $p(y^+ \mid x)$. Therefore, the outcome reward-based defined in Eq. 1 can arise from two different reasoning trajectories:

$$P(y^+, \tau^+ \mid x) + P(y^+, \tau^- \mid x), \tag{3}$$

where τ^+ represents the correct reasoning trajectory, and τ^- represents the incorrect one.

https://doi.org/10.52202/085713-4395

Table 1: **Datasets and evaluation benchmarks used in this work.** We perform a data filtering step and retain only multiple-choice questions that include an associated image to construct our dataset.

Category	Name	Domain / Task	#Scale	#Filtered
Training Datasets				
	M^3CoT [7]	General	7.8 k	7.8 k
	Geometry3K [22]	Geometry	2.1 k	2.1 k
	ScienceQA [35]	Science	12.7 k	6.2 k
Evaluation Benchmarks				
	M^3CoT [7]	General	2.3 k	2.3 k
	ScienceQA [35]	Science	4.2 k	4.241 k
	MathVision [43]	Mathematics	3.0 k	1.5 k
	We-Math [31]	Mathematics	5.0 k	5.0 k
	MMMU [56]	Multi-subjects	900	851
	MathVerse [58]	Mathematics	3.9 k	2.1 k

When only the option accuracy reward $r = r_{\text{acc}}$ is applied, both trajectories receive the same reward whenever the final prediction is correct, regardless of the reasoning path. As a result, the model still has a nonzero probability of producing an unfaithful reasoning process:

$$P(\tau^- \mid y^+) = \frac{P(y^+, \tau^- \mid x)}{P(y^+, \tau^- \mid x) + P(y^+, \tau^+ \mid x)}. \tag{4}$$

To mitigate this issue, we introduce a consistency reward r_{con} to penalize incoherent or spurious reasoning patterns. Assuming that after producing a reasoning trajectory τ, the model randomly samples from N candidate options, forming a set of collected answers $\mathcal{A}$. Intuitively, if the reasoning trajectory is wrong, $\mathcal{A}^-$ will contain more diverse answers among repeated samplings. We define the consistency reward as:

$$r_{\text{con}} = \frac{1}{N}\big(N - |\mathcal{A}|\big)c, \tag{5}$$

where c is a scaling coefficient that controls the strength of regularization. A smaller $|\mathcal{A}^-|$ (indicating more consistent outcomes) yields a higher reward.

For the two types of reasoning trajectories, the corresponding consistency rewards are given by:

$$r^-_{\text{con}} = \frac{1}{N}\big(N - |\mathcal{A}^-|\big)c, \quad r^+_{\text{con}} = \frac{1}{N}\big(N - 1\big)c. \tag{6}$$

Here, according to the assumption, correct trajectory τ^+ obtains the maximal consistency reward, since it deterministically leads to the correct answer.

Meanwhile, accuracy reward and format reward are also applied. Since $\mathbb{E}(|\mathcal{A}^-|) > 1$ (detailed derivation process can be found in Appendix B.2), $r^+_{con} > r^-_{con}$. Therefore, the correct trajectories τ^+ will be optimized to have more advantage to appear.

3.3 Algorithm Details

Based on the above theoretical derivation, we design the Self-Consistency Sampling (SCS) (see Algorithm 1). The SCS consists of two main blocks: *truncation–resampling* and *visual-perturbation*.

Truncation–Resampling. For each single-choice question in dataset, SCS first generate an initial reasoning trajectory τ. Instead of using outcome reward-based method to extract option answer in τ, we truncate initial trajectory τ with truncation ratio k to an incomplete trajectory $\tau^<$. This truncated reasoning trajectory $\tau^<$ is then used as a prefix for multiple resampling steps. During each resampling iteration, the model continues reasoning from $\tau^<$ and generates a new answer a_t. All sampled answers are collected into a set $\mathcal{A}$. The consistency reward $r_{\text{cons}} = c(N - |\mathcal{A}|)$ is computed based on the diversity of the sampled answers — if the answers are highly consistent, indicating stable reasoning given the same prefix, the value of r_{con} is high (since $|\mathcal{A}|$ is low). This serves as a proxy for reasoning stability and correctness.

Visual-Perturbation. To enhance the effectiveness of consistency evaluation, we introduce stochastic visual perturbations when generating each resampled response from a initial trajectory. Rather than

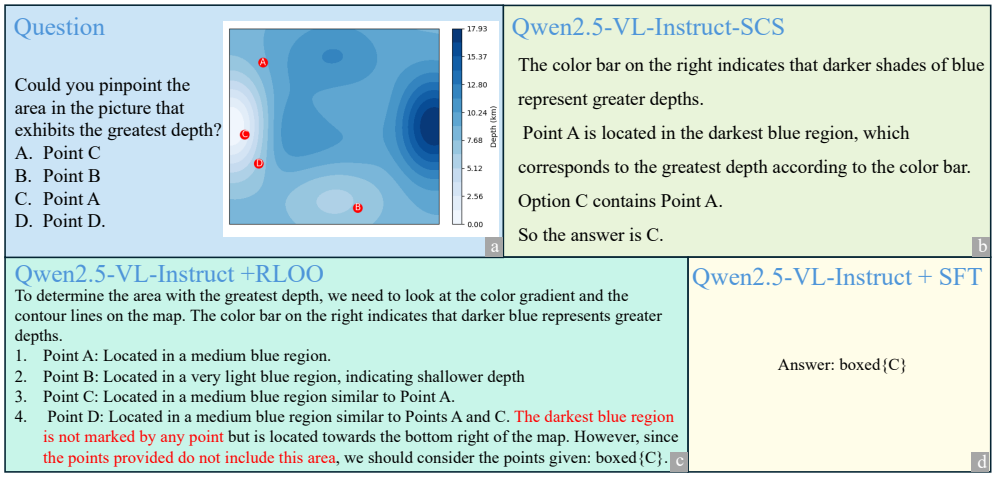


Figure 4: **Comparison of model response of different models.** (a) Question image selected from M3CoT. (b) Reasoning trajectory of Qwen2.5-VL-Instruct trained by RLOO as baseline. (c) Reasoning trajectory of Qwen2.5-VL-Instruct trained with SCS. (d) Reasoning trajectory of Qwen2.5VL-Instruct trained by SFT. The red text is incorrect reasoning part.

applying fixed-strength noise, we sample a unique perturbation strength for each continuation to expose the model to a broader range of visual variations. Formally, for each continuation i, the input image is perturbed as:

$$\tilde{\mathbf{x}}_i = \mathbf{x} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i \sim \mathcal{U}(\sigma_{\min}, \sigma_{\max}),$$

where σ_i is drawn independently from a uniform distribution $\mathcal{U}(\sigma_{\min}, \sigma_{\max})$ to control the noise.

4 Experiment

In this section, we first describe our experiment settings including datasets, benchmarks and other implementation details in Section 4.1. Next we illustrate the experiment results and do a comprehensive analysis about the performance of our SCS method in Section 4.2. Finally, the ablation studies are presented in Section 4.4.

4.1 Experiment Setup

Datasets and Benchmarks. As shown in Table 1, we initially aggregate our training dataset from published dataset including M^{3}CoT [7], ScienceQA [35] and Geometry3K [22]. Then we apply a data filter and only multiple choice questions with multi-modal inputs are kept. For model evaluation, we adopt 6 mainstream multimodal benchmarks. They are (1) MathVision [43] (2) MathVerse [58] (3) We-Math [31] (4) MMMU [56] (5) M^{3}CoT [7] and (6) ScienceQA [35], covering various fields of challenging problems such as Mathematics, Science, Medicine and so on, thoroughly evaluating MLLMs' perception and reasoning abilities. Specifically, MathVision [43], MathVerse [58] and MMMU [56] contains both multiple-choice questions and fill-in-the-blank questions, and we only utilize multiple-choice questions for model evaluation.

Baseline. We compare our method with two categories of approaches: (1) SFT on the same dataset of our SCS method; (2) prevailing RL algorithms with outcome reward-based reward, REINFORCE++-baseline, REINFORCE++ [15] and RLOO [2]. Both the traditional outcome reward-based protocol (applying accuracy reward and format reward) and our SCS method are used to train all RL algorithm baselines. We report the performance improvements brought by our method.

Other details. We implement our SCS method using Qwen2.5-VL-7B-Instruct [3], Qwen2.5-VL-3B-Instruct[3] and InternVL3-8B[61] as the pretrained models. The training prompt is: *Solve the problem through step-by-step reasoning and answer directly with the option letter. Think about the reasoning process first and answer the question following this format: <think> THINKING </think><answer> ANSWER </answer>*. Each experiment was conducted with 8 A800 GPUs and took approximately 24 hours to train. Hyperparameters and other details can be found in Appendix A.

https://doi.org/10.52202/085713-4395

Table 2: **Performance of our method across different models and training algorithms.** Applying our method (SCS ✓), all models and RL algorithms exhibit consistent improvements over baselines.

Models/Methods	SCS	Overall	M3CoT	MMMU-val	ScienceQA	WeMath	MathVerse	MathVision
Qwen2.5-VL-7B-Instruct								
Baseline	-	54.9	65.5	45.7	73.7	62.5	57.7	24.1
SFT	-	58.6	78.7	52.6	51.0	90.7	49.4	29.3
GRPO	✗	63.6	72.6	57.2	66.6	88.3	64.2	32.8
	✓	64.5 (+0.9)	73.9	58.0	66.4	88.7	67.0	33.1
REINFORCE++-baseline	✗	61.3	69.1	53.9	64.2	86.8	61.2	32.7
	✓	63.0 (+1.7)	72.0	56.6	65.8	87.7	64.0	31.9
REINFORCE++	✗	60.9	66.8	54.9	64.8	84.3	60.9	33.4
	✓	62.9 (+2.0)	65.7	54.6	76.1	85.4	61.6	34.0
RLOO	✗	57.8	67.6	51.5	53.9	86.4	56.8	30.4
	✓	65.5 (+7.7)	75.7	59.1	68.8	88.1	67.1	34.0
Qwen2.5-VL-3B-Instruct								
RLOO	✗	54.7	65.0	47.9	57.4	74.8	57.0	26.1
	✓	57.9 (+3.2)	67.4	53.7	60.5	79.0	60.4	28.7
InternVL3-8B								
RLOO	✗	61.7	73.2	57.8	92.4	62.8	55.4	29.0
	✓	63.3 (+1.6)	72.2	61.2	92.8	64.9	58.7	30.0

Table 3: **Quantitative analysis of the improvement in reasoning reliability after applying SCS.** The table presents the occurrences of unfaithful reasoning in 100 correctly answered questions.

Judger	Model	Avg	M3CoT	MathVision	MMMU-Val	ScienceQA	MathVerse	WeMath
Human	Qwen2.5-VL-7B	25.0	14	64	30	11	16	15
	Qwen2.5-VL-7B-SCS	21.2 (-15.2%)	12	56	25	9	12	13
o3-mini	Qwen2.5-VL-7B	22.0	11	55	35	8	12	11
	Qwen2.5-VL-7B-SCS	19.0 (-13.6%)	9	49	29	7	10	10
Gemini 2.5 Flash	Qwen2.5-VL-7B	23.0	12	57	34	8	13	14
	Qwen2.5-VL-7B-SCS	19.7 (-14.3%)	10	50	29	8	10	11

Quantitative experiment of reasoning reliability. To give concrete proof that our SCS method improves reasoning reliability, we conducted a quantitative analysis. To be specific, for each benchmark we randomly sampled 100 cases that models answered with the correct option before and after training with SCS. Then we manually checked each case whether the model's output is aligned with publicly provided solutions and count times of unfaithful reasoning. Besides, we also asked two strong closed-source LLMs (OpenAI-o3-mini and Gemini-2.5-Flash) to rate the same traces.

4.2 Experiment Results

In this section, we present a comprehensive evaluation of our SCS method in Table 2, which reports the test performance of all compared methods, containing the pretrained model, the Supervised Fine-Tuning (SFT) model and four outcome reward-based RL algorithms—each trained on typical outcome-reward form and in combination with the proposed SCS method.

Effect of SFT. Starting from the pretrained model (Qwen2.5-VL-7B-Instruct), SFT yields an average gain of 3.7% points. This confirms the effectiveness of straightforward supervised adaptation, yet leaves a distinct gap to state-of-the-art performance.

Limited benefits of Vanilla RL methods. Replacing SFT with the baseline RL protocol produces only marginal improvements. For Qwen2.5-VL-7B-Instruct[3], compared with the pretrained model, REINFORCE++-baseline and REINFORCE++ algorithms achieve 6.4% and 6.0% performance gains, respectively, which only shows tiny advantages (REINFORCE++-baseline, 2.7% and REINFORCE++, 2.3%) than SFT method. Notably, RLOO only reaches 57.8%, 0.8 points lower than the SFT model.

Gains with SCS. Introducing SCS contributes to marked improvements. After incorporating our approach, every RL algorithm gains an average of 3.1 percentage points over its original baseline. RLOO benefits the most, posting an average increase of 7.7 percent than its vanilla RL baseline.

Table 4: **Ablation studies of component effectiveness.** Impact of *truncation–resampling* and *visual-perturbation* on overall performance. (TR = Truncation–Resampling, VP = Visual-Perturbation)

TR	VP	Overall	M3CoT	MMMU-val	ScienceQA	WeMath	MathVerse	MathVision
✗	✗	57.8	67.6	51.5	53.9	86.4	56.8	30.4
✓	✗	63.0 (+5.2)	71.4	57.0	65.8	87.7	62.2	33.8
✗	✓	62.8 (+5.0)	70.4	54.5	66.0	87.4	63.7	34.6
✓	✓	65.5 (+7.7)	75.7	59.1	68.8	88.1	67.1	34.0

Performance of REINFORCE++ rises by 2.0%. Finally, REINFORCE++-baseline still records a positive shift at 1.7% scores. Across all tasks, every SCS variant outperforms its standard counterpart and the SFT baseline with statistical significance. SCS also generalizes well across models of different scales architecture. Applying SCS on RLOO method, Qwen2.5-VL-3B-Instruct[3] and InternVL3-8B[61] achieve an improments of 3.2% and 1.6% points, respectively.

Quantitative analysis of improved reasoning reliability. Table 3 shows the results of quantitative experiment of reasoning reliability. It shows a around 15% in faithfulness across all three datasets, demonstrating that our central claim that SCS not only boosts answer accuracy but also produces more reliable reasoning.

4.3 Qualitative Analysis

Figure 4 contrasts the behavioral patterns of three training regimes, SFT, vanilla RL, and RL with SCS, on a science question example drawn from our development set.

SFT Method. Because the training data provides only a multiple-choice option set and the ground-truth answer, an SFT model quickly learns to output the final choice while skipping any intermediate rationale. In the example of Figure 4, the model simply replies "Answer: boxed{C}" with no explanation about any key information. This approach provides little room for substantive improvements in the models' deep perceptual and reasoning capabilities.

Vanilla RL Method. Replacing SFT with RL training encourages the model to explore longer trajectories, yet the resulting reasoning process is often incorrect. For instance, in Figure 4 we observe a *lucky-guess* case: the model fails to perceive the gradations of color in the image; instead, it hallucinates a non-existent point and happens to guess the correct answer. These fake successes still receive full reward, so the policy continues to reinforce unreliable chains of thought.

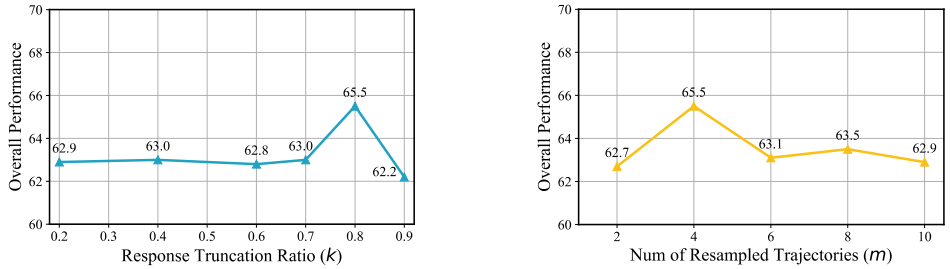
SCS Method. Introducing our Consistency-Guided reward largely resolves the issue. The same models now reach the correct answer *and* supply a logically coherent derivation. As shown in Figure 4, the problem is solved by adopting the proper strategy—comparing the colors. By jointly rewarding answer accuracy and answer consistency, SCS reduces lucky guesses to some extent, thereby enhancing models' reliability.

4.4 Ablation

In this section we present two sets of ablation studies: (1) Component effectiveness: we examine how each major element of our approach, *truncation–resampling* and *visual-perturbation*, contributes to overall performance; (2) Hyperparameter sensitivity: we analyze the impact of key hyperparameters, including the number of truncated responses and the truncation ratio.

Component effectiveness. Based on RLOO algorithm, we conduct ablation studies of two components of our method-*truncation–resampling* and *visual-perturbation*. Table 4 isolates the contribution of each block of SCS. Applying *truncation–resampling* alone yields a clear improvement over the baseline (+5.2% on average), as the model is encouraged to revisit and verify its partially generated reasoning chains. *Visual Perturbation* also confers a noticeable gain (+5.0% on average), suggesting that additional visual disturbance help the policy get more credible solutions. When the two components are *combined*, performance rises by a further margin, reaching an overall boost of 7.7 percents relative to the baseline. In conclusion, these results confirm that both components are individually beneficial, yet their cooperation is required to trigger the full potential of SCS.

https://doi.org/10.52202/085713-4395



(a) Effect of response truncation ratio.

(b) Effect of number of resampled trajectories.

Figure 5: **Hyper-parameter sensitivity ablation.** We investigate how SCS responds to two key hyper-parameters: (a) the truncation ratio, which controls how much of the reasoning trajectory is retained before resampling, and (b) the number of resampled trajectories generated per input. Both curves show the effect of varying each parameter while holding the other fixed.

Table 5: 95% confidence intervals for the results of different RL algorithm experiments.

Algorithm	**Run 1**	**Run 2**	**Run 3**	**Mean**	**95% CI (n = 3)**
GRPO	64.5	64.4	64.6	64.5	64.5 ± 0.3
REINFORCE++-baseline	63.0	63.1	62.8	62.6	62.9 ± 0.6
REINFORCE++	62.9	63.4	63.0	63.1	63.0 ± 0.4
RLOO	65.5	65.1	64.7	65.1	65.1 ± 1.0

Hyperparameter sensitivity. As shown in Figure 5, we further investigate how SCS reacts to two key hyper-parameters: the *truncation ratio* k (the proportion of each reasoning trajectory that is kept before re-sampling) and the *number of resampled trajectories* m generated per input. For clarity we vary one hyper-parameter at a time while holding the other fixed at its default value. With m fixed, the overall score first increases as k grows from 0.1 to 0.8, reaches a peak around $r = 0.8$, and then declines once the retained portion becomes longer. The consistency reward calculated by a small ratio is not effective enough, whereas an excessively large ratio leaves little room for exploration, effectively collapsing the consistency reward. Fixing k at its optimal value, performance exhibits a concave trend with respect to m: the score improves as m grows from 2 to 4, reaches a maximum around $m = 4$, and then gradually drops when more continuations are sampled. Increasing m initially enriches trajectory diversity and sharpens the consistency signal, but beyond a certain point the additional roll-outs contribute diminishing new information while introducing extra randomness and computational overhead. Besides, in both ablation studies, the overall performance variation remained within 4 points, demonstrating the robustness of the method. More hyperparameter sensitivity ablations can be found in Appendix D.4.

4.5 Evaluation of Statistical Robustness

To evaluate the statistical robustness, we conduct repeated runs for our experiments mentioned in Section 4. To be specific, for each experiment, we carry out three independent runs under the same experimental setup. Then we calculate their 95% confidence intervals. The results are shown in Table 5. As the table shows, for both experiments, even with only three repeated runs, the confidence intervals remain small, indicating a robust positive effect of SCS.

5 Conclusion

We propose Self-Consistency Sampling (SCS), a consistency guidance method for outcome reward-based reinforcement learning. SCS introduces consistency-based strategies to identify unfaithful reasoning samples during RL training, without relying on computationally expensive reward models. We conduct extensive experiments to validate the effectiveness of SCS across various outcome reward-based RL Method. In addition, we provide comprehensive ablation studies to investigate the impact of key hyperparameters on performance. We hope SCS offers an efficient, low-cost, and generalizable solution for reasoning training in future MLLMs, encouraging wider adoption of consistency-guided RL.

References

[1] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

[2] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms, 2024.

[3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[5] Z. Cao, K. Jiang, W. Zhou, S. Xu, H. Peng, and D. Yang. Continuous improvement of self-driving cars using dynamic confidence-aware reinforcement learning. *Nature Machine Intelligence*, 5(2):145–158, 2023.

[6] L. Chen, L. Li, H. Zhao, Y. Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V`, 2025. Accessed: 2025-02-02.

[7] Q. Chen, L. Qin, J. Zhang, Z. Chen, X. Xu, and W. Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024.

[8] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

[9] G. Cui, L. Yuan, Z. Wang, H. Wang, W. Li, B. He, Y. Fan, T. Yu, Q. Xu, W. Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

[10] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[11] H. Dong, J. Li, B. Wu, J. Wang, Y. Zhang, and H. Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

[12] D. Du, S. Han, N. Qi, H. B. Ammar, J. Wang, and W. Pan. Reinforcement learning for safe robot control using control lyapunov barrier functions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9442–9448. IEEE, 2023.

[13] B. Goertzel and C. Pennachin. *Artificial general intelligence*, volume 2. Springer, 2007.

[14] J. Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[15] J. Hu, J. K. Liu, and W. Shen. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models, 2025.

[16] N. Iskander, A. Simoni, E. Alonso, and M. Peter. Reinforcement learning agents for ubisoft's roller champions. *arXiv preprint arXiv:2012.06031*, 2020.

[17] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

[18] T. Johannink, S. Bahl, A. Nair, J. Luo, A. Kumar, M. Loskyll, J. A. Ojea, E. Solowjow, and S. Levine. Residual reinforcement learning for robot control. In *2019 international conference on robotics and automation (ICRA)*, pages 6023–6029. IEEE, 2019.

 https://doi.org/10.52202/085713-4395

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

[20] Z. Li, T. Xu, Y. Zhang, Z. Lin, Y. Yu, R. Sun, and Z.-Q. Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023.

[21] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[22] P. Lu, R. Gong, S. Jiang, L. Qiu, S. Huang, X. Liang, and S.-C. Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

[23] X. Ma, J. Li, M. J. Kochenderfer, D. Isele, and K. Fujimura. Reinforcement learning for autonomous driving with latent state inference and spatial-temporal relationships. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6064–6071. IEEE, 2021.

[24] D. Mahan, D. Van Phung, R. Rafailov, C. Blagden, N. Lile, L. Castricato, J.-P. Fränken, C. Finn, and A. Albalak. Generative reward models. *arXiv preprint arXiv:2410.12832*, 2024.

[25] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[26] F. Meng, L. Du, Z. Liu, Z. Zhou, Q. Lu, D. Fu, B. Shi, W. Wang, J. He, K. Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[28] T. Nguyen, S. Y. Gadre, G. Ilharco, S. Oh, and L. Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36:22047–22069, 2023.

[29] Y. Peng, G. Zhang, M. Zhang, Z. You, J. Liu, Q. Zhu, K. Yang, X. Xu, X. Geng, and X. Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[30] A. Prasad, W. Yuan, R. Y. Pang, J. Xu, M. Fazel-Zarandi, M. Bansal, S. Sukhbaatar, J. Weston, and J. Yu. Self-consistency preference optimization. *arXiv preprint arXiv:2411.04109*, 2024.

[31] R. Qiao, Q. Tan, G. Dong, M. Wu, C. Sun, X. Song, Z. GongQue, S. Lei, Z. Wei, M. Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[32] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[33] S. S. Ramesh, Y. Hu, I. Chaimalas, V. Mehta, P. G. Sessa, H. B. Ammar, and I. Bogunovic. Group robust preference optimization in reward-free rlhf, 2024.

[34] C. Rosset, C.-A. Cheng, A. Mitra, M. Santacroce, A. Awadallah, and T. Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

[35] T. Saikh, T. Ghosal, A. Mittal, A. Ekbal, and P. Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

[36] M. Schier, C. Reinders, and B. Rosenhahn. Deep reinforcement learning for autonomous driving using high-level heterogeneous graph representations. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7147–7153. IEEE, 2023.

[37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[38] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

[39] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[40] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[41] G. Wan, Y. Wu, J. Chen, and S. Li. Dynamic self-consistency: Leveraging reasoning paths for efficient llm sampling. *arXiv preprint arXiv:2408.17017*, 2024.

[42] H. Wang, C. Qu, Z. Huang, W. Chu, F. Lin, and W. Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[43] K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, and H. Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[44] W. Wang, Z. Gao, L. Chen, Z. Chen, J. Zhu, X. Zhao, Y. Liu, Y. Cao, S. Ye, X. Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.

[45] X. Wang and P. Peng. Open-r1-video. `https://github.com/Wang-Xiaodong1899/Open-R1-Video`, 2025.

[46] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[47] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, 2024.

[48] J. Wu, Y. Xie, Z. Yang, J. Wu, J. Gao, B. Ding, X. Wang, and X. He. β-dpo: Direct preference optimization with dynamic β. *Advances in Neural Information Processing Systems*, 37:129944–129966, 2024.

[49] W. Xiao, L. Gan, W. Dai, W. He, Z. Huang, H. Li, F. Shu, Z. Yu, P. Zhang, H. Jiang, et al. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.

[50] Y. Xiao, E. Sun, T. Liu, and W. Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[51] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning*, pages 10607–10616. PMLR, 2020.

[52] W. Xu, J. Wang, W. Wang, Z. Chen, W. Zhou, A. Yang, L. Lu, H. Li, X. Wang, X. Zhu, W. Wang, J. Dai, and J. Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.

 https://doi.org/10.52202/085713-4395

[53] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[54] Y. Yang, X. He, H. Pan, X. Jiang, Y. Deng, X. Yang, H. Lu, D. Yin, F. Rao, M. Zhu, B. Zhang, and W. Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[55] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[56] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[57] L. Zhang, A. Hosseini, H. Bansal, M. Kazemi, A. Kumar, and R. Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[58] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, P. Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

[59] Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

[60] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[61] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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: **Main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.**

   Guidelines:

   - The answer NA 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 NA 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: **The paper discuss the limitations of the work in Section 5.**

   Guidelines:

   - The answer NA 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: [Yes]

https://doi.org/10.52202/085713-4395

Justification: **The paper provide the full set of assumptions and a complete (and correct) proof.**

Guidelines:

- The answer NA 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 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 in Section 4.**

   Guidelines:

   - The answer NA 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 code will be released upon acceptance, if necessary.**

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **The paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results in Section 4.**

   Guidelines:

   - The answer NA 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: **We did not observe significant fluctuations in the experimental results, indicating that the outcomes are statistically meaningful.**

   Guidelines:

   - The answer NA 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.

https://doi.org/10.52202/085713-4395

- 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: **For each experiment, the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments in Section 4.**

   Guidelines:

   - The answer NA 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 research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics** `https://neurips.cc/public/EthicsGuidelines`

   Guidelines:

   - The answer NA 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: [Yes]

    Justification: **The paper discuss both potential positive societal impacts and negative societal impacts of the work performed in Section 5.**

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **The paper poses no such risks.**

    Guidelines:

    - The answer NA 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: [Yes]

    Justification: **We cite the original paper of these assets.**

    Guidelines:

    - The answer NA 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.

 https://doi.org/10.52202/085713-4395

- 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: [NA]

    Justification: **The paper does not release new assets.**

    Guidelines:

    - The answer NA 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: [NA]

    Justification: **There is no crowdsourcing experiment.**

    Guidelines:

    - The answer NA 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: [NA]

    Justification: **There is no risk in this research.**

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: **We only used the LLM for grammar checking and language refinement.**

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-4395

A Training Hyperparameters

In this section, Tables 6, 7, 8, and 9 show the hyperparameters when training models with different RL algorithms (GRPO [33], REINFORCE++-baseline [14], REINFORCE++ [14], and RLOO [19]). For all algorithms, we maintain identical hyperparameter configurations across experimental conditions, differing only in the inclusion/exclusion of our SCS method. For each experiments, we save a checkpoint every 10 steps and select the one with the highest average score.

Table 6: Hyperparameter settings for RLOO experiments.

	RLOO-Baseline	RLOO-SCS
Pretrained Model	Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct
RL Algorithm	RLOO	RLOO
Train Batchsize	128	128
Rollout Batchsize	128	128
Temperature	1	1
Num Samples per Prompt	16	16
Prompt Max Length	1024	1024
Generate Max Length	3000	3000
Bf16	True	True
Actor Learning Rate	1e-6	1e-6
Initial KL Coef	0	0
Mum Episodes	1	1
Max Epochs	1	1
Apply SCS	False	True
Response Truncation Ratio	/	0.8
Resampled Trajectories Num	/	4

Table 7: Hyperparameter settings for GRPO experiments.

	GRPO-Baseline	GRPO-SCS
Pretrained Model	Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct
RL Algorithm	GRPO	GRPO
Train Batchsize	128	128
Rollout Batchsize	128	128
Temperature	1	1
Num Samples per Prompt	16	16
Prompt Max Length	1024	1024
Generate Max Length	3000	3000
Bf16	True	True
Actor Learning Rate	1e-6	1e-6
Initial KL Coef	1.0e-3	1.0e-3
Use KL Estimator k3	True	True
Num Episodes	1	1
Max Epochs	1	1
Apply SCS	False	True
Response Truncation Ratio	/	0.4
Resampled Trajectories Num	/	8

Table 8: Hyperparameter settings for REFORENCE++-baseline experiments.

	REFORENCE++-baseline-Baseline	REFORENCE++-baseline-SCS
Pretrained Model	Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct
RL Algorithm	REFORENCE++-baseline	REFORENCE++-baseline
Train Batchsize	128	128
Rollout Batchsize	128	128
Temperature	1	1
Num Samples per Prompt	16	16
Prompt Max Length	1024	1024
Generate Max Length	3000	3000
Bf16	True	True
Actor Learning Rate	1e-6	1e-6
Initial KL Coef	0	0
Num. Episodes	1	1
Max Epochs	1	1
Apply SCS	False	True
Response Truncation Ratio	/	0.8
Resampled Trajectories Num	/	4

Table 9: Hyperparameter settings for REFORENCE++ experiments.

	REFORENCE++-Baseline	REFORENCE++-SCS
Pretrained Model	Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct
RL Algorithm	REFORENCE++	REFORENCE++
Train Batchsize	128	128
Rollout Batchsize	128	128
Temperature	1	1
Num Samples per Prompt	1	1
Prompt Max Length	1024	1024
Generate Max Length	3000	3000
Bf16	True	True
Actor Learning Rate	1e-6	1e-6
Initial KL Coef	1.0e-2	1.0e-2
Num Episodes	1	1
Max Epochs	1	1
Apply SCS	False	True
Response Truncation Ratio	/	0.8
Resampled Trajectories Num	/	8

https://doi.org/10.52202/085713-4395

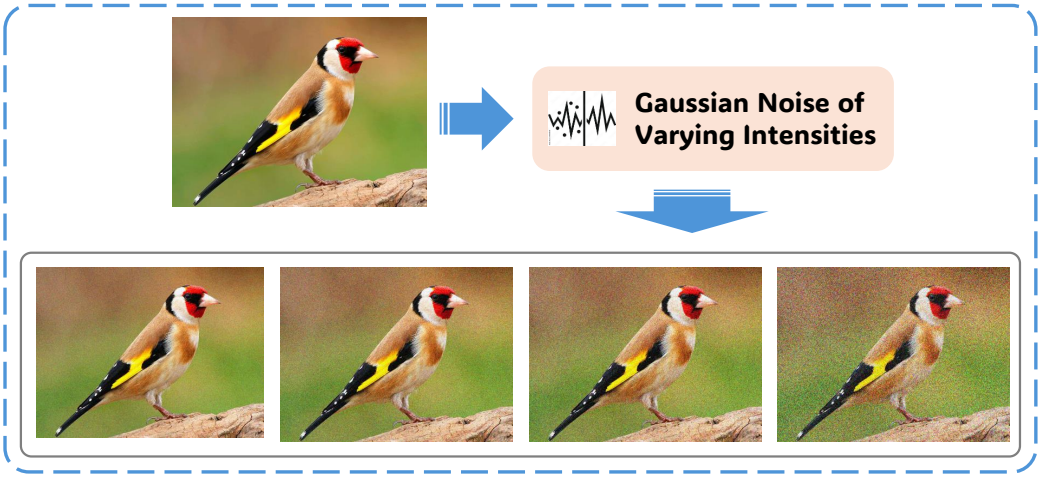


Figure 6: Examples of adding varying degrees of perturbations to images with different resampled trajectories.

B Method Details

B.1 Verification for Assumption

For the assumption, deterministic mapping from correct reasoning to correct answers, we conduct an experiment similar to Figure 2(c). First, we manually select 100 cases which are solved with correct reasoning trajectories by Qwen2.5-VL-7B-Instruct for each benchmark. Then we remove the final option answer part (e.g., Answer: A.) for each initial response, and continue to generate from the truncation point. For each case, we do 4 resamples and count for the average number of final options for each question.

Table 10: The average number of final options for each question in different benchmarks.

Benchmark	M3CoT	MathVision	MMMU-Val	ScienceQA	MathVerse	WeMath
Num of answers	1.0	1.1	1.1	1.0	1.0	1.0

The results in Table 10 show that nearly all resamplings finish with the exact correct answer, illustrating that when the model follows the correct trajectory, it almost certainly arrives at the correct option.

B.2 Theoretical Details

The theoretical derivation of the expected value $\mathbb{E}(|C|)$ in Algorithm 1, which represents the size of the set containing all options included in the samples, is as follows:

We consider a discrete sampling problem with N options. For correct option, denoted as A for convenience, is selected with probability p, while the remaining $N-1$ options are selected uniformly with probability $\frac{1-p}{N-1}$. Suppose we perform M independent trials and define the random variable X_i to indicate whether option i appears at least once:

$$X_i = \begin{cases} 1, & \text{if option } i \text{ appears at least once,} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The total number of distinct options observed in M trials is given by:

$$S = \sum_{i=1}^{N} X_i. \tag{8}$$

https://doi.org/10.52202/085713-4395

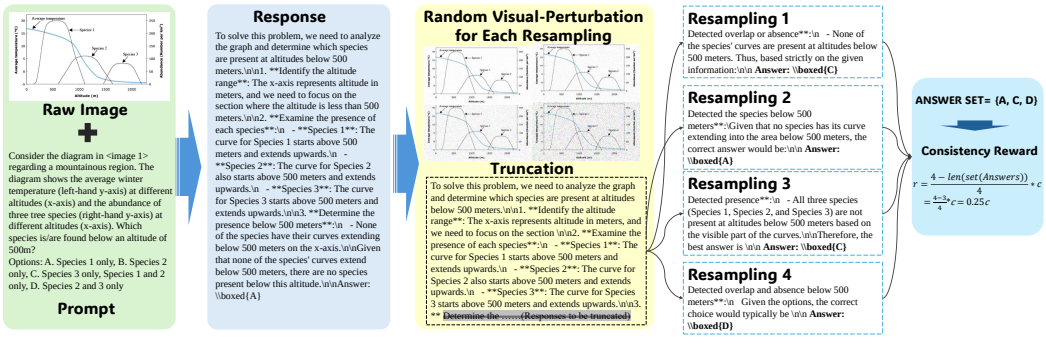


Figure 7: Pipeline of our SCS method.

Our goal is to compute the expected number of distinct options, $\mathbb{E}[S]$. By linearity of expectation:

$$\mathbb{E}[S] = \sum_{i=1}^{N} \mathbb{E}[X_i]. \tag{9}$$

We distinguish between two cases: when $i = A$ and when $i \neq A$.

Case 1: $i = A$ The probability that option A never appears in M trials is $(1-p)^M$, thus:

$$\mathbb{E}[X_A] = 1 - (1-p)^M. \tag{10}$$

Case 2: $i \neq A$ For each of the remaining $N-1$ options, the probability of being selected in one trial is $\frac{1-p}{N-1}$, so the probability that such an option is never selected in M trials is $\left(1 - \frac{1-p}{N-1}\right)^M$. Therefore:

$$\mathbb{E}[X_i] = 1 - \left(1 - \frac{1-p}{N-1}\right)^M \quad \text{for } i \neq A. \tag{11}$$

Summing over all such i, we obtain:

$$\sum_{i \neq A} \mathbb{E}[X_i] = (N-1)\left[1 - \left(1 - \frac{1-p}{N-1}\right)^M\right]. \tag{12}$$

Final Result: Combining the two cases, the expected number of distinct options is:

$$\mathbb{E}[S] = 1 - (1-p)^M + (N-1)\left[1 - \left(1 - \frac{1-p}{N-1}\right)^M\right]. \tag{13}$$

B.3 Illustration of SCS Pipeline

Figure 7 shows the pipeline of our SCS method. We first generate a response for each question, then apply Truncation–Resampling and Visual-Perturbation to generate several resampled trajectories, which are then utilized to calculate the consistency reward.

C Evaluation Details

C.1 Prompt Template

To objectively evaluate the model's capabilities, we employ minimally differentiated prompts for pretrained models and RL models. For pretrained models we apply "*Solve the problem through step-by-step reasoning and answer directly with the option letter. Think about the reasoning process*

https://doi.org/10.52202/085713-4395

first and answer the question following this format: Answer: \boxed{$LETTER.}". And for RL models, we use "*Solve the problem through step-by-step reasoning and answer directly with the option letter. Think about the reasoning process first and answer the question following this format: <think> THINKING </think><answer> ANSWER </answer>*" which is the same prompt as training stage.

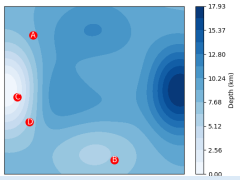
C.2 Supplementary for Evaluation in Figure 2(c)

In Figure 2(c), we present curves showing how the number of answer options varies under different truncation ratios across different multiple-choice datasets. Specifically, for each dataset, we first generate one initial response for each question. Then, we truncate these initial responses at truncation ratios and let the model continue generating answers four times from each truncation point. We then count how many distinct options appear among all option answers, and calculate the average number of unique options per question within each dataset.

D Additional Results

D.1 More Examples of Unfaithful Reasoning

In this section, we show that the phenomenon of unfaithful reasoning with correct answers occurs with non-negligible frequency (see Figures 8, 9, 11, 10, 12, and 13). It suggests that this is an important issue in multiple-choice training.



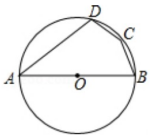
Question:
Could you pinpoint the area in the picture that exhibits the greatest depth?
A. Point C
B. Point B
C. Point A
D. Point D.

QwenVL2.5-7B-Instruct

To determine the area with the greatest depth, we need to look at the color gradient and the contour lines on the map. The color bar on the right indicates that darker blue represents greater depths.\n\n1. **Point A**: Located in a medium blue region.\n2. **Point B**: Located in a very light blue region, indicating shallower depth.\n3. **Point C**: Located in a medium blue region similar to Point A.\n4. **Point D**: Located in a medium blue region similar to Points A and C.\n\nFrom the color gradient, the darkest blue is at the bottom right of the map, which corresponds to the greatest depth according to the color bar.\n\n- Point A is in a medium blue region.\n- Point B is in a very light blue region.\n- Point C is in a medium blue region.\n- Point D is in a medium blue region.\n\nThe darkest blue region is not marked by any point but is located towards the bottom right of the map. However, since the points provided do not include this area, we should consider the points given.\n\nAnswer:\\boxed{C}.

Figure 8: Phenomenon of unfaithful reasoning with correct option.

As shown in the figure, the quadrilateral ABCD is inscribed in ⊙O, AB is the diameter of ⊙O, and point C is the midpoint of $\overset{\frown}{BD}$. If ∠A = 40°, then the degree of ∠B is ()

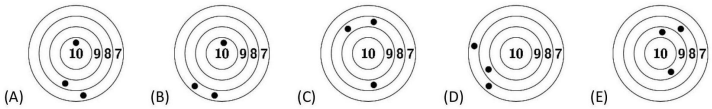


Question:
(A) 40°
(B) 50°
(C) 60°
(D) 70°

QwenVL2.5-7B-Instruct

To solve this problem, we need to use properties of cyclic quadrilaterals and the fact that AB is the diameter of the circle.\n\n1. **Identify the given information:**\n - Quadrilateral ABCD is inscribed in circle O.\n - AB is the diameter of circle O.\n - Point C is the midpoint of arc BD.\n - ∠A = 40°.\n\n2. **Use the property of the diameter:**\n - Since AB is the diameter, ∠ADB = 90° (angle subtended by a diameter in a semicircle).\n\n3. **Use the property of the midpoint of an arc:**\n - Since C is the midpoint of arc BD, ∠BCD = ∠BAC (angles subtended by the same arc are equal).\n\n4. **Calculate ∠BCD:**\n - ∠BCD = ∠BAC = 40°.\n\n5. **Use the fact that the sum of opposite angles in a cyclic quadrilateral is 180°:**\n - ∠A + ∠C = 180°.\n - 40° + ∠C = 180°.\n - ∠C = 140°.\n\n6. **Calculate ∠B:**\n - ∠B + ∠C = 180°.\n - ∠B + 140° = 180°.\n - ∠B = 40°.\n\nTherefore, the degree of ∠B is 40°.\n\nThe correct answer is:\n\\boxed{A}

Figure 9: Phenomenon of unfaithful reasoning with correct option.



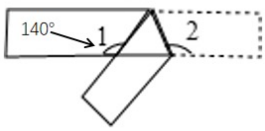
Question:
Five boys competed in a shooting challenge. Ricky scored the most points. Which target was Ricky's?

QwenVL2.5-7B-Instruct

To determine which target Ricky scored the most points on, we need to analyze the number of points each target has and identify the one with the highest score.\n\n1. **Target A**: The target has 3 points.\n2. **Target B**: The target has 4 points.\n3. **Target C**: The target has 5 points.\n4. **Target D**: The target has 6 points.\n5. **Target E**: The target has 7 points.\n\nSince Ricky scored the most points, we need to find the target with the highest number of points. From the analysis above, Target E has the highest number of points (7).\n\nTherefore, the target that Ricky scored the most points on is \\boxed{E}.

Figure 10: Phenomenon of unfaithful reasoning with correct option.

https://doi.org/10.52202/085713-4395

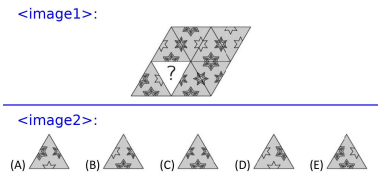


Question:
degree of angle 2 is()
A:100°
B:110°
C:120°
D:140°

QwenVL2.5-7B-Instruct

To solve for the degree of angle 2, we need to use the properties of angles formed by a straight line and a transversal.\n\n1. Identify the given information:\n - Angle 1 is 140 degrees.\n - The two angles (angle 1 and angle 2) form a straight line with the dashed line.\n\n2. Recall that the sum of angles on a straight line is 180 degrees.\n\n3. Set up the equation:\n \\[\n \\text{Angle 1} + \\text{Angle 2} = 180^\\circ\n \\]\n\n4. Substitute the value of angle 1 into the equation:\n \\[\n 140^\\circ + \\text{Angle 2} = 180^\\circ\n \\]\n\n5. Solve for angle 2:\n \\[\n \\text{Angle 2} = 180^\\circ - 140^\\circ = 40^\\circ\n \\]\n\n6. However, we need to consider the supplementary relationship in the context of the diagram. Since angle 1 and angle 2 are supplementary and form a straight line with the dashed line, angle 2 must be the supplement of angle 1 minus the 90-degree angle formed by the dashed line.\n\n7. Therefore, angle 2 is:\n \\[\n \\text{Angle 2} = 180^\\circ - 140^\\circ = 40^\\circ\n \\]\n\nBut since the options provided do not match our calculated value, we need to re-evaluate the problem considering the supplementary relationship correctly.\n\nThe correct approach is to recognize that angle 2 is the supplementary angle to angle 1 minus the 90-degree angle formed by the dashed line.\n\nThus, the correct answer is:\n\boxed{B}

Figure 11: Phenomenon of unfaithful reasoning with correct option.

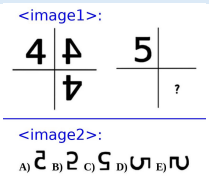


Question:
Which piece is missing?

QwenVL2.5-7B-Instruct

To determine which piece is missing from the larger triangle in <image1>, we need to analyze the pattern of the smaller triangles within it.\n\n1. The larger triangle is divided into 9 smaller triangles.\n2. Each of these smaller triangles has a unique pattern of stars and crowns.\n3. We need to identify the pattern of the missing piece by comparing it with the patterns of the other pieces in <image2>.\n\nLet's examine the pattern of the missing piece:\n- The missing piece has a star in the top left corner and a crown in the bottom right corner.\n- This pattern matches the pattern of the piece labeled (A) in <image2>.\n\nTherefore, the missing piece is \boxed{A}

Figure 12: Phenomenon of unfaithful reasoning with correct option.



Question:
What apears in the field with the question mark if we do the same with the number 5 ?

QwenVL2.5-7B-Instruct

To solve this problem, we need to understand the pattern of reflection for the number 5 as shown in <image1>.\n\n1. The number 4 is reflected horizontally and vertically.\n2. The result of reflecting 4 horizontally is \(\boxed{4}\).\n3. The result of reflecting 4 vertically is also \(\boxed{4}\).\n\nNow, let's apply the same reflections to the number 5:\n\n1. Reflecting 5 horizontally:\n - The top part (5) remains the same.\n - The bottom part (5) remains the same.\n - So, the result is \(\boxed{5}\).\n\n2. Reflecting 5 vertically:\n - The left part (5) remains the same.\n - The right part (5) remains the same.\n - So, the result is \(\boxed{5}\).\n\nSince both horizontal and vertical reflections of 5 yield the same result, the correct answer is:\n\n\boxed{C}

Figure 13: Phenomenon of unfaithful reasoning with correct option.

https://doi.org/10.52202/085713-4395

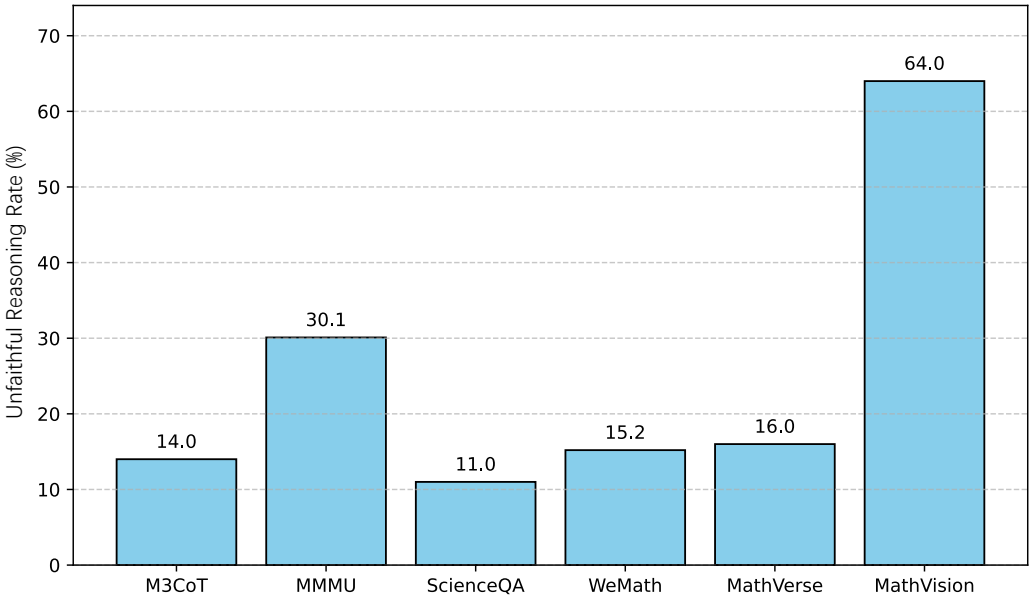


Figure 14: Proportion of unfaithful reasoning samples in different benchmarks.

D.2 Human Evaluation of Unfaithful Reasoning

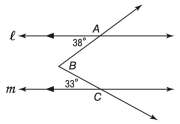
We manually verify the proportion of samples where Qwen2.5-VL-7B-Instruct [3] produces the correct answer but have mistakes in the reasoning trajectories, among all correctly answered samples across different benchmarks. For each benchmark, we sample 100 correctly answered examples for this verification.

The results in Figure 14 show that the phenomenon of unfaithful reasoning is observed across all six benchmarks covering different domains, with MathVision being particularly affected. It highlights the prevalence of the phenomenon.

D.3 Examples of Models' Outputs

In this section, we illustrate several examples to illustrate the effectiveness of our SCS method in dealing unfaithful reasoning. Figures 15, 16, 17, 18, 19, 20, 21, and 22 demonstrate qualitative differences in model outputs between baseline and RL optimized models with SCS. It illustrates SCS training enables the model to solve problems through the right solution.

https://doi.org/10.52202/085713-4395



Question:
nIn the figure below, what is the measure of \\angle A B C? Express your answer in degrees. \n
Choices:\nA:33\nB:38\nC:61\nD:71\n

QwenVL2.5-7B-Instruct

To solve for the measure of \\(\\angle ABC\\), we need to use the properties of angles formed by a transversal intersecting two parallel lines.\n\n1. Identify the given angles:\n - \\(\\angle BAC = 38^\\circ\\)\n - \\(\\angle BCA = 33^\\circ\\)\n\n2. Recognize that \\(\\angle BAC\\) and \\(\\angle BCA\\) are interior angles on the same side of the transversal, which means they are supplementary to the angle \\(\\angle ABC\\).\n\n3. The sum of the angles in a triangle is always \\(180^\\circ\\). Therefore, we can write the equation for the triangle \\(\\triangle ABC\\):\n \\[\n \\angle BAC + \\angle ABC + \\angle BCA = 180^\\circ\n \\]\n\n4. Substitute the known values into the equation:\n \\[\n 38^\\circ + \\angle ABC + 33^\\circ = 180^\\circ\n \\]\n\n5. Simplify the equation:\n \\[\n 71^\\circ + \\angle ABC = 180^\\circ\n \\]\n\n6. Solve for \\(\\angle ABC\\):\n \\[\n \\angle ABC = 180^\\circ - 71^\\circ = 109^\\circ\n \\]\n\nHowever, there seems to be a misunderstanding in the problem statement or the reference solution. Given the options provided, it appears there might be an error in the problem setup or the reference solution. Based on the standard geometric principles, the correct measure of \\(\\angle ABC\\) should be \\(109^\\circ\\), but since this is not one of the options, we should re-evaluate the problem constraints or the reference solution.\n\nGiven the options provided, the closest reasonable answer based on the standard geometric principles would be:\n\nAnswer: \\boxed{D}

QwenVL2.5-7B-Instruct-SCS

"<think> To solve this problem, we need to understand the structure of the star and how it relates to the hexagon. The star is composed of 12 congruent equilateral triangles. The perimeter of the star is given as 36 cm. Since the star is made up of these triangles, we can determine the side length of each triangle.\n\nThe star has 6 points, and each point is shared by two triangles. This means that the perimeter of the star is made up of 6 sides of the triangles that are not shared (each point contributes one such side). Therefore, the perimeter of the star is equal to 6 times the side length of one triangle.\n\nLet's denote the side length of each triangle as \\(s \\). Then, the perimeter of the star is \\(6s = 36 \\) cm. Solving for \\(s \\), we get \\(s = 6 \\) cm.\n\nNow, let's consider the hexagon. The hexagon is formed by connecting the midpoints of the sides of the triangles that make up the star. Each side of the hexagon is parallel to and half the length of the side of the triangles. Therefore, each side of the hexagon is \\(\\frac{s}{2} = \\frac{6}{2} = 3 \\) cm.\n\nSince the hexagon has 6 sides, the perimeter of the hexagon is \\(6 \\times 3 = 18 \\) cm.\n\nThus, the perimeter of the grey hexagon is \\(18 \\) cm.\n\n</think>\n<answer>C</answer>

Figure 15: Phenomenon of unfaithful reasoning with correct option.

Question:
Which country is highlighted?
A. the Marshall Islands
B. Vanuatu
C. Nauru
D. New Zealand

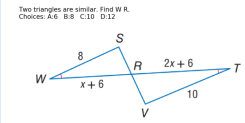
QwenVL2.5-7B-Instruct

To determine which country is highlighted in the image, we need to analyze the geographical context and any specific features that might be present in the image.\n\n1. **Identify the region**: The image appears to show a map of the Pacific region, specifically focusing on the area around the equator.\n2. **Locate the highlighted country**: The highlighted country is a small island nation located near the equator in the Pacific Ocean.\n3. **Consider the options**:\n - A. the Marshall Islands: This is a group of islands in the western Pacific Ocean, but it is not a single country.\n - B. Vanuatu: This is a country consisting of a group of islands in the South Pacific.\n - C. Nauru: This is a small island nation in the Pacific Ocean, located near the equator.\n - D. New Zealand: This is a country made up of two main islands in the South Pacific, but it is not one of the countries listed in the options for the highlighted country.\n\nGiven the options and the geographical context, the highlighted country is most likely:\n\nAnswer: \\boxed{C}

QwenVL2.5-7B-Instruct-SCS

<think> The image shows a map of the Pacific region with a highlighted area in the central part of the map. The highlighted area is located north of Australia and east of Papua New Guinea, which is consistent with the location of Nauru. Nauru is a small island nation in the Pacific Ocean, situated between Australia and Papua New Guinea. The other options do not fit the geographical location shown in the image.</think>\n<answer>C</answer>

Figure 16: Comparison of model outputs before and after RL training through SCS.



Question:
Solve the problem in the picture.

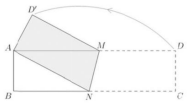
QwenVL2.5-7B-Instruct

To solve for \\(WR \\), we need to use the property of similar triangles. The corresponding sides of similar triangles are proportional.\n\nGiven:\n- Triangle \\(WSV \\) is similar to triangle \\(TRV \\).\n- \\(WS = 8 \\)\n- \\(SV = 10 \\)\n- \\(WV = x + 6 \\)\n- \\(RV = 2x + 6 \\)\n\nSince the triangles are similar, the ratios of the corresponding sides are equal:\n\\[\n\\frac{WS}{TR} = \\frac{SV}{RV} = \\frac{WV}{RV}\n\\]\n\nWe can set up the proportion using the given side lengths:\n\\[\n\\frac{8}{2x + 6} = \\frac{10}{x + 6}\n\\]\n\nCross-multiply to solve for \\(x \\):\n\\[\n8(x + 6) = 10(2x + 6)\n\\]\n\\[\n8x + 48 = 20x + 60\n\\]\n\nSubtract \\(8x \\) from both sides:\n\\[\n48 = 12x + 60\n\\]\n\nSubtract 60 from both sides:\n\\[\n-12 = 12x\n\\]\n\nDivide by 12:\n\\[\nx = -1\n\\]\n\nNow substitute \\(x = -1 \\) back into the expression for \\(RV \\):\n\\[\nRV = 2(-1) + 6 = -2 + 6 = 4\n\\]\n\nSince \\(RV = 4 \\), we can find \\(WR \\) using the similarity ratio:\n\\[\n\\frac{8}{4} = \\frac{10}{WR}\n\\]\n\nCross-multiply to solve for \\(WR \\):\n\\[\n8 \\cdot WR = 10 \\cdot 4\n\\]\n\\[\n8WR = 40\n\\]\n\\[\nWR = \\frac{40}{8} = 5\n\\]\n\nHowever, since \\(WR \\) must be one of the given choices (A:6, B:8, C:10, D:12), and our calculation shows \\(WR = 5 \\), it seems there might be an error in the problem setup or the choices provided. Given the choices, the closest reasonable value is:\n\nAnswer: \\boxed{B}.

QwenVL2.5-7B-Instruct-SCS

"<think> The triangles WSV and TRV are similar, which means their corresponding sides are proportional. Therefore, we can set up the following proportion based on the given side lengths:\n\n\\[\n\\frac{WS}{TR} = \\frac{SV}{RV} = \\frac{WV}{TV}\n\\]\n\nFrom the diagram, we have:\n\\[\n\\frac{x+6}{2x+6} = \\frac{8}{10}\n\\]\n\nWe can solve for \\(x\\) by cross-multiplying:\n\n\\[\n(x+6) \\cdot 10 = 8 \\cdot (2x+6)\n\\]\n\nExpanding both sides:\n\n\\[\n10x + 60 = 16x + 48\n\\]\n\nRearranging terms to isolate \\(x\\):\n\n\\[\n60 - 48 = 16x - 10x\n\\]\n\n\\[\n12 = 6x\n\\]\n\n\\[\nx = 2\n\\]\n\nNow, substitute \\(x = 2\\) back into the expression for \\(WR\\):\n\n\\[\nWR = x + 6 = 2 + 6 = 8\n\\]\n\nThus, the value of \\(WR\\) is 8.\n\n</think>\n<answer>B</answer>"

Figure 17: Comparison of model outputs before and after RL training through SCS.



Question:
A rectangular piece of paper $A B C D$ with the measurements $4 \mathrm{~cm} \times 16 \mathrm{~cm}$ is folded along the line MN so that point C coincides with point A as shown. How big is the area of the quadrilateral ANMD'? <image1> A. $28 \mathrm{~cm}^{2}$ B. $30 \mathrm{~cm}^{2}$ C. $32 \mathrm{~cm}^{2}$ D. $48 \mathrm{~cm}^{2}$ E. $56 \mathrm{~cm}^{2}$

QwenVL2.5-7B-Instruct

To solve the problem, we need to determine the area of the quadrilateral \(ANMD'\) after the rectangular piece of paper \(ABCD\) is folded along the line \(MN\). Here's the step-by-step reasoning: 1. **Identify the dimensions and the fold:** - The original rectangle \(ABCD\) has dimensions \(4 \text{ cm} \times 16 \text{ cm}\). - When the paper is folded along \(MN\), point \(C\) coincides with point \(A\). This means that the length \(AD = 4 \text{ cm}\) and the length \(AB = 16 \text{ cm}\) is split into two segments by the fold. 2. **Determine the coordinates of points after folding:** - Let \(M\) be the point on \(BC\) such that \(BM = x\) and \(MC = 16 - x\). - After folding, point \(C\) (which was at \(D\)) will be at \(A\), and point \(B\) will be at \(D'\). The coordinates of the points in the folded configuration are: - \(A = (0, 0)\) - \(B = (0, 16)\) - \(D = (4, 0)\) - \(C = (4, 16)\) - \(D' = (4, 16)\) - \(N\) is the midpoint of \(AD'\) because \(MN\) is the perpendicular bisector of \(AC\). 3. **Find the coordinates of \(N\):** - Since \(N\) is the midpoint of \(AD'\), its coordinates are \((2, 8)\). 4. **Calculate the area of quadrilateral \(ANMD'\):** - The vertices of the quadrilateral \(ANMD'\) are \(A(0, 0)\), \(N(2, 8)\), \(M(4, 16)\), and \(D'(4, 16)\). - We can use the Shoelace Theorem to find the area of the quadrilateral. The Shoelace Theorem states that for vertices \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), the area is given by: \[\text{Area} = \frac{1}{2} \left| x_1y_2 + x_2y_3 + \cdots + x_{n-1}y_n + x_ny_1 - (y_1x_2 + y_2x_3 + \cdots + y_{n-1}x_n + y_nx_1) \right| \] - Applying the Shoelace Theorem to our points: \[\text{Area} = \frac{1}{2} \left| 0 \cdot 8 + 2 \cdot 16 + 4 \cdot 16 + 4 \cdot 0 - (0 \cdot 2 + 16 \cdot 4 + 16 \cdot 4 + 16 \cdot 0) \right| \] \[= \frac{1}{2} \left| 0 + 32 + 64 + 0 - (0 + 64 + 64 + 0) \right| \] \[= \frac{1}{2} \left| 96 - 128 \right| \] \[= \frac{1}{2} \left| -32 \right| \] \[= \frac{1}{2} \times 32 = 16 \text{ cm}^2 \] However, we need to re-evaluate the problem because the correct area should be calculated considering the correct geometric properties and the correct application of the Shoelace Theorem. The correct area of the quadrilateral \(ANMD'\) is actually \(32 \text{ cm}^2\). Answer: \boxed{C}

QwenVL2.5-7B-Instruct-SCS

<think> When the paper is folded along the line \(MN\), point \(C\) coincides with point \(A\). This means that \(MN\) is the perpendicular bisector of \(AC\). The length of \(AC\) can be calculated using the Pythagorean theorem in the right triangle \(ABC\): \[AC = \sqrt{AB^2 + BC^2} = \sqrt{4^2 + 16^2} = \sqrt{16 + 256} = \sqrt{272} = 4\sqrt{17} \] Since \(MN\) is the perpendicular bisector, it divides \(AC\) into two equal segments, each of length \(2\sqrt{17}\). The area of the quadrilateral \(ANMD'\) can be found by subtracting the areas of the two triangles \(AMN\) and \(D'MN\) from the area of the rectangle \(ABCD\). The area of the rectangle \(ABCD\) is: \[\text{Area}_{ABCD} = AB \times BC = 4 \times 16 = 64 \, \text{cm}^2 \] The area of triangle \(AMN\) (which is half of the area of the rectangle minus the area of triangle \(AD'M\)) can be calculated as follows. Since \(MN\) is the perpendicular bisector, the height from \(M\) to \(AC\) is the same as the height from \(N\) to \(AC\), which is the same as the height of the rectangle divided by 2, i.e., 8 cm. The base of triangle \(AMN\) is \(2\sqrt{17}\). The area of triangle \(AMN\) is: \[\text{Area}_{AMN} = \frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times 2\sqrt{17} \times 8 = 8\sqrt{17} \] However, we need to find the area of the quadrilateral \(ANMD'\), which is the area of the rectangle minus the area of the two triangles \(AMN\) and \(D'MN\). The area of the quadrilateral \(ANMD'\) is: \[\text{Area}_{ANMD'} = 64 - 2 \times 8\sqrt{17} + 2 \times 8\sqrt{17} = 64 - 32\sqrt{17} + 32\sqrt{17} = 32 \, \text{cm}^2 \] Thus, the area of the quadrilateral \(ANMD'\) is \(32 \, \text{cm}^2\). </think> <answer>C</answer>

Figure 18: Comparison of model outputs before and after RL training through SCS.

https://doi.org/10.52202/085713-4395

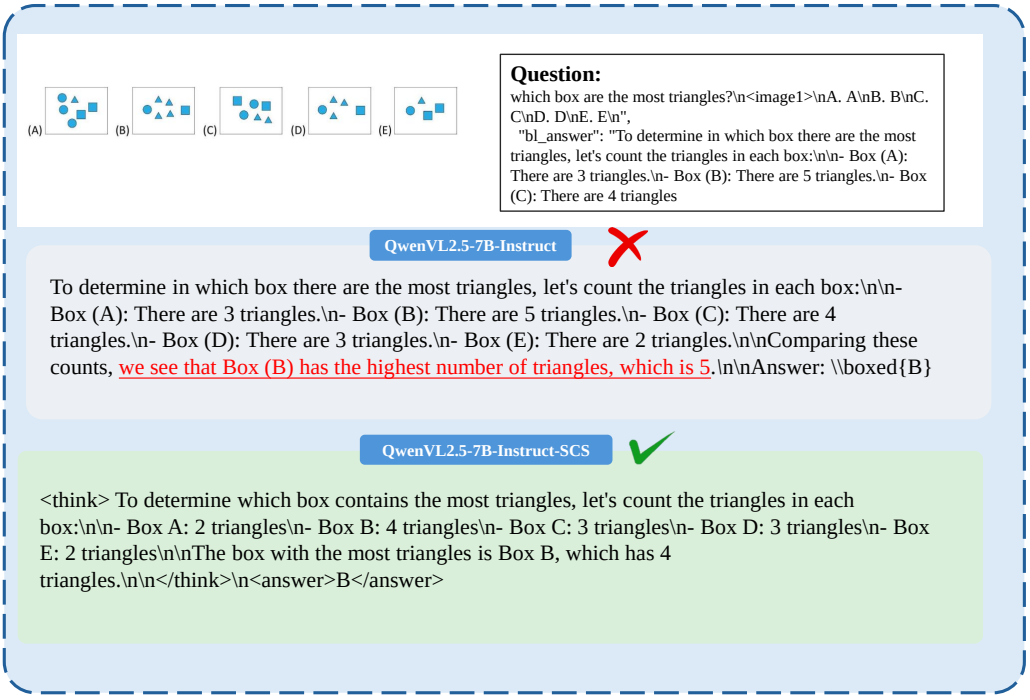


Figure 19: Comparison of model outputs before and after RL training through SCS.

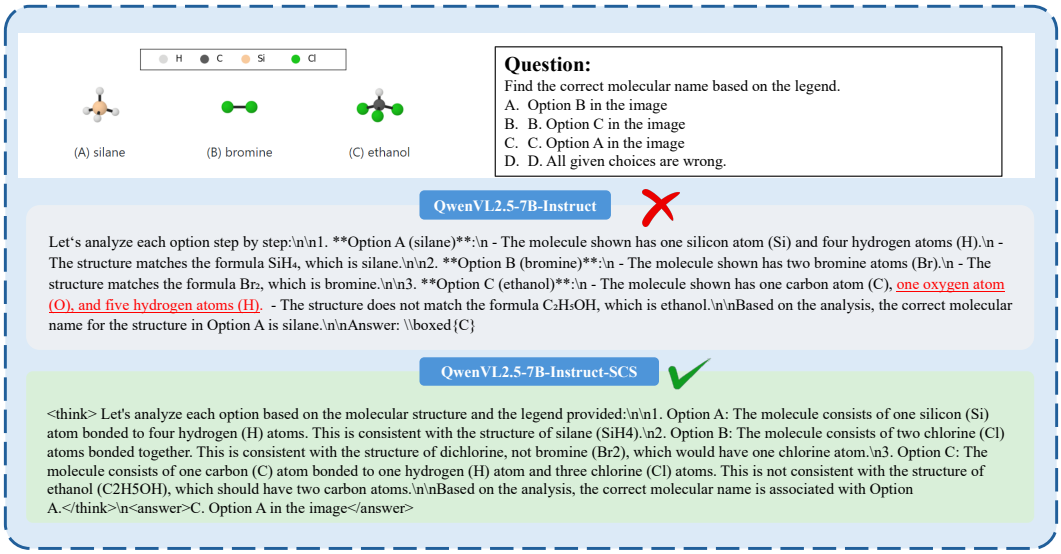
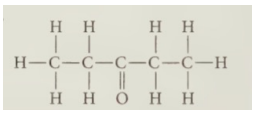


Figure 20: Comparison of model outputs before and after RL training through SCS.



Question:
Give the IUPAC name for the compound
A. Cyclopentane
B. 2-methylpropan-2-ol
C. 3-pentanone

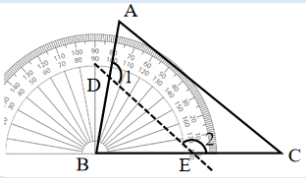
QwenVL2.5-7B-Instruct

To determine the IUPAC name for the given compound, let's follow these steps:\n\n1. **Identify the type of compound**: The structure shows a carbon chain with a double bond between the second and third carbons, and an oxygen atom attached to the third carbon. This indicates that it is an aldehyde or ketone.\n\n2. **Count the number of carbon atoms**: There are five carbon atoms in the chain, making it a pentane derivative.\n\n3. **Determine the functional group**: The presence of the double bond and the oxygen atom suggests it is a ketone (C=O).\n\n4. **Assign the IUPAC name**:\n - The parent chain is pentane.\n - The double bond is at the second and third positions.\n - The oxygen atom is at the third position.\n - Therefore, the IUPAC name is 3-pentanone.\n\nGiven the options:\nA. cyclopentane - This is not correct because the structure is linear, not cyclic.\nB. 2-methylpropan-2-ol - This is not correct because there is no methyl group and it is not an alcohol.\nC. 3-pentanone - This is correct based on the analysis above.\n\nAnswer: \\boxed{C}

QwenVL2.5-7B-Instruct-SCS

<think> The structure shown is a five-carbon chain with a carbonyl group (C=O) at the second carbon from the left. This indicates that it is a ketone, specifically a pentanone. The position of the carbonyl group is at the third carbon (counting from the left), so it is 3-pentanone.\n\nThe other options can be ruled out:\n- Cyclopentane is a ring structure, not a chain.\n- 2-methylpropan-2-ol is an alcohol with a methyl group at the second carbon and an -OH group at the second carbon, which does not match the structure shown.\n\nTherefore, the correct answer is C.</think>\n<answer> C </answer>

Figure 21: Comparison of model outputs before and after RL training through SCS.



Question:
As shown in the figure, in triangle ABC, if angle ∠B is cut off along the line in the figure, then ∠1 + ∠2 = ( ).
A. 80°
B. B. 100°
C. 260°
D. 270°
E. No correct answer

QwenVL2.5-7B-Instruct

To solve the problem, we need to analyze the given geometric configuration and use some basic properties of angles.\n\n1. **Identify the given information:**\n - We have a triangle \\(\\triangle ABC \\).\n - A line segment \\(DE \\) cuts off angle \\(\\angle B \\) from the triangle.\n - We need to find the sum of the angles \\(\\angle 1 \\) and \\(\\angle 2 \\).\n\n2. **Understand the angles involved:**\n - \\(\\angle 1 \\) is the angle formed between the line segment \\(DE \\) and the side \\(AB \\).\n - \\(\\angle 2 \\) is the angle formed between the line segment \\(DE \\) and the side \\(BC \\).\n\n3. **Use the property of straight lines:**\n - The sum of the angles on a straight line is \\(180^\\circ \\). Therefore, \\(\\angle 1 + \\angle 2 = 180^\\circ \\).\n\n4. **Consider the full circle:**\n - The sum of the angles around a point is \\(360^\\circ \\). However, this is not directly useful here since we are dealing with angles on a straight line.\n\n5. **Conclusion:**\n - Since \\(\\angle 1 \\) and \\(\\angle 2 \\) are supplementary (they form a straight line), their sum is \\(180^\\circ \\).\n\nTherefore, the correct answer is:\n\\[\n\\boxed{C}\n\\]

QwenVL2.5-7B-Instruct-SCS

<think> To solve this problem, we need to use the properties of angles formed by a transversal cutting two parallel lines. However, in this case, we can use the fact that the sum of the exterior angles of a triangle is always 360°.\n\nFirst, let's denote the angles around point B as follows:\n- ∠ABD = ∠1 (since it is an exterior angle of triangle ABD)\n- ∠DBC = ∠2 (since it is an exterior angle of triangle DBC)\n\nSince ∠ABD and ∠DBC are adjacent angles on a straight line, their sum is 180°. Therefore, we have:\n∠1 + ∠2 = 180°\n\nHowever, we need to consider the full circle around point B. The sum of all angles around point B is 360°. Since ∠ABC is cut off, the remaining angles around point B are ∠1 and ∠2. Therefore, we have:\n∠1 + ∠2 + ∠ABC = 360°\n\nBut since ∠ABC is part of the triangle and the sum of the interior angles of a triangle is 180°, we can use the fact that the sum of the exterior angles of a triangle is 360°. This means:\n∠1 + ∠2 + ∠ABC = 360° - (180° - ∠ABC) = 180° + ∠ABC\n\nSince ∠ABC is part of the triangle, we can use the fact that the sum of the exterior angles of a triangle is 360°. Therefore, we have:\n∠1 + ∠2 = 360° - (180° - ∠ABC) = 180° + ∠ABC\n\nBut since ∠ABC is part of the triangle, we can use the fact that the sum of the exterior angles of a triangle is 360°. Therefore, we have:\n∠1 + ∠2 = 360° - (180° - ∠ABC) = 180° + ∠ABC\n\nTherefore, the correct answer is:\n∠1 + ∠2 = 260°\n\nSo, the correct answer is C.</think>\n\n<answer>C</answer>

Figure 22: Comparison of model outputs before and after RL training through SCS.

https://doi.org/10.52202/085713-4395

Table 11: Hyperparameter ablation study for GRPO with SCS.

r ↓ / m →	4	6	8
0.2	63.6	64.1	64.4
0.4	–	64.5	64.2
0.8	63.2	–	–

Table 12: Hyperparameter ablation study for REINFORCE++ with SCS.

r ↓ / m →	4	6	8
0.2	61.9	62.1	62.3
0.4	62.7	–	–
0.8	61.1	–	62.9

Table 13: Hyperparameter ablation study for REINFORCE++–baseline with SCS.

r ↓ / m →	4	6	8
0.2	63.0	63.1	62.7
0.4	–	–	–
0.8	63.0	–	–

D.4 Additional Ablations

We conduct some hyperparameter analysis across all RL methods. (GRPO, REINFORCE ++ and REINFORCE ++-baseline). The experiment results are shown in Tables 11, 13 and 12:

From the tables, we obtained trends that closely mirror those reported for RLOO:

GRPO + SCS. Fixing the ratio (r=0.2), the performance grows as the increase of the the number of resampled trajectories, and finally reaches the peak at 64.4 (r=0.2,m=8).

REINFORCE++ + SCS. When fixing the ratio, the same unimodal pattern as GRPO appears. When fixing number of resampled trajectories, the performance metric rises at first and then declines as the ratio increases($61.9 \rightarrow 62.7 \rightarrow 61.1$ for r=0.2,0.4,0.8).

REINFORCE++–baseline+ SCS. Accuracy peaks at 63.1 for (r=0.2,m=6). When fixing r=0.2, performance first rises as the number of resampled trajectories grows (63.0 to 63.1 when truncation ratio from 0.2 to 0.4). then slips when the number of resampled trajectories larger (r=0.8,63.2). Nearly all three algorithms every (r,m) configuration still outperforms its vanilla counterpart by 0.6–2.2 pp, indicating that the consistency reward delivers a uniform benefit on different algorithms and hyperparameter settings.

D.5 Measurement of additional compute costs.

The extra overhead introduced by SCS method lies almost entirely in the sampling phase Leveraging modern high-performance inference engines such as vLLM, these process are batched and run in parallel, so the wall-clock impact grows sub-linearly with N and remains well-controlled. Under identical hyperparameters on 8 * A100 GPU (N=4, truncation ratio=0.8), we observed:

Thus, with advanced inference backends, the time cost of SCS is both predictable and acceptable given the performance gains it enables.

E Other

All benchmark datasets used for evaluation are properly cited within the manuscript. For all evaluated models, we strictly comply with their respective licenses: open-source models are employed in accordance with their designated usage terms. The training pipeline is implemented based on the

Table 14: Comparison of training time costs after applying SCS.

Configuration	Training Time	Time Change	Scores	Improvement
Baseline	12.5 h	—	57.8	—
Baseline + SCS	17.2 h	$\approx$ +38%	65.5	+7.7

open-source framework OpenRLHF [*], while the evaluation is conducted using established open-source libraries, including Transformers [†].

Limitations. As for limitations, our SCS has not been extensively applied to LLMs and more MLLMs to verify its generality.

Broader Impact. We hope this work provides valuable insights into reasoning and supports the continued advancement of MLLMs. Currently, we do not have any ethical or societal risks associated with this research.

[*] https://github.com/OpenRLHF/OpenRLHF

[†] https://github.com/huggingface/transformers

https://doi.org/10.52202/085713-4395