

Stable and Efficient Single-Rollout RL for Multimodal Reasoning

Supplementary Material

5.1. Preliminaries

Group Relative Policy Optimization. In the multimodal group-based approach GRPO [27], given an input $x = (x_{\text{text}}, x_{\text{image}})$, a group of responses $\{o_i\}_{i=1}^G$ are sampled from the old policy $\pi_{\theta_{\text{old}}}$, each associated with a reward r_i . Then the normalized advantage for response o_i is defined as
$$A_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^G)}{\text{std}(\{r_i\}_{i=1}^G)}.$$

GRPO uses clipped importance sampling to stabilize policy updates. Let $\rho_i(\theta) = \frac{\pi_{\theta}(o_i|x)}{\pi_{\theta_{\text{old}}}(o_i|x)}$ denote the probability ratio between the new and old policies. The GRPO objective is to maximize the following:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{o_i\} \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^G \min(\rho_i(\theta) A_i, \text{clip}(\rho_i(\theta), 1 - \epsilon, 1 + \epsilon) A_i) \right], \quad (1)$$

where ϵ is the clipping hyperparameter.

5.2. Prompt Templates

We list below the prompt used to instruct the model to produce the structured outputs.

System Prompt

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST be enclosed within `<think></think>` tags. The final answer MUST be put in `\boxed{\}`.

5.3. Adapting Text-Only Single-Rollout Approach

We adapt the text-only Single-stream Policy Optimization (SPO) [35] to multimodal settings with the Qwen2.5-VL-7B model, and present the training accuracy and model entropy in Figure 7. However, training is unstable, with declining accuracy and entropy collapse, highlighting the instability due to multimodality.

5.4. Baseline Estimate

During MSSR training with the 7B model, we track the mean and standard deviation of the baseline estimate, modeled as a Beta distribution. As shown in Figure 8, the mean steadily increases over training while the standard deviation decreases.

5.5. Training Cost

We measure the training cost for both 3B and 7B models in terms of minutes per training step. As shown in Table 3, MSSR maintains a per-step cost similar to GRPO, introducing only minimal overhead due to the initial baseline estimation, while requiring far fewer total steps to achieve the final performance level of GRPO, as shown in Figure 1.

Table 3. Training cost across methods on Qwen2.5-VL 3B and 7B models, measured as the average training time per step (min/step).

Model	Train Cost (mins/step)
Qwen2.5-VL-3B [2]	–
+ GRPO [27]	3.9
+ RLOO [1]	3.9
+ REINFORCE++ [13]	4.8
+ MSSR	3.9
Qwen2.5-VL-7B [2]	–
+ GRPO [27]	6.1
+ RLOO [1]	6.2
+ REINFORCE++ [13]	8.3
+ MSSR	6.9

5.6. Ablation on Entropy Loss Coefficients

We vary the coefficients over $\{0.01, 0.05, 0.10, 0.15\}$ for MVSR and track the resulting model entropy. As illustrated in Figure 9, larger coefficients initially increase entropy, but collapse occurs in all cases. MSSR, in contrast, maintains entropy effectively, demonstrating superior stabilization.

5.7. Ablation on Fixed η schedule.

We compare MSSR with a fixed $\eta = \frac{\eta_{\text{min}} + \eta_{\text{max}}}{2}$ and the adaptive schedule. As shown in Figure 10, both yield stable and similar training accuracy, while adaptive η achieves better final validation accuracy. The results demonstrate the robustness of MSSR to the update rule in the baseline estimate.

5.8. Training on Different Dataset.

We train with the 7B model on another dataset MMRL30K [44], which contains around 30K samples. We show the training and validation accuracy curves in Figure 11. MSSR still remains stable and improves steadily, whereas MVSR is unstable and collapses. MSSR also reaches a similar final



Figure 7. **Training accuracy and model entropy of Qwen2.5-VL-7B when adapting the text-only single-rollout approach SPO to multimodal settings.** Training is unstable, characterized by declining accuracy and entropy collapse, highlighting the instability due to multimodality.

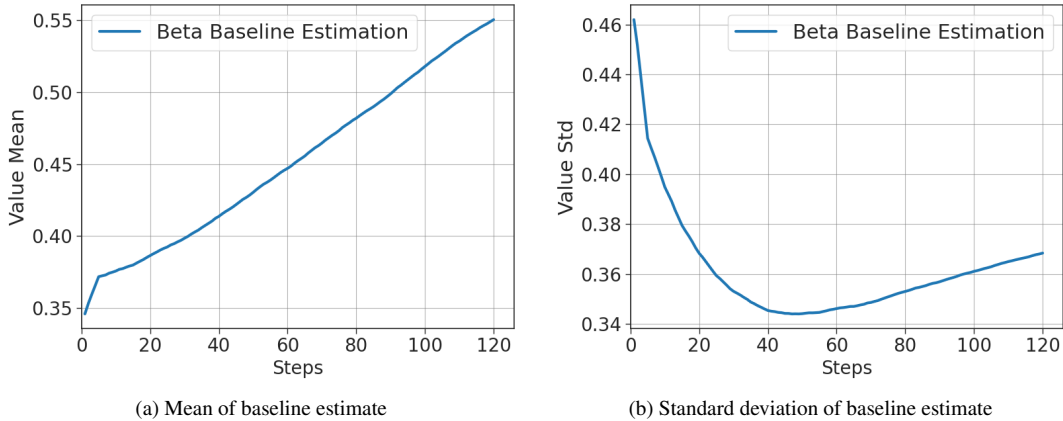


Figure 8. **Mean and standard deviation of the Beta baseline estimate during MSSR training with Qwen2.5-VL-7B.** The mean steadily increases over training while the standard deviation decreases, indicating a reliably learned baseline.

validation accuracy to GRPO with about half the training steps, reinforcing its superior training compute efficiency.

5.9. Training Accuracy for Sensitivity Analysis

In Section 4.4, we study how the sliding window size N and the target KL divergence KL_{target} affect validation accuracy. We find that $N = 20$ and $KL_{\text{target}} = 0.01$ yield the strongest performance. For completeness, we also report the corresponding training accuracy curves in Figure 12. The training trends across different choices of N and KL_{target} remain broadly similar.

5.10. Additional Performance Comparison

We compare MVSR and MSSR on Qwen2.5-VL 3B and 7B models for their generalization performance on diverse multimodal reasoning benchmarks. We present the results in Table 4. Due to the training instability, MVSR provides

limited gains over the base model for the 7B setting, and for the 3B model, its fine-tuned performance is often lower than the base model across most benchmarks. In contrast, MSSR consistently improves generalization performance across all benchmarks.

5.11. Reasoning Case Studies

In Section 4 Experiments, we present a representative case study in Figure 5 comparing the reasoning outputs of GRPO and MSSR. MSSR arrives at the correct solution, whereas GRPO fails. We mark the erroneous reasoning steps of GRPO in red and the key steps enabling MSSR’s correct prediction in green.

For illustration, we provide additional qualitative examples in Figures 13, 14, and 15. Across these cases, GRPO gives incorrect answers, while MSSR successfully solves the problems, further showcasing its stronger and more re-

Table 4. **Model generalization performance on diverse multimodal reasoning benchmarks.** We compare MVSR and MSSR on Qwen2.5-VL 3B and 7B models. As discussed earlier, MVSR’s training instability leads to limited gains for the 7B model and, for the 3B model, often results in performance below the base model across most benchmarks. In contrast, MSSR consistently improves generalization performance across all benchmarks.

Model	MathVerse	MathVista	MMK12	R1-Onevision Bench	HallusionBench	Avg.
Qwen2.5-VL-3B [2]	33.3	59.5	42.5	27.6	59.9	44.6
+ MVSR	31.3	52.9	41.3	15.2	64.8	41.1
+ MSSR	39.6	63.0	49.2	29.0	66.6	49.5
Qwen2.5-VL-7B [2]	45.8	67.2	48.1	34.6	68.4	52.8
+ MVSR	43.9	69.0	51.7	35.5	68.7	53.8
+ MSSR	49.8	71.1	62.5	39.2	70.6	58.6

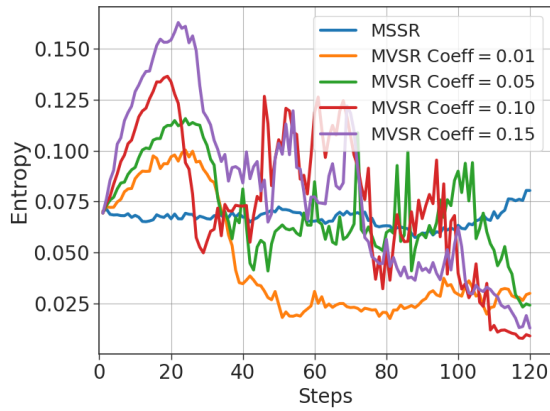


Figure 9. **Ablation on entropy loss coefficients for combing with MVSR.** Larger coefficients initially increase entropy for MVSR, but collapse occurs in all cases. In contrast, MSSR effectively maintains entropy, demonstrating superior stability.

liable reasoning capability.

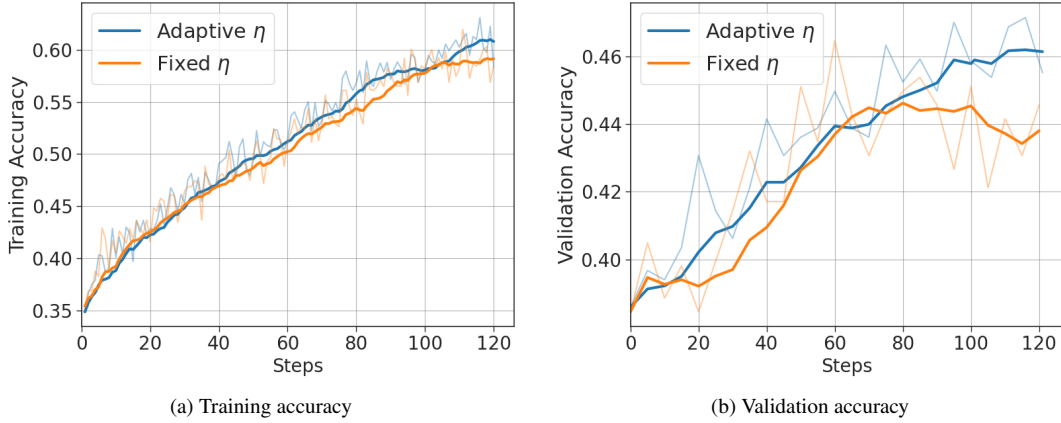


Figure 10. **Ablation on the η schedule in the Beta baseline estimate.** We compare MSSR with a fixed $\eta = \frac{\eta_{\min} + \eta_{\max}}{2}$ and the adaptive schedule. Both yield stable and similar training accuracy, while adaptive η achieves better final validation accuracy. This demonstrates the robustness of MSSR to the update rule for the baseline estimate.

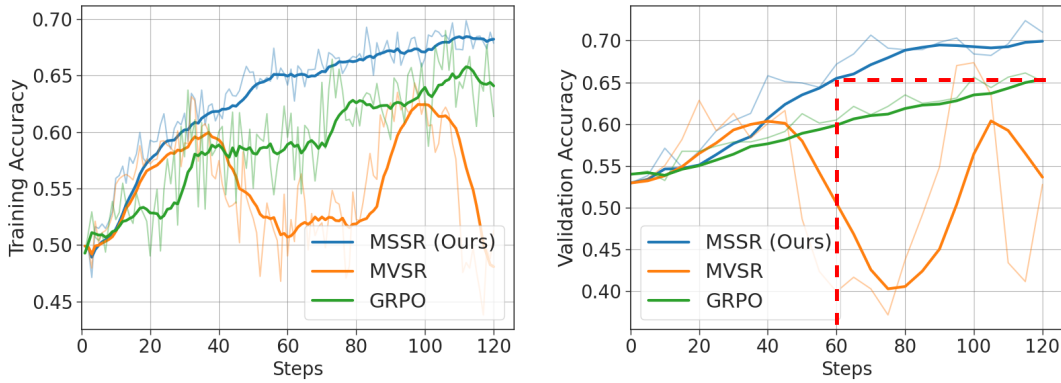


Figure 11. **Training and validation accuracy on the MMRL30K dataset.** We train with the 7B model on another dataset MMRL30K [44]. MSSR still remains stable and improves steadily, whereas MVSR is unstable and collapses. MSSR also reaches a similar final validation accuracy to GRPO with about half the training steps, reinforcing its superior training compute efficiency.

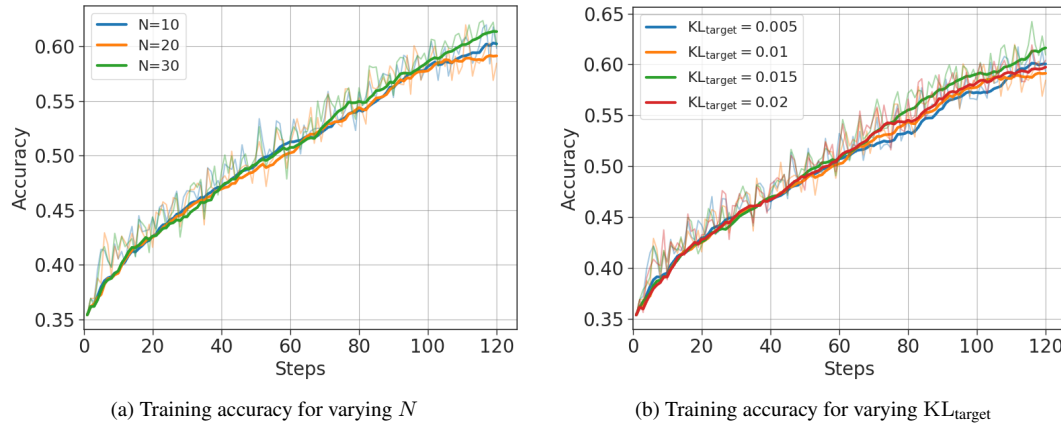


Figure 12. **Sensitivity analysis for effects of sliding window size N and target KL value KL_{target} on adjusting the discount factor η and their impact on training accuracy.** (a) Training accuracy under different sliding window sizes N . (b) Training accuracy under different target KL values KL_{target} . Across all settings, the training accuracy curves remain largely similar.

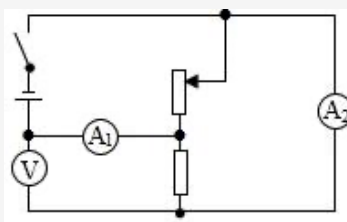
η	-2	-1	0	1	2	3
P	0.1	0.2	0.2	0.3	0.1	0.1

Question: If the distribution of the random variable η is as follows: then when $P(\eta < x) = 0.8$, the range of the real number x is (). A. $x < 1$. B. $1 < x < 2$. C. $1 < x < 2$. D. $1 < x < 2$. Answer: C

GRPO: To determine the range of the real number x such that $P(\eta < x) = 0.8$, ... This means that $P(\eta < 1) = 0.8$ and $P(\eta < 2) = 0.9$. Therefore, the value of x must be in the range $1 \leq x < 2$. The correct answer is \boxed{D} .

MSSR: To determine the range of the real number x for which $P(\eta < x) = 0.8$, ... Therefore, $P(\eta < 1) = 0.5$ and $P(\eta < 2) = 0.8$ Since we need $P(\eta < x) = 0.8$, the value of x must be in the range where the cumulative probability is at least 0.8, which is $1 < x \leq 2$. Thus, the correct answer is: \boxed{C}

Figure 13. Example of reasoning outputs. Comparing the Qwen2.5-VL-7B model fine-tuned with GRPO and MSSR. While GRPO produce incorrect answers, MSSR successfully solves the problem, demonstrating its superior reasoning capability. We highlight the critical reasoning steps that lead to GRPO's incorrect answer in red, and the key steps enabling MSSR's correct prediction in green.

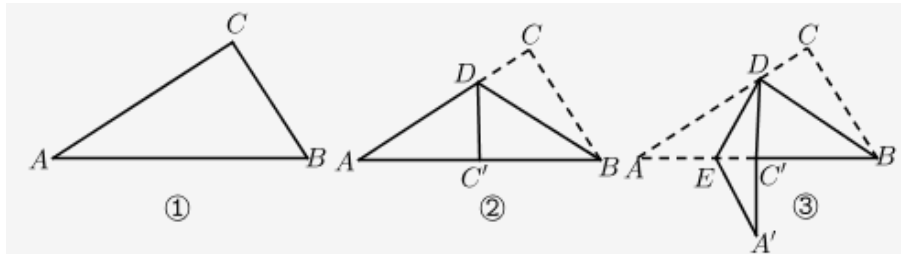


As shown in the figure, the power supply voltage remains constant. After closing the switch, when the slider of the variable resistor moves upwards, which of the following statements is correct? A. The readings of all three meters increase B. The readings of all three meters decrease C. The reading of meter A_1 decreases, while the readings of meters V and A_2 remain unchanged D. The readings of meters A_1 and A_2 decrease, while the reading of meter V remains unchanged Answer: C.

GRPO: When the slider of the variable resistor moves upwards, the resistance in the parallel branch increases. ... Since both currents are decreasing, the total current through A_1 will also decrease ... Since the total current through A_1 is decreasing, the current through A_2 will also decrease. Therefore, the readings of meters A_1 and A_2 will decrease, while the reading of meter V will remain unchanged. The correct answer is \boxed{D} .

MSSR: To answer this question, we need to analyze the circuit behavior when the slider of the variable resistor moves upwards. ... The total current (measured by A_1) will decrease because the current through the variable resistor decreases. ... Determine the effect on the voltage: - The voltage across the parallel combination (measured by the voltmeter V) remains unchanged because the voltage source is constant. Based on this analysis: - The reading of A_1 (the current through the variable resistor) decreases. - The readings of V (the voltage across the parallel combination) and A_2 (the current through the fixed resistor) remain unchanged. The correct statement is: \boxed{C}

Figure 14. Example of reasoning outputs. Comparing the Qwen2.5-VL-7B model fine-tuned with GRPO and MSSR. While GRPO produce incorrect answers, MSSR successfully solves the problem, demonstrating its superior reasoning capability. We highlight the critical reasoning steps that lead to GRPO's incorrect answer in red, and the key steps enabling MSSR's correct prediction in green.



As shown in Figure 1, it is a right-angled triangular paper piece, $\angle A = 30^\circ$, $BC = 4 \text{ cm}$, and it is folded so that point C lands on point C' on the hypotenuse, with the fold line being BD , as shown in Figure 2. Then, Figure 2 is folded along DE , so that point A lands on point A' on the extension of BC' , as shown in Figure 3. What is the length of the fold line DE ?
 A. $\frac{8}{3} \text{ cm}$ B. $2\sqrt{3} \text{ cm}$ C. $2\sqrt{2} \text{ cm}$ D. 3 cm

Answer: A

GRPO: <think>\nTo solve this problem, we need to analyze the geometric transformations step by step. ... $(BC = 4)$, so $(AB = 4\sqrt{3})$ and $(AC = 8)$. Since (BD) is the angle bisector, we can use the fact that (BD) splits (AC) into segments proportional to the adjacent sides. However, a simpler way is to use the fact that (BD) is the altitude of the 30-60-90 triangle, so $(AD = 4\sqrt{3} \cdot \frac{\sqrt{3}}{2} = 6)$.
 The length of (DE) is the same as the length of the altitude of the 30-60-90 triangle, which is $(2\sqrt{3})$.
 Therefore, the length of the fold line (DE) is (\boxed{B}) .

MSSR: To solve this problem, we need to analyze the geometric properties and the fold lines in the right-angled triangle. ... The hypotenuse (AB) is (8 cm) (since $(AB = 2 \times BC = 2 \times 4 = 8 \text{ cm})$).
 3. **Analyze the first fold:** ... In $(\triangle BDC')$, $(\angle BDC' = 90^\circ)$ and $(\angle DBC' = 30^\circ)$.
 - The length of (BD) can be found using the properties of the 30-60-90 triangle.
 5. **Analyze the second fold:** ... In $(\triangle ADE)$, $(\angle ADE = 90^\circ)$ and $(\angle DAE' = 30^\circ)$.
 - The length of (DE) can be found using the properties of the right triangle.
 The length of (DE) is $(\frac{8}{3} \text{ cm})$.
 The length of (DE) is (\boxed{A}) .

Figure 15. **Example of reasoning outputs.** Comparing the Qwen2.5-VL-7B model fine-tuned with GRPO and MSSR. While GRPO produce incorrect answers, MSSR successfully solves the problem, demonstrating its superior reasoning capability. We highlight the critical reasoning steps that lead to GRPO's incorrect answer in red, and the key steps enabling MSSR's correct prediction in green.